Internet Draft
Document: draft-pate-pwe3-framework-01.txt
Expires: January 13, 2002

Prayson Pate
Overture Networks
XiPeng Xiao
Photuris Inc.
Tricci So
Caspian Networks
Craig White
Level 3 Communications, LLC.
Kireeti Kompella
Juniper Networks, Inc.
Andrew G. Malis
Vivace Networks
Thomas K. Johnson
Litchfield Communications

Framework for Pseudo Wire Emulation Edge-to-Edge (PWE3)

draft-pate-pwe3-framework-01.txt

Status of this Memo

Abstract

This document describes a framework for Pseudo Wires Emulation Edge-to-Edge (PWE3). It discusses the emulation of circuits (such as T1, E1, T3, E3 and SONET/SDH) and services (such as ATM and Frame Relay) over packet switched networks (PSNs) using IP, L2TP or MPLS. It presents an architectural framework for pseudo wires (PWs), defines terminology, specifies the various protocol elements and their functions, overviews some of the services that will be supported and discusses how PWs fit into the broader context of protocols.

Copyright Notice

Table of Contents

1. Introduction

This document describes a framework for Pseudo Wires Emulation Edge-to-Edge (PWE3). It discusses the emulation of circuits (such as T1, E1, T3, E3 and SONET/SDH) and services (such as ATM and Frame Relay) over packet switched networks (PSNs) using IP, L2TP or MPLS. It presents an architectural framework for pseudo wires (PWs), defines terminology, specifies the various protocol elements and their functions, overviews the services supported and discusses how PWs fit into the broader context of protocols.

1.1. What Are Pseudo Wires?

1.1.1. Definition

PWE3 is a mechanism that emulates the essential attributes of a service (such as a T1 leased line or Frame Relay) over a PSN. The required functions of PWs include encapsulating service-specific bit-streams or PDUs arriving at an ingress port, and carrying them across a path or tunnel, managing their timing and order, and any other operations required to emulate the behavior and characteristics of the service as faithfully as possible.

From the customer perspective, the PW is perceived as an unshared link or circuit of the chosen service. However, there may be deficiencies that impede some applications from being carried on a PW. These limitations should be fully described in the appropriate service-specific Applicability Statements (ASes).

1.1.2. Functions

PWs provide the following functions in order to emulate the behavior and characteristics of the desired service.

- Encapsulation of service-specific PDUs or circuit data arriving at an ingress port (logical or physical).

- Carrying the encapsulated data across a tunnel.

- Managing the signaling, timing, order or other aspects of the service at the boundaries of the PW.

- Service-specific status signaling and alarm management.

ASes for each service will describe any shortfalls of the emulation's faithfulness.

1.2. Goals of This Document

- Description of the motivation for creating PWs, and some background on how they may be deployed.

- Description of an architecture and terminology for PWs.

- Description of the relevant services that will be supported by PWs, including any relevant service-specific considerations.

- Description of methods to ensure in-order final PDU delivery,

- Description of methods to perform clock recovery, as needed or appropriate.

- Description of methods to perform edge-to-edge/inband signaling functions across the PSN, as needed or appropriate.

- Description of the statistics and other network management information needed for tunnel operation and management.

- Description of the security mechanisms to be used to protect the control of the PW technology. The protection of the encapsulated content (e.g., payload encryption) of the PW is outside of scope.

- Description of a mechanism to exchange encapsulation control information at an administrative boundary of the PSN, including security methods.

- Whenever possible, relevant requirements from existing IETF documents and other sources will be incorporated by reference.

1.3. Non-Goals

The following are out of scope:

- The protection of the encapsulated content of the PW.

- Any multicast service not native to the emulated medium. Thus, Ethernet transmission to a "multicast" IEEE-48 address is in scope, while multicast services like MARS that are implemented on top of the medium are out of scope.

- Methods to signal the underlying PSN.

2. Background and Motivation

Many of today's service providers are struggling with the dilemma of moving to an optical network based on IP and/or MPLS. How do they realize the capital and operational benefits of a new packet-based optical infrastructure, while leveraging the existing base of SONET (Synchronous Optical Network) gear, and while also protecting the large revenue stream associated with this equipment? How do they move from mature Frame Relay or ATM networks, while still being able to provide these lucrative services? One possibility is the emulation of circuits or services via PWs. Circuit emulation over ATM and interworking of Frame Relay and ATM have already been standardized. Emulation allows existing circuits and/or services to be carried across the new infrastructure, and thus enables the interworking of disparate networks. [ATMCES] provides some insight into the requirements for such a service:

> There is a user demand for carrying certain types of

> constant bit rate (CBR) or "circuit" traffic over Asynchronous Transfer Mode (ATM) networks. As ATM is essentially a packet- rather than circuit-oriented transmission technology, it must emulate circuit characteristics in order to provide good support for CBR traffic.
>
> A critical attribute of a Circuit Emulation Service (CES) is that the performance realized over ATM should be comparable to that experienced with the current PDH/SDH technology.

Section 4 of [ANAVI] gives more background on why such emulation is desirable:

> The simplicity of TDMoIP translates into initial expenditure and operational cost benefits. In addition, due to its transparency TDMoIP can support mixed voice, data and video services. It is transparent to both protocols and signaling, irrespective of whether they are standards based or proprietary with full timing support and the capability of maintaining the integrity of framed and unframed DS1 formats.

2.1. Current Network Architecture

2.1.1. Multiple Networks

For any given service provider delivering multiple services, the current "network" usually consists of parallel or "overlay" networks. Each of these networks implements a specific service, such as voice, Frame Relay, Internet access, etc. This is quite expensive, both in terms of capital expense as well as in operational costs. Furthermore, the presence of multiple networks complicates planning.

Service providers wind up asking themselves these questions:

- Which of my networks do I build out?

- How many fibers do I need for each network?

- How do I efficiently manage multiple networks?

2.1.2. Convergence Today

There are some examples of convergence in today's network:

- Frame Relay is frequently carried over ATM networks using [FRF.5] interworking.

- T1, E1 and T3 circuits are sometimes carried over ATM networks using [ATMCES].

- Voice is carried over ATM (using AAL2), Frame Relay (using FRF.11 VoFR), IP (using VoIP) and MPLS (using VoMPLS) networks.

Deployment of these examples range from limited (ATM CES) to fairly common (FRF.5 interworking) to rapidly growing (VoIP).

2.2. The Emerging Converged Network

Many service providers are finding that the new IP-based and MPLS-based switching systems are much less costly to acquire, deploy and maintain than the systems that they replace. The new systems take advantage of advances in technology in these ways:

- The newer systems leverage mass production of ASICs and optical interfaces to reduce capital expense.

- The bulk of the traffic in the network today originates from packet sources. Packet switches can economically switch and deliver this traffic natively.

- Variable-length switches have lower system costs than ATM due to simpler switching mechanisms as well as elimination of segmentation and reassembly (SAR) at the edges of the network.

- Deployment of services is simpler due to the connectionless nature of IP services or the rapid provisioning of MPLS applications.

2.3. Transition to a IP-Optimized Converged Network

The greatest assets for many service providers are the physical communications links that they own. The time and costs associated with acquiring the necessary rights of way, getting the required governmental approvals, and physically installing the cabling over a variety of terrains and obstacles represents a significant asset that

is difficult to replace. Their greatest on-going costs are the operational expenses associated with maintaining and operating their

networks. In order to maximize the return on their assets and minimize their operating costs, service providers often look to consolidate the delivery of multiple service types onto a single networking technology.

The first generation converged network is based on TDM (time-division multiplexing) technology. Voice, video, and data traffic has been carried successfully across TDM/DACS-based networks for decades. TDM technology has some significant drawbacks as a converged networking technology. Operational costs for TDM networks remain relatively high because the provisioning of end-to-end TDM circuits is typically a tedious and labor-intensive task. In addition, TDM switching does not make the best use of the communications links. This is because fixed assignment of timeslots does not allow for the statistical multiplexing of bursty data traffic (i.e. temporarily unused bandwidth on one timeslot cannot be dynamically re-allocated to another service).

The second generation of converged network is based on ATM technology. Today many service providers convert voice, video, and data traffic into fixed-length cells for carriage across ATM-based networks. ATM improves upon TDM technology by providing the ability to statistically multiplex different types of traffic onto communications links. In addition, ATM SPVC technology is often used to automatically provision end-to-end services, providing an additional advantage over traditional TDM networks. However, ATM has several significant drawbacks. One of the most frequently cited problems with ATM is the so-called cell-tax, which refers to the 5 bytes out of 53 used as an ATM cell header. Another significant problem with ATM is the AAL5 SAR, which becomes extremely difficult to implement above 1 Gbps. There are also issues with the long-term scalability of ATM, especially as a switching layer beneath IP.

As IP traffic takes up a larger and larger portion of the available network bandwidth, it becomes increasingly useful to optimize public networks for the Internet Protocol. However, many service providers are confronting several obstacles in engineering IP-optimized networks. Although Internet traffic is the fastest growing traffic segment, it does not generate the highest revenue per bit. For example, Frame Relay traffic currently generates a higher revenue per bit than do native IP services. Private line TDM services still generate even more revenue per bit than does Frame Relay. In addition, there is a tremendous amount of legacy equipment deployed within public networks that does not communicate using the Internet Protocol. Service providers continue to utilize non-IP equipment to deploy a variety of services, and see a need to interconnect this legacy equipment over their IP-optimized core networks.

To maximize the return on their assets and minimize their operational costs, many service providers are looking to consolidate the delivery

of multiple service offerings and traffic types onto a single IP-optimized network.

In order to create this next-generation converged network, standard methods must be developed to emulate existing telecommunications formats such as Ethernet, Frame Relay, ATM, and TDM over IP-optimized core networks. This document describes a framework accomplishing this goal.

3. Architecture of Pseudo Wires

3.1. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALLNOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119. Below are the definitions for the terms used throughout the document.

Packet Switched Network
: A Packet Switched Network (PSN) is a network using IP, MPLS or L2TP as the unit of switching.

Pseudo Wire Emulation Edge to Edge
: Pseudo Wire Emulation Edge to Edge (PWE3) is a mechanism that emulates the essential attributes of a service (such as a T1 leased line or Frame Relay) over a PSN.

Customer Edge
: A Customer Edge (CE) is a device where one end of an emulated service originates and terminates. The CE is not aware that it is using an emulated service rather than a "real" service.

Provider Edge
: A Provider Edge (PE) is a device that provides PWE3 to a CE.

Pseudo Wire
: A Pseudo Wire (PW) is a connection between two PEs carried over a PSN. The PE provides the adaptation between the CE and the PW.

PW End Service
: A Pseudo Wire End Service (PWES) is the interface between a PE and a CE. This can be a

	physical interface like a T1 or Ethernet, or a virtual interface like a VC or VLAN.
Pseudo Wire PDU	A Pseudo Wire PDU is a PDU sent on the PW that contains all of the data and control information necessary to provide the desired service.

PSN Tunnel	A PSN Tunnel is a tunnel inside which multiple PWs can be nested so that they are transparent to core network devices.
Pseudo Wire Domain	A PW Domain (PWD) is a collection of instances of PWs that are within the scope of a single homogenous administrative domain (e.g. PW over MPLS network or PW over IP network etc.).
Path-oriented PW	A Path-oriented PW is a PW for which the network devices of the underlying PSN must maintain state information.
Non-path-oriented PW	A Non-path-oriented PW is a PW for which the network devices of the underlying PSN need not maintain state information.
Interworking	Interworking is used to express interactions between networks, between end systems, or between parts thereof, with the aim of providing a functional entity capable of supporting an end-to-end communication. The interactions required to provide a functional entity rely on functions and on the means to select these functions.
Interworking Function	An Interworking Function (IWF) is a functional entity that facilitates interworking between two dissimilar networks (e.g., ATM & MPLS, ATM & L2TP, etc.). A PE performs the IWF function.
Service Interworking	In Service Interworking, the IWF (Interworking Function) between two dissimilar protocols (e.g., ATM & MPLS, Frame Relay & ATM, ATM & IP, ATM & L2TP, etc.) terminates the protocol used in one network and translates (i.e. maps) its

	Protocol Control Information (PCI) to the PCI of the protocol used in other network for User, Control and Management Plane functions to the extent possible. In general, since not all functions may be supported in one or other of the networks, the translation of PCI may be partial or non-existent. However, this should not result in any loss of user data since the payload is not affected by PCI conversion at the service interworking IWF.
Network Interworking	In Network Interworking, the PCI (Protocol Control Information) of the protocol and the payload information used in two similar networks are transferred transparently by an IWF of the PE across the PSN. Typically the

---

	IWF of the PE encapsulates the information which is transmitted by means of an adaptation function and transfers it transparently to the other network.
Applicability Statement	Each PW service will have an Applicability Statement (AS) that describes the particulars of PWs for that service, as well as the degree of faithfulness to that service.
Outbound	The traffic direction where information from a CE is adapted to a PW, and PW-PDUs are sent into the PSN.
Inbound	The traffic direction where PW-PDUs are received on a PW from the PSN, re-converted back in the emulated service, and sent out to a CE.
CE Signaling	CE (end-to-end) Signaling refers to messages sent and received by the CEs. It may be desirable or even necessary for the PE to participate in or monitor this signaling in order to effectively emulate the service.
PE/PW Signaling	PE/PW Signaling is signaling used by the PEs to set up and tear down the PW. It may be coupled

with CE signaling in order to effectively manage the PW.

PSN Tunnel Signaling PSN Tunnel Signaling is used to set up, maintain and remove the underlying PSN tunnel. An example would be LDP in MPLS for maintaining LSPs. This type of signaling is not within the scope of PWE3.

<Editor's Note: The following figure is temporary. It is intended to facilitate discussion of the preceding set of terms versus those used in [MARTINI].>

[MARTINI]	This Draft
MPLS Network	PSN (includes MPLS)
Tunnel LSP	PSN Tunnel
VC LSP	PW
Edge LSR, R1, R2	PE

Figure 1: Comparison of Terms

3.2. Reference Models

3.2.1. Network Reference Model

Figure 2 below shows the network reference model for PWs.

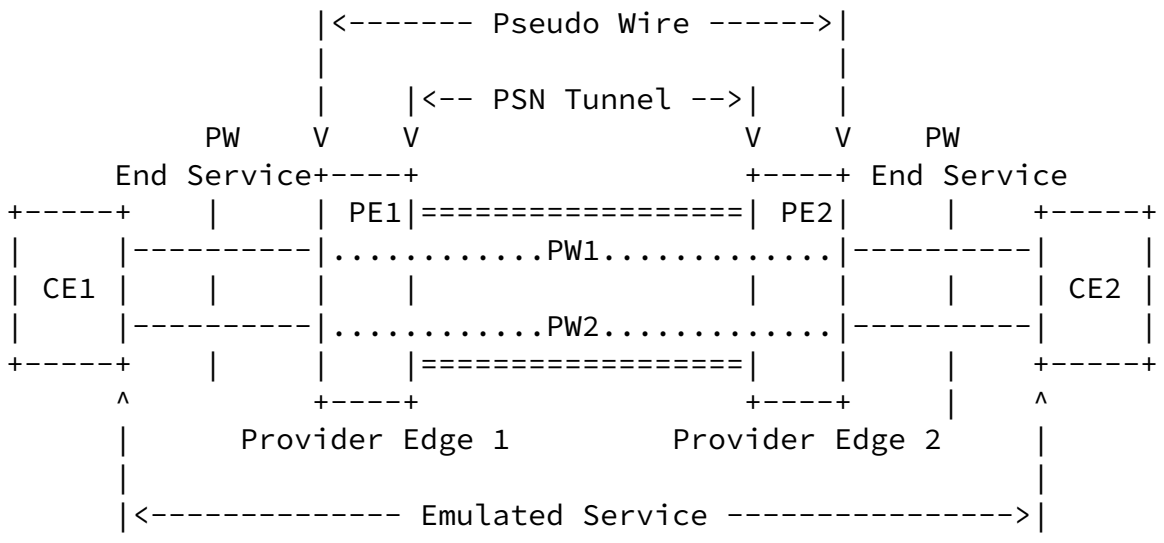


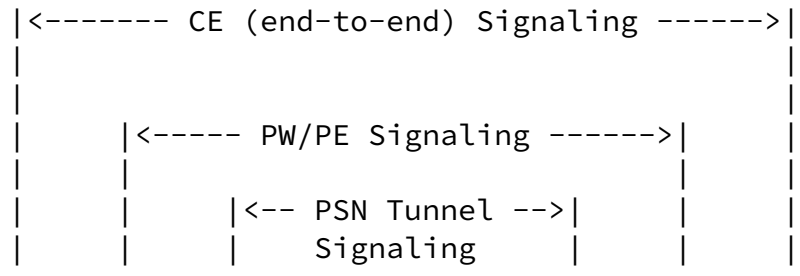
Figure 2: PWE3 Network Reference Model

As shown, the PW provides an emulated service between the customer edges (CEs). Any bits or packets presented at the PW End Service (PWES) are encapsulated in a PW-PDU and carried across the underlying network. The PEs perform the encapsulation, decapsulation, order management, timing and any other functions required by the service. In some cases the PWES can be treated as a virtual interfaces into a further processing (like switching or bridging) of the original service before the physical connection to the CE. Examples include Ethernet bridging, SONET cross-connect, translation of locally-significant identifiers such as VCI/VPI, etc. to other service type, etc.

The underlying PSN is not involved in any of these service-specific operations.

3.2.2. Signaling Reference Model

Figure 3 below shows the signaling reference model for PWs.



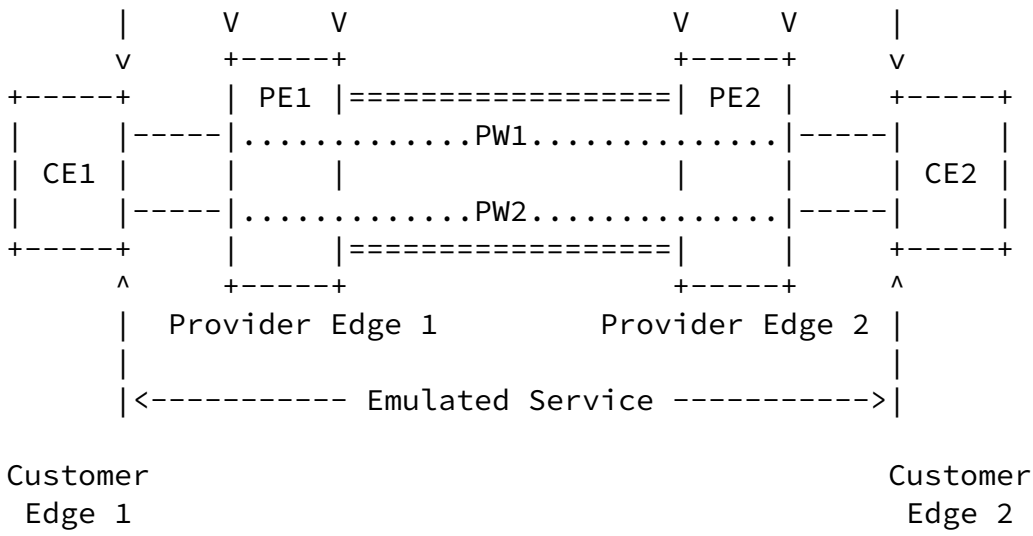


Figure 3: PWE3 Signaling Reference Model

- The CE (end-to-end) signaling is between the CEs. This signaling includes Frame Relay PVC status signaling, ATM SVC signaling, etc.

- The PW/PE signaling is used between the PEs to set up and tear down PWs, including any required coordination of parameters between the two ends.

- The PSN Tunnel signaling controls the underlying PSN. An example would be LDP in MPLS for maintaining LSPs. This type of signaling is not within the scope of PWE3.

3.2.3. Protocol Stack Reference Model

Figure 4 below shows the protocol stack reference model for PWs. The PW provides the CE with what appears to be a connection to its peer

at the far end. Bits or PDUs from the CE are passed through an encapsulation layer.

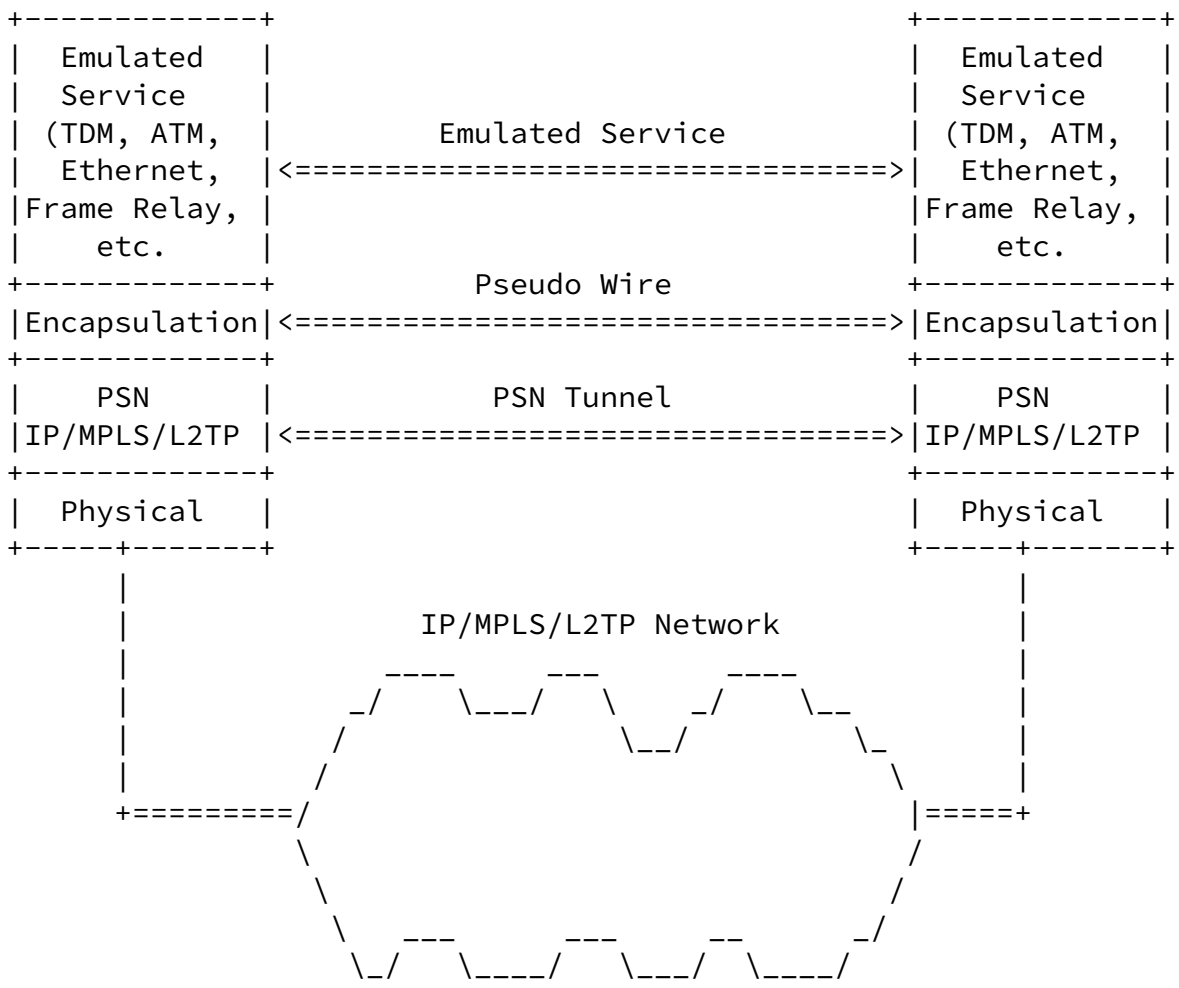


Figure 4: PWE3 Protocol Stack Reference Model

3.3. Architecture Assumptions

1) The current design is focused on a point-to-point and same-to-same service interface at both end of the PW. Only network interworking will be performed at the edge or the PW. Support for service interworking is for further study.

2) The initial design of PWE3 is focused on a single homogenous administrative PWD (e.g. PW over MPLS or PW over IP etc. ONLY). Interworking between different PW types and the support of inter-domain PWs are for further study.

3) The design of PW will not perfectly emulate the characteristics of the native service. It will be dependent on both the emulated service, as well as on the network implementation. An AS shall be

created for each service to describe the degree of faithfulness of a PW to the native service.

4) Only the permanent emulated circuit type (e.g. PVC/PVP) is considered initially. The switched emulated circuit type (e.g. SVC/SVP) will be for further study.

5) The creation and placement of the PSN tunnel to support the PW is not within the scope.

6) The current PW encapsulation approach considerations are focused on IPv4, IPv6, L2TP and MPLS. Other encapsulation approach is for further study.

7) Current PW service applications are focused on Ethernet (i.e. Ethernet II (DIX), 802.3 "raw", Ethernet 802.2, Ethernet SNAP, 802.3ac VLAN), Frame Relay, ATM, TDM (e.g. DS1, DS3, E1, SONET/SDH etc.) and MPLS.

8) Within the single administrative PWD, the design of the PW assumes the inheritance of the security mechanism that has been applied to the emulated services. No PW specific security mechanism will be specified.

3.4. Suitable Applications for PWE3

<Editor's Note: This section will discuss the attributes that make an application suitable (or not) for PWE3 emulation. This section is currently under revision. >

When considering PWs as a means of providing a service, the following questions regarding the application must be considered.

- Preservation of Order - Does the application require in-order delivery of data? Emulation of an application that requires in-order delivery over a PSN that does not guarantee such delivery may be difficult.

- Preservation of Timing - Does the application require fine-grain preservation of timing? If so, the adaptation may be complicated by providing such timing where it is not normally available.

- Natural Delineation - What is the "natural" boundary for delineation of data for encapsulation? (Note: For bit/byte-oriented services, such as TDM emulation, this "natural" delineation may not necessarily be the overriding consideration for determining the best "chunk" for packetizing the service.)

- Packet Size - Are the encapsulated packets variable or fixed in size?

- Data Rate - Is the data rate presented at the interface fixed or variable?

Figure 5 below shows a summary of the applications relevant to PWs, along with a comparison of their attributes.

Attribute -> Application	Preserve Order	Preserve Timing	"Natural" Delineation	Packet Size	Data Rate
T1/E1/T3/E3	yes	yes	125 us frame	fixed	fixed
SONET/SDH	yes	yes	125 us frame	fixed	fixed
Frame Relay	yes	no	frame	variable	variable
ATM AAL1	yes	yes	cell	fixed	fixed
ATM AAL2	yes	yes	cell	fixed	variable
ATM AAL5	yes	no	cell or PDU	variable	variable
Ethernet	yes	no	frame	variable	variable

Figure 5: Summary of Applications and Attributes

4. Layer 1 (Circuit) Applications

For circuit applications the entire bit stream (or at least the payload) needs to be recreated at the far end of the PW. As with ATM CES, the physical layer coding is terminated and re-generated on the far end. In addition, framing may be terminated and regenerated, depending on the application.

4.1. Reference Model

Figure 6 below shows a pair of T1s being carried over a TDM/SONET network. The node marked "M" is an M13 multiplexer, while the nodes marked "S" are SONET Add-Drop Multiplexers (ADMs). Note that the

physical interface of the circuit may change without affecting the circuit. For example, the T1s in Figure 6 below enter the network as physical T1s but exit the network as Virtual Tributaries (VTs) in a physical OC12.

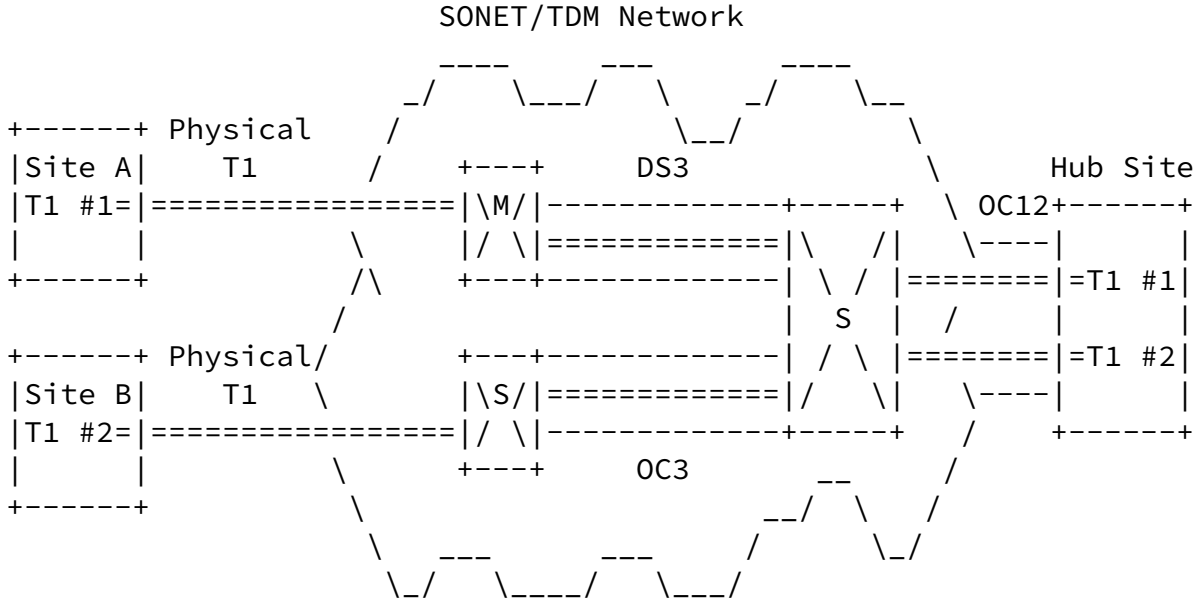
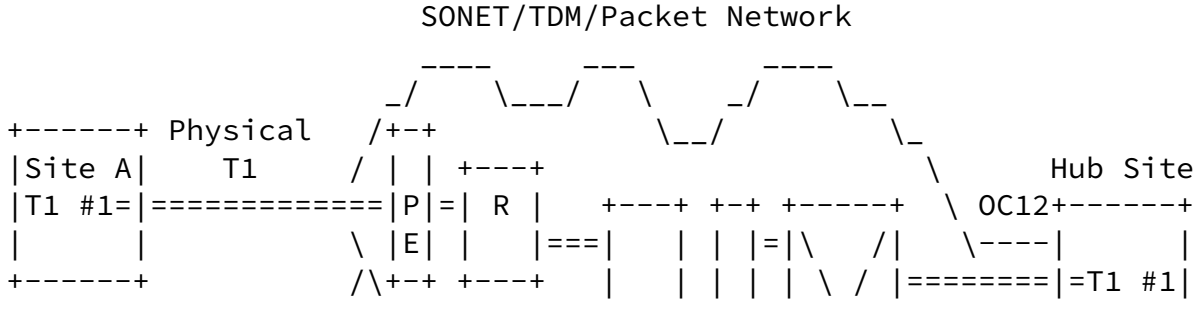


Figure 6: T1/SONET Example Diagram

Figure 7 below shows the same pair of T1s being carried over a packet network. Here the emulation is performed by the PEs marked "PE", and the routers marked "R" carry the resulting packets. Note that the PE, routing and/or SONET functions could be combined in the same device.



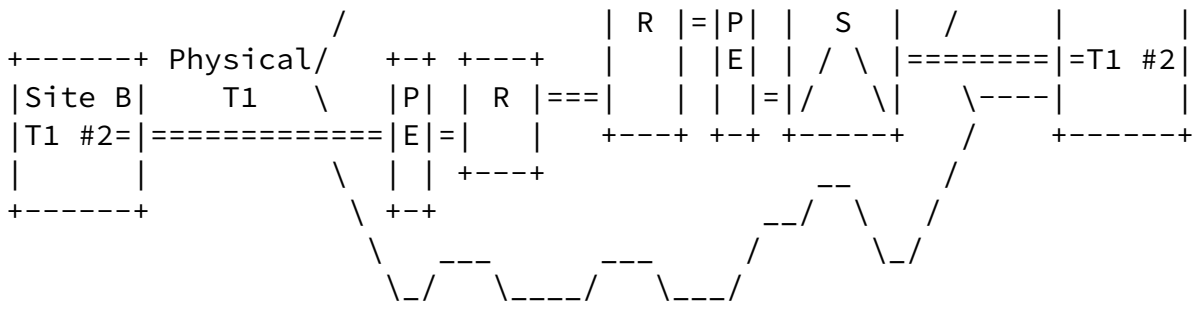


Figure 7: T1 Emulation Example Diagram

4.2. Operational Considerations

4.2.1. Transparency

Circuits such as T1/E1/T3/E3/SONET/SDH lines need a greater degree of transparency than Layer 2 services. These circuits may be carrying the services described in the section on Layer 2 services, but in the Layer 1 scenario the higher layer application is irrelevant and is ignored. In general, these are "bits in, bits out" applications.

In this application a circuit or bit stream is encapsulated in fixed-size frames that are sent at a fixed rate. The emulated stream must be delivered in a reliable and predictable fashion to the far end. Absolute delay and delay variation (also called jitter or wander) must be minimized. Excess delay and delay variation may cause problems with the application carried by the TDM/SONET CEs.

This encapsulation of TDM data must be transparent. The emulated circuit could be carrying one or more types of data (ATM, Frame Relay, TCP/IP, etc.), voice traffic, video or anything else. The data is not interpreted; it is simply transported.

4.2.2. Structured Versus Unstructured Mode For TDM Circuits

As discussed in [ATMCES], emulation of a T1, E1 or other circuit can be done in a structured (framed) mode or in an unstructured (unframed) mode. This same distinction can be applied to higher rate circuits such as DS3, E3, and SONET/SDH.

Unstructured mode generally involves collecting all bits received from a physical port (including transport framing), and packing them into packets for transport through the PSN. The fact that the received bit stream contains a framed signal is more or less irrelevant to the adaptation function.

Structured mode requires the use of a framer to identify and terminate the incoming transport framing, and delineate logical TDM channels within the TDM bit stream for emulation. In addition, TDM framers are generally needed to detect maintenance signals such as Alarm Indication Signal (AIS) and Remote Defect Indication (RDI). Framers are also used to measure various performance parameters such as Errored Seconds, Frame Errored Seconds, etc. Lastly, a framer is needed to generate and terminate the Facility Data Link (FDL) as well as the SONET/SDH Data Communications Channels (DCCs).

The capabilities described in the rest of this section (except for LOS) are predicated on the presence of a framer.

4.2.3. Fractional T1/E1

A fractional T1 or E1 is composed of a number of concatenated DS0s and is sometimes referred to as NxDS0. It may be emulated by replicating the contents of the relevant DS0s at the other end of the tunnel. The value of the other timeslots and/or framing are irrelevant and are not transported in leased line application. Even though the framing is not transported, a framer is still needed to delineate the timeslots for encapsulation.

4.2.4. STS-1/Nc

The SONET/SDH equivalent to Structured T1/E1 services are STS-1/Nc and their SDH equivalents. For STS-1/Nc services a single SONET

---

timeslot or a concatenation of multiple timeslots is used to carry a single logical circuit. As with structured T1/E1 services, the transport framing (i.e. SONET Section and Line Overhead) is terminated, and only the relevant SONET timeslots are carried through the packet network. A single physical SONET interface can be the source of multiple STS-1/Nc services, each of which may be emulated as an independent PWE3 service.

4.2.5. Loopbacks

It could be useful for a PE to process loopback messages as defined in [T1.403]. This would allow for isolation of faults in a network. It also facilitates the certification of equipment for operation in a carrier's network.

There are also inband loopback commands that are used for voice

equipment. These loopback commands are triggered by patterns carried in with the data itself. Voice is limited in the patterns it can present, so it won't falsely mimic the inband loopback command. These inband commands are falling out of favor due to their incompatibility with data services. The inband pattern for the loopback may inadvertently appear in a data stream due to its arbitrary nature. A PE device that implements inband loopbacks must have the capability to disable them.

4.2.6. Performance Processing

[T1.403] defines a Network Performance Report Message (NPRM) that carry periodic reports on the performance of the link. It would be useful for a PE to generate these messages, as they are frequently used for surveillance and trouble-shooting.

4.2.7. LOS/LOF/AIS

Figure 8 shows an example for the generation of AIS and RAI.

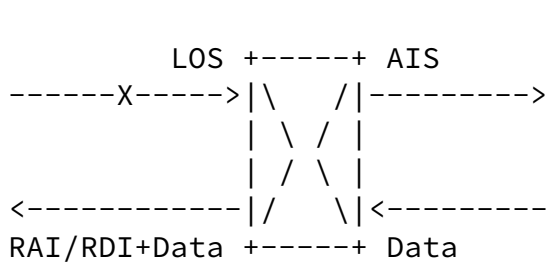


Figure 8: Generation of AIS and RAI/RDI

A TDM multiplexer, SONET ADM, switch or other line terminating equipment (LTE) must respond to an LOS (Loss of Signal), LOF (Loss of Frame) or AIS (Alarm Indication Signal) condition (traditionally known as "red alarms") by generating AIS in the "downstream" direction i.e. the same direction in which the LOS was detected. AIS

is conveyed by sending a certain pattern in the data stream. It may also send RAI (or RDI in SONET) in the "upstream" direction i.e. the opposite direction from that where the LOS was detected. See section 9 of [T1.403] for more information on T1 AIS and RAI. See section section 6.2.1 in [GR253] for more information on the SONET AIS-L and RDI-L signals.

Bandwidth can be saved by suppressing the AIS signal in the emulated

stream and sending instead an indication in the control overhead. This also applies to a received AIS signal. [MALIS] discusses the propagation of AIS using the pointer bits in the TDM control word.

A device emulating TDM circuit must either replicate the AIS indication or indicate this condition in the control overhead.

4.2.8. SONET/SDH STS Unequipped

The "STS Unequipped" indication may be treated in a fashion similar to that for LOS/AIS. As discussed in [MALIS], bandwidth can be saved by suppressing the payload in the emulated stream and sending instead an indication in the control overhead.

4.3. Encapsulations

Encapsulation options for TDM services may be compared on the following criteria.

- Timing - TDM services are very sensitive to timing and timing variations ("jitter"). The encapsulation may need to provide additional information (such as [RTP] timestamps) to help convey timing across the PW.

- Line Signals - The encapsulation should provide a means to convey signals such as AIS and line conditions such as LOF.

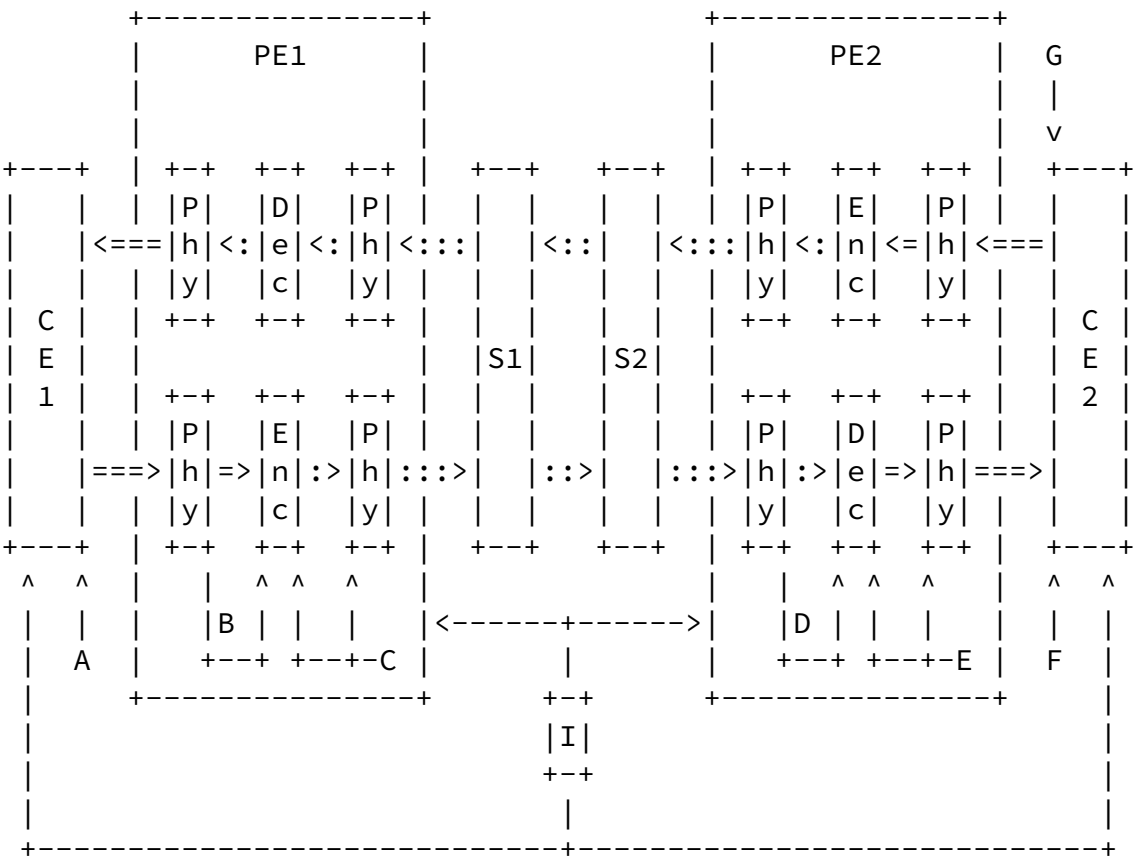
- The encapsulation should minimize overhead.

4.4. Timing

In the recent Ken Burns Jazz television series, it was said of Louis Armstrong that he was very economical with his notes, but that the timing of those notes was everything. The timing of the reconstructed bit stream is similarly important. This section describes the various approaches to this problem. A summary is also provided at the end of the section.

4.4.1. Reference Model

Consider the example network shown in Figure 9. In this network, CE1 and CE2 are connected by a PW provided by PE1, S1, S2 and PE2.



Where:
"CEn" is the TDM edge device
"PEn" is the PE adaptation device
"Sn" is a core switch
"A" - "I" are clocks
"=" is the T1 Bit Stream
":" is the switched connection
"Phy" is a physical interface
"Enc" is a PWE3 encapsulation device
"Dec" is a PWE3 decapsulation device

Figure 9: Timing Recovery Reference Diagram

For this application to work, CE2 needs to be clocked (by clock E) at the same frequency as CE1 (which is being clocked by clock A). A jitter correction buffer at PE2 can handle short-term differences between these two clocks, but over time any absolute difference is going to cause this buffer to overflow or underflow.

Bits are clocked into an encapsulation function in PE1 according to clock B, which is recovered from the incoming data stream. Clocks A

and B will have the same frequency.

The bits are packed into frames and clocked out according to clock C. Clock C could be related to clock B, but in most cases these clocks are completely independent.

The frames arrive at switch S1, and are clocked out according to the internal oscillator on the output interface of switch S1. The frames will depart at the same average rate at which they arrived, but the instantaneous rate will be different on each side of S1. Note that there could be short-term variations due to congestion, but S1 can't experience long-term congestion with respect to the frames carrying emulated data, or the service won't work.

The frames travel on to switch S2, which again forwards them. Note that switch S2 introduces yet another clock, which it uses to transmit the packets toward PE2. Again the average rate is preserved, but the instantaneous rate will vary.

The frames arrive at PE2 and are clocked into the decapsulation function when they arrive (using clock D). Note that clock D will have the same average frequency as A and B, but will have short-term variations. The bits are clocked out of the FIFO according to clock E. Clock F at CE2 is recovered from the bit stream and therefore runs at the same frequency as clock E.

4.4.2. Recreating the Timing

The big question is: where does clock E in Figure 9 come from? There are 5 possibilities, and these are detailed in the following sections.

1) Clock E is derived from an external source such as clock I or B (indirectly via A) at CE1 and G (indirectly via H) at CE2. This method is described in the "External Timing" section below.

2) Clock E could be derived from Clock I and the pointers. This approach is described in the "SONET Pointer Justification" section below.

3) Clock E is derived from the average rate of Clock D. This is the "Adaptive Timing" scenario described in a subsequent section.

4) Clock E is derived from a combination of the local oscillator at PE2 and received SRTS timestamps. The "Differential (SRTS)"

section below describes this approach.

5) Clock E is derived from inband RTP timestamps. This method is discussed in the "RTP" section below.

4.4.2.1. External Timing

The simplest method for communicating timing from one end of a system to the other is an external timing source, such as clock I in Figure

---

9. This external timing source is normally a T1 or E1, but it could be delivered via SONET or a GPS receiver. Its 8 KHz frame rate is extracted and used to directly clock the reconstructed data streams, or as an input to a phase-locked loop to synthesize the desired clock. The drawback to this method is that a common external clock is not commonly available in a data network or in a multi-carrier network.

Note that clock I may actually be two separate clocks of a particular accuracy or stratum. The difference in frequency will eventually cause the FIFO to slip, but if the clock is of a high enough accuracy then the slips will be very infrequent. For example, a stratum 1 clock is accurate to one part in 10**11 [G.811]. This gives a frequency slip rate of 15.4**-6 bit-slip/sec:

```
slip rate = 1.544 Mbps * 10**-11
          = 15.4**-6 bit-slip/sec
```

Taking the reciprocal yields 18 hours/bit-slip:

```
bit slip period = 1 / ( 15.4**-6 bit-slip/sec * 3600 s/h)
                = 18 hours/bit-slip.
```

A typical multiplexer has a buffer that is two frames deep. Assuming that it starts out centered, the expected time for a slip would be almost 5 months:

```
frame slip period = 18 hours/bit-slip * 193 bits/frame
                  = 3474 hours
                  = 145 days
                  = 4.8 months
```

This slip rate could be higher or lower depending on the bit rate, clock accuracy and the depth of the FIFO.

4.4.2.2. SONET Pointer Justification

SONET defines layers of pointers that allow for the multiplexing and transmission of asynchronous signals. These pointers convey the timing of the carried signal with respect to the timing of the encapsulating signal. Each SONET ADM must manipulate these pointers to preserve the timing. This method has the advantage of being well-defined and understood.

One way to apply this method to a packet-based network would be to ensure that all of the links on a given path are synchronous. This would be difficult for Gigabit Ethernet or POS links.

Another way would be for each router to update the pointers as the packet traversed the router. This would be compute intensive.

The method defined in [MALIS] requires pointer manipulation only at the end points. It does require an external clock as a reference for the pointer adjustments.

4.4.2.3. Adaptive Timing

Adaptive timing is used when an external reference is not available. [ATMCES] describes adaptive timing as follows:

> The adaptive technique does not require a network-wide synchronization signal to regenerate the input Service clock at the output IWF. A variety of techniques could be used to implement Adaptive clock recovery. For example, the depth of the reassembly buffer in the output IWF could be monitored:
>
> 1. When the buffer depth is too great or tends to increase with time, the frequency of the Service clock could be increased to cause the buffer to drain more quickly.
>
> 2. When the buffer contains fewer than the configured number of bits, the Service clock could be slowed to cause the buffer to drain less quickly. Wander may be introduced by the Adaptive clock recovery technique if there is a low-frequency component to the Cell Delay Variation inserted by the ATM network carrying cells from the input to output IWF.

> Careful design is required to make adaptive timing work well. The instantaneous buffer depth must be filtered to extract the average frequency and to reject the jitter and wander.

Adaptive timing is ideal for many network applications where there is no external timing reference available (needed for SRTS), and where the packet rate is decoupled from the line rate (as in a routed network). Adaptive timing may not meet the requirements of [G.823], [G.824] and other similar specifications.

4.4.2.4. Differential (SRTS)

[ATMCES] describes the SRTS (Synchronous Residual Time Stamp) method:

> The SRTS technique measures the Service Interface input clock frequency against a network-wide synchronization signal that must be present in the IWF, and sends difference signals, called Residual Time Stamps, in the AAL1 header to the reassembly IWF. At the output IWF, the differences can be combined with the network-wide synchronization signal to re-create the input Service Interface clock.

The requirement for a network-wide signal is reasonable in a Telco or SONET environment, where such clocks are commonly available. It may be problematic in a packet network.

Here is the correspondence between the clocks in Figure 9 and.[I.363.1].

Description	I.363.1	Figure 9
Service Clock	Fs	A, B
Network Clock	Fn	C
Derived Reference	Fnx	Based on C

4.4.2.5. RTP

[RTP] uses an absolute timestamp to play out the bits at the same rate that they were received and packetized. RTCP (RTP Control Protocol) provides a means to synchronize the transmit and receive clocks.

4.4.2.5.1. RTP Timestamps

Section 4 of [RTP] defines a timestamp that is either 32-bits or 64-bits in size:

> Wallclock time (absolute time) is represented using the timestamp format of the Network Time Protocol (NTP), which is in seconds relative to 0h UTC on 1 January 1900 [NTP]. The full resolution NTP timestamp is a 64-bit unsigned fixed-point number with the integer part in the first 32 bits and the fractional part in the last 32 bits. In some fields where a more compact representation is appropriate, only the middle 32 bits are used; that is, the low 16 bits of the integer part and the high 16 bits of the fractional part. The high 16 bits of the integer part must be determined independently.

A 32-bit absolute time stamp with a 16-bit fractional part would give a 15 us granularity (= 1/65535), which is too coarse for circuit emulation. This means that the 64-bit timestamp must be used, with a granularity of 23 ns.

The transmit timestamps are created according to clocks B and C at PE1 and interpreted according to clocks D and E at PE2. These two oscillators will vary by some amount, even if they are very accurate. This drift means that RTCP, NTP or some other means must be used to synchronize the clocks at each end.

4.4.3. Summary of Timing Recovery Methods

All of the previously described methods for timing recovery can be made to work for Layer 1 circuit services. How then can we compare

---

them? Here are some criteria:

- Is an external timing source required? This might be a direct timing source as described in "External Timing", or it could be an indirect source as with SRTS.

- Must the PE synthesize clocks? Synthesis of clocks generally requires a Phase-Locked Loop (PLL) to create one clock from another.

- Is the method provably correct? Some methods such as external

timing and SRTS can be proven to meet specifications. The performance of others, such as adaptive timing, is more dependent on particular implementations.

Figure 10 below shows a summary of the methods for timing recovery.

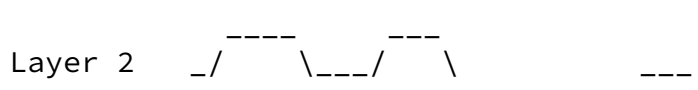
Method-> / Parameter	Ext. Source	SONET Ptr	SRTS	RTP/RTCP	Adaptive
External timing required?	yes	yes	yes	no	no
Clock synthesis?	no	yes	yes	yes	yes
Provably correct?	yes	yes	yes	?	?

Figure 10: Summary of Timing Recovery Methods

5. Layer 2 (Packet/Cell) Applications

5.1. Layer 2 PW Reference Model

Figure 11 below shows the reference model for Layer 2 PWs. The Layer 2 emulated protocols/services include ATM VCC, ATM VPC, Frame Relay DLCI, IEEE 802.1Q VLAN, IEEE 802.3x link, etc. The nodes marked "S" are protocol-specific switches e.g. Frame Relay switches.



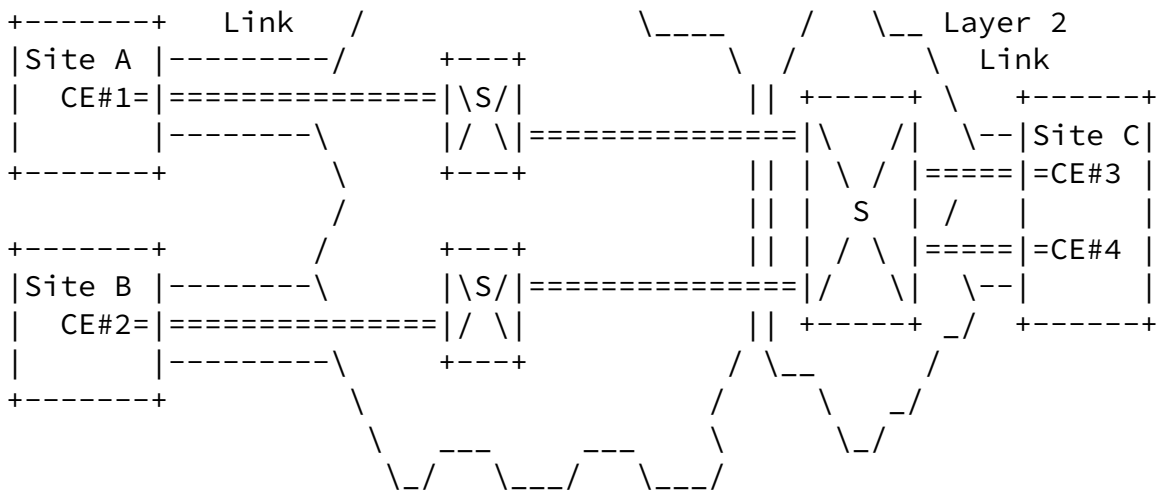


Figure 11: Layer 2 Interconnect Reference Model

Figure 12 below shows the reference model for PW emulation of Layer 2 services.

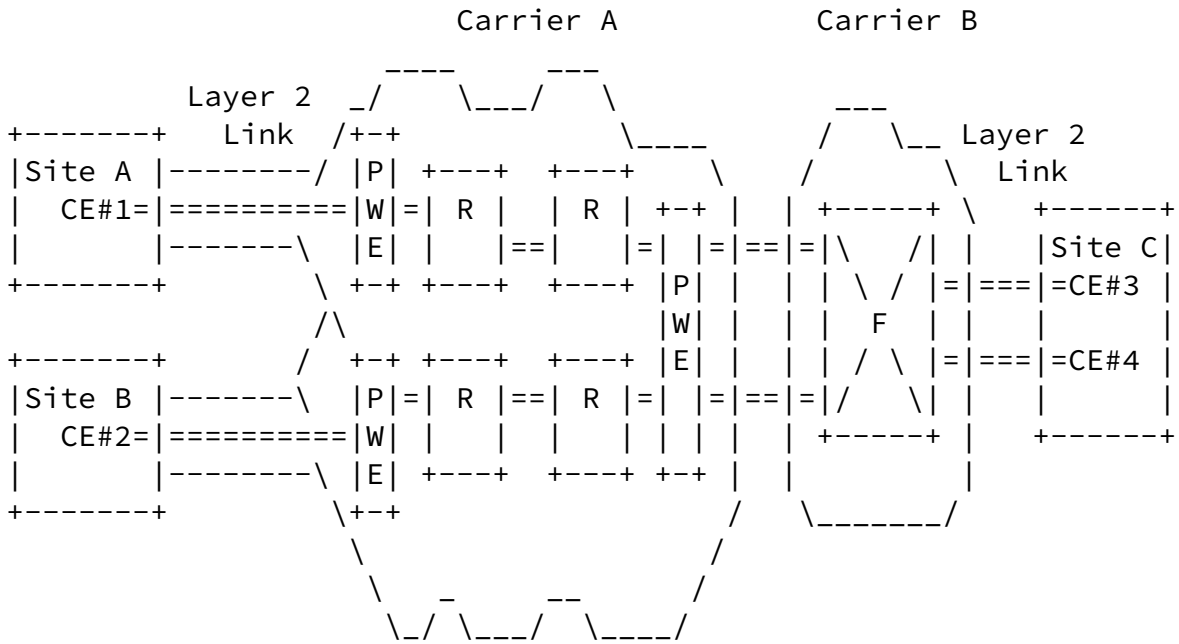


Figure 12: Layer 2 PW Emulation

5.2. Ethernet

5.2.1. Reference Model Scope

PW carriage of Ethernet operates as point-to-point trunking in a non-shared medium. The Ethernet interface can operate in a half-duplex or full-duplex mode. Control functions such as IEEE 802.3 Carrier Sense Multiple Access with Collision Detection (CSMA/CD)[802.3] and IEEE 802.1D Spanning Tree [802.1D] are not applicable nor within the scope of PWs. However, the PW shall conform to the service

definitions as defined in IEEE 802.1P,Q [802.1Q], as required. Also, it shall support all Ethernet framing i.e. Ethernet Frame and IEEE 802.3x including IEEE 802.3ac VLAN Tagging [802.3] as well as Jumbo Frame.

5.2.2. Operational Considerations

5.2.2.1. Operational Modes

The design of the Ethernet PW must consider the support of the two operational modes in this framework. Both modes shall be supported for all Ethernet interfaces, i.e. from 10 Mbps to 10,000 Mbps, and the design of the Ethernet PW functions shall be agnostic to the Ethernet's link capacity. Both modes shall transparently support the address resolution protocols, i.e. ARP, InverseARP, proxy ARP and Ethernet-based control protocol (e.g. Generic Attribute Registration Protocol (GARP), GARP VLAN Registration Protocol (GVRP) etc.).

- Opaque Trunking - In this mode, the ingress PE shall relay all of the traffic from an Ethernet port into the PW.

- Transparent Trunking - This mode is particularly designed for support of Virtual LANs (VLAN). VLAN types include Port-based VLANs, MAC-Address-Based VLANs, IP-Based VLANs, 802.1Q Tag-based VLANs and 802.10 Security-based VLANs.

  The ingress PE may pay attention to the MAC header and other relevant VLAN classification information based on the configuration policy. The Ethernet PW shall carry multiple VLANs traffic and can extend VLANs across the PWD. In the case when 802.1Q Tag-based VLAN is configured, if the received frame is tagged with a NULL VLAN_ID, it will be associated with the VLAN equal to the Port's default VLAN. At frame transmission, all frames that are associated with 802.1Q Tag-based VLAN shall be tagged except for those assigned for the default VLAN.

  The PE may provide translation of the VLAN_ID in order to facilitate deployment. Note that this does not increase the VLAN_ID space, so it has no effect on scalability.

5.2.2.2. Quality Of Service Support Considerations

The Ethernet AS shall describe the faithfulness of the PW with respect to these attributes described in IEEE 802.1p [802.1Q].

- Service Availability - Service availability is measured as ratio between times when MAC service is unavailable and when it is

available.

- Frame Loss - The MAC service does not provide guaranteed delivery of service data units. However, the Ethernet PW system should consider monitoring frame loss.

- Frame Misorder - The MAC service does not permit reordering frames within the same user-priority for a source and destination address pair.

- Frame Duplication - The MAC service does not permit duplicating frames.

- Transit Delay Performance - IEEE 802.1p [802.1Q] defines frame transit delay is the elapsed time between an MA_UNITDATA.request and corresponding MA_UNITDATA.indication on a successful transfer.

- Undetected Frame Error Rate - By using the Frame Checksum (FCS) calculation for each frame, the undetected frame error rate should be low.

- Maximum Service Data Unit Size - The maximum service data unit size is dependent on the access media used. In general, it is the lowest common denominator of the two adjacent Ethernet interface.

- Priority Tags and Traffic Classes - IEEE 802.1p defines eight traffic classes (priority). The PRI bits on the VLAN fields should be carried transparently over the PW. COS differentiation on the PW level based on the received 802.1p bits is possible but is out-of-scope.

5.3. Frame Relay

5.3.1. Reference Model

Frame Relay service offerings often have a different physical format and speed at each end of the link. For example, a hub and spoke deployment of Frame Relay might provide fractional T1 access at the spokes and a clear channel T3 to the hub. The Virtual Circuits (VCs) are aggregated by switches in the Frame Relay network. This is shown in figure 13 below, where the Frame Relay switches are marked with an "F".

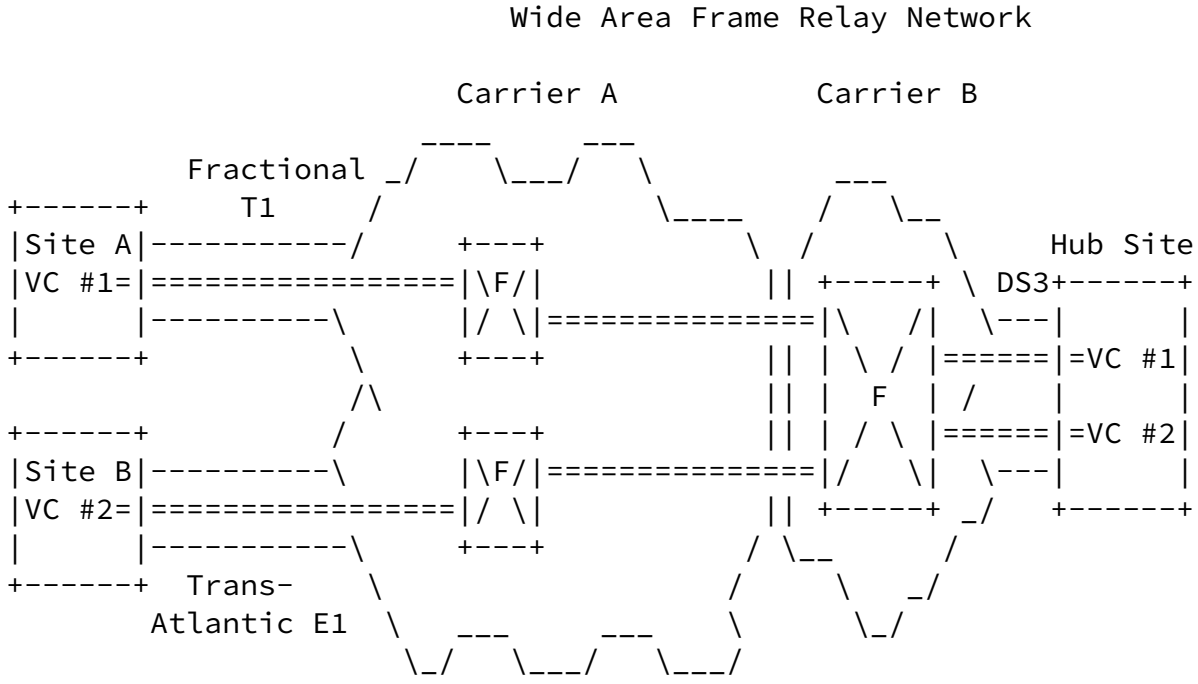
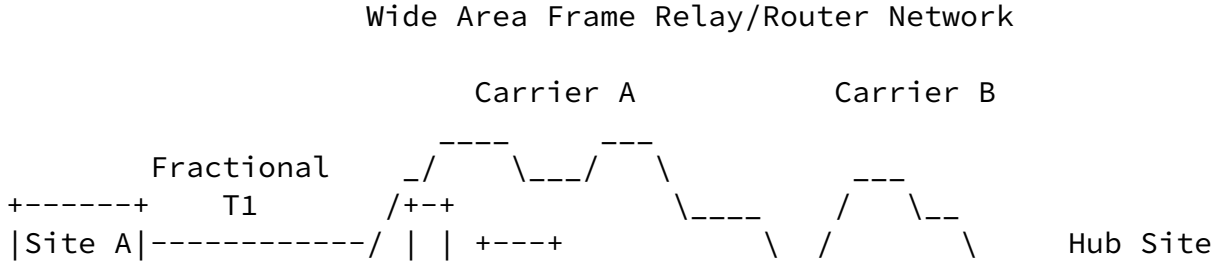


Figure 13: Frame Relay Example Model

Figure 14 shows an emulated network, where Carrier "A" is providing a transparent Frame Relay emulation connection to Carrier "B".



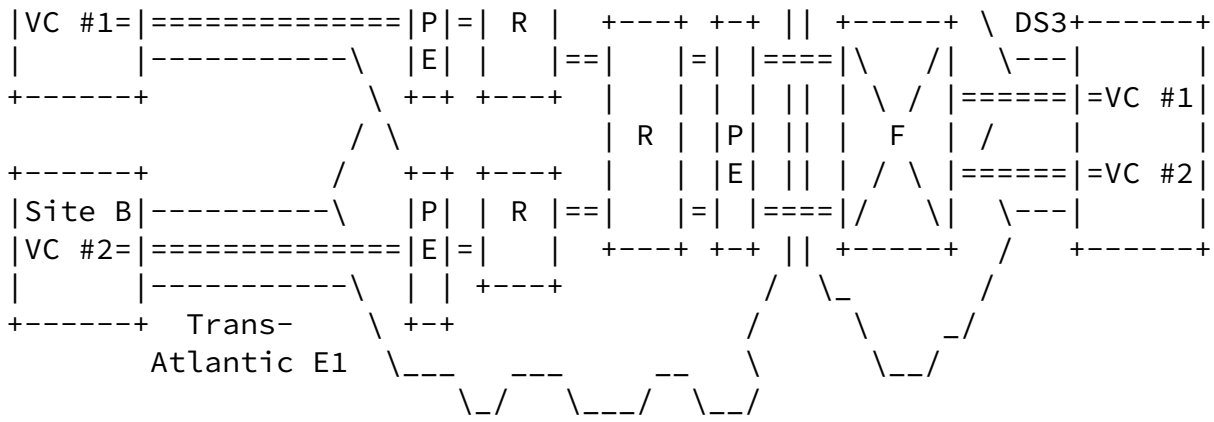


Figure 14: Frame Relay Emulation Example Diagram For Transparent Emulation

Here the emulation is performed by the PEs marked "PE", and the resulting packets are carried by the routers marked "R". In this case, the emulated VCs can transparently carry the PVC status signaling (if any) and need not perform any higher layer function. Also, note that the emulation and routing functions could be combined in the same device.

If the Frame Relay switches are to be completely eliminated (as shown in figure 15 below), then the emulation service must implement Frame Relay PVC status signaling and/or connection signaling for SVCs. As previously noted, the PE and routing functions could be combined in the same device.

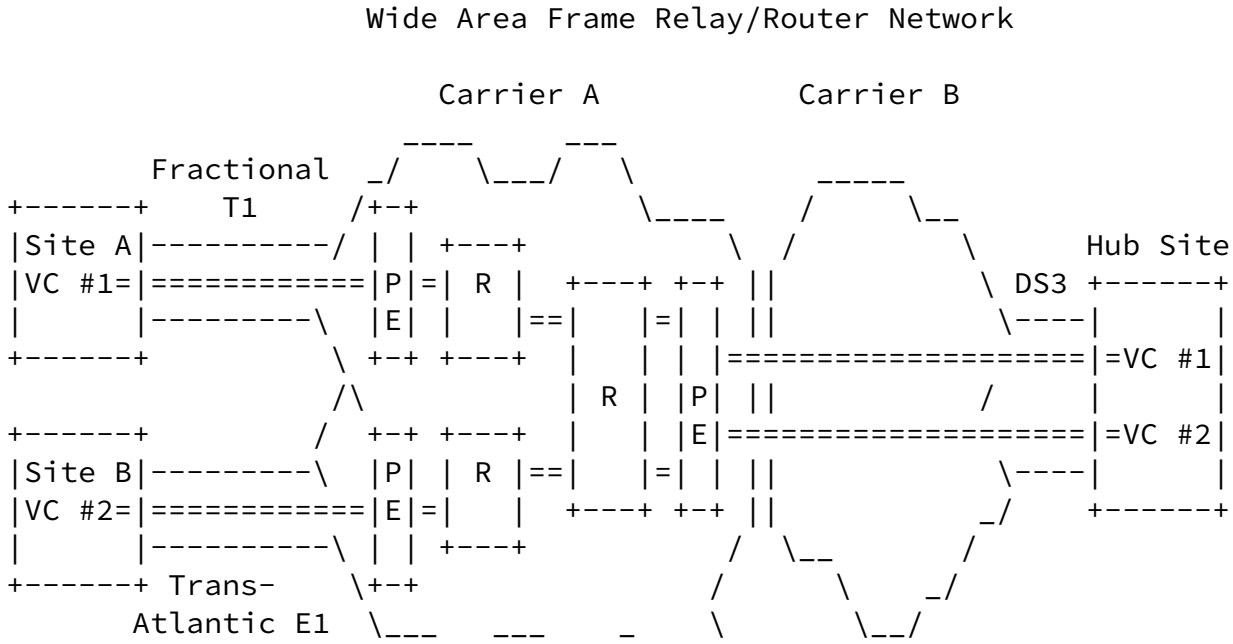


Figure 15: Frame Relay Emulation Example Diagram For Non-Transparent Emulation

5.3.2. Operational Considerations

Frame Relay provides a connection-oriented circuit-based carriage of variable-sized frames. There are two types of virtual circuits supported in Frame Relay: Permanent Virtual Circuits (PVCs) and Switched Virtual Circuits (SVCs). The following sections describe the considerations to support the operation of Frame Relay over the PW.

5.3.2.1. Frame Sequence

The PW must deliver frames in the proper sequence.

5.3.2.2. Frame Size

In general, the maximum frame size for Frame Relay is 1600 bytes per [FRF.1.2]. This can be made larger in some implementations. If the MTU of the PW is less than (1600 bytes - size of PW headers), a fragmentation and reassembly mechanism may be needed.

5.3.2.3. End-to-End Characteristics

[FRF.5] and [FRF.13] define a set of traffic parameters to support Service Level Agreements (SLAs). The design of the PW may be

---

required to preserve these end-to-end transport characteristics.

5.3.2.4. Connection Management and Congestion Control

Each Frame Relay header contains some connection management information, including

- a command/response (CR) bit

- a discard eligibility (DE) bit

- a connection ID (DLCI)

- an address extension indicator (EA)

- Forward/Backward Congestion Notification (FECN/BECN). Figure 16 shows an example of how BECN and FECN are sent.

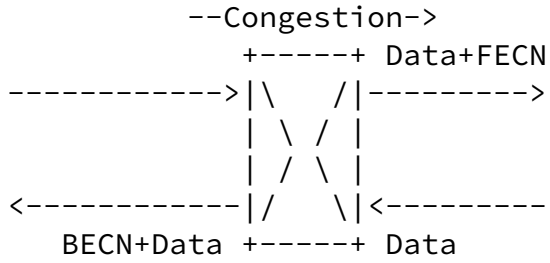


Figure 16: Generation of BECN and FECN

All of this information is vital to the integrity of the Frame Relay circuit. The Frame Relay PW AS must define a means to preserve such information across the PWD.

5.3.2.5. Link Management Support

Frame Delay defines a set of link management functions for PVC Status Management as specified in [T1.617D] and [Q.933A]. Link Management runs on a dedicated PVC; therefore, its operation does not impact actual user data. The management functions include:

- a heartbeat exchange that verifies that the link is operational

- a report regarding the status of one or more individual DLCIs

For some networks, such as the one shown in Figure 15, this link management is the only means to verify the end-to-end integrity of the Frame Relay virtual circuit. The PE may required to emulate such functions. These functions will be transparent to the underlying network.

Another important consideration is that there should be some coordination between the PW's link status and the associated Frame Relay VCs. For example, it might be necessary to tear down the VCs

---

in the presence of a network fault.

5.3.2.6. DLCI Association

There are two scopes of DLCI addressing that have been defined by ANSI and ITU-T: Local and Global DLCIs.

- Local DLCI addressing means that DLCI numbers are only significant at one end of a Frame Relay virtual circuit.

- Global DLCI addressing is an extension to PVC status management that allows a DLCI number to have universal significance. A global DLCI identifies the same VC at both ends of the network.

In the case when the DLCI is locally significant, the management of the PWD must provide a mechanism to coordinate the DLCIs at the two ends of the PW. The association can be done via signaling or configuration.

5.3.2.7. Multiplexing VCs over PWs

To preserve PW tunnel space and to enhance scalability of PWs, it would be very valuable to allow one or more VCs to be multiplexed onto the same PW. One scenario might be to associate an entire Frame Relay logical interface to a PW. Another possibility is that the assignment of VCs to the PW could be done via signaling or management. In either of these cases, the DLCI for each frame would need to be preserved across the PW.

If such multiplexing approach is used, the earlier discussion related to the packet sequencing, end-to-end characteristics, SLA preservation and link status management, shall be addressed with the same considerations.

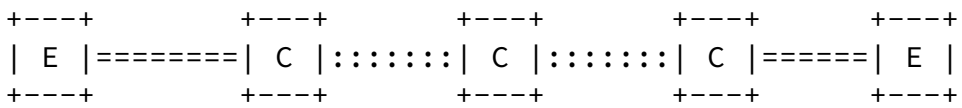
5.3.2.8. Signaling Transparency

Since Frame Relay supports SVCs, the PE may need to support signaling interworking at the PWES. InverseARP frames should be passed on without interpretation. In either case, these frames shall be transparent to the underlying PSN.

5.3.2.9. Soft PVC Support

One type of connection service that is provided by Frame Relay networks is called a Soft PVC (SPVC). A SPVC may be considered to be composed of three parts: two peer-to-peer PVCs at each side of the core, and a SVC between them.

Figure 17 shows the SPVC interconnection example.



Where:
"E" is the edge switch
"C" is the core switch
"=" is the permanent connection
":" is the switched connection

Figure 17: Example of an SPVC Over a Frame Relay Network

The creation of the SPVC within the core is triggered by detecting the existence of the PVCs at the edges. This detection is done either by network management or by some proprietary signaling.

Now consider the case where the core switches are replaced by PEs as shown in Figure 14. The SVC connection within the core is replaced by the PW. The PVCs configuration which are maintained by the ingress and the egress CEs of the PWD should remain the same. The ingress and egress PEs shall maintain the SPVC behavior such that it is transparent to the CEs.

5.4. ATM

5.4.1. Reference Model

As far as PWs are concerned, ATM is very similar to Frame Relay. We will use the same reference network (Figure 13) for ATM as we did for Frame Relay. Of course, the Frame Relay switches would be ATM switches instead. Likewise, the PE networks shown in Figures 14 and 15 are applicable to ATM.

5.4.2. Operational Considerations

Like Frame Relay, ATM provides connection-oriented circuit-based carriage of fixed-size cells. There are two types of virtual circuit supported in ATM: PVC and SVC. In addition to virtual circuit connections (VCCs), ATM also supports virtual path connections (VPCs). There are also permanent virtual paths (PVPs) and switched virtual paths (SVPs).

ATM carries data in fixed size (53 byte) frames called "cells". Higher layer frames are adapted to these fixed size cells via ATM Adaptation Layers (e.g. AAL1, AAL2 and AAL5) and SAR. Different types of AALs have different cell header formats, and the cells may contain signaling information.

The following sections describe the set of considerations for PW support of ATM.

5.4.2.1. Cell Sequence

The PW must deliver cells in the proper sequence.

5.4.2.2. End-to-end Transport Characteristics

The ITU-T [I.356] and the ATM Forum [TM4.0] each define a set of traffic and QoS parameters. The AS for ATM PWs should specify how the PW will maintain the end-to-end characteristics for such VCs.

5.4.2.3. ATM SLA

The ITU-T [I.371] defines performance targets for managing ATM traffic and congestion control. These targets may be used by some service providers to define their ATM SLAs. The AS for ATM PWs should specify how SLA transparency will be achieved.

5.4.2.4. Connection Management and Congestion Control

The ATM header contains some connection management and congestion control information, as defined in [I.371]:

- Cell Loss priority (CLP)

- Connection Identifier (VPI/VCI)

- Payload Type Identifier (PTI) - distinguishes between an OAM (Operations, Administration and Management) cell and a user cell

- Explicit Forward Congestion Indication (EFCI)

All of this information is vital to the integrity of the ATM circuit. The ATM PW AS must define a means to preserve such information across the PWD.

5.4.2.5. OAM Support

ATM OAM functions are defined in the ITU-T standard [I.610]. OAM cells are used to provide functions like fault management, performance management and continuity checks.

OAM is implemented differently in VCCs and VPCs. In the case of a

VCC, the OAM cell is sent along the same VC as the user cells. For a VPC, the OAM cell is sent over a dedicated VC within the VPC. OAM flows are also classified as end-to-end flows (covering the entire virtual connection) or as segment flows (covering only parts of the virtual connection).

The PE may emulate the end-to-end OAM flows by encapsulating the OAM cells in a PW-PDU. A PE that supports the OAM function should support coordination between the OAM behavior between the PE peers. For example, an OAM AIS cell at one end can result in PW signaling

that causes the PW link to go down at the far end. If the PE does support OAM, the emulation of the OAM function shall be transparent to the underlying network.

5.4.2.6. ILMI Support

The Integrated Local Management Interface [ILMI] protocol facilitates network deployment and management in several ways:

- ILMI allows an ATM node to determine the various features supported by its neighbors (such as type of signaling, size of connection space etc).

- ILMI allows for smoother administration of ATM addresses through address registration.

- ILMI facilitates automatic configuration of private interfaces.

- ILMI supports procedures for detecting loss of connectivity through periodic polling.

For some networks, such as the one shown in Figure 15, ILMI is the only means to verify the end-to-end integrity of the ATM virtual circuit. It may be desirable for the PE to emulate such functions. If supported, these functions must be transparent to the underlying network.

5.4.2.7. VC Associations

Each ATM connection identifier has local significance only. Local significance means that each ATM connection identifier (VPI and/or VCI) is only significant at a local ATM interface, and is independent from the identifier at the other end of the link. The management of the PWD must provide a mechanism to coordinate the identifiers at the

two ends of the PW. The association can be done via signaling or configuration.

5.4.2.8. Multiplexing ATM VCs over PWs

See the discussion in the "Multiplexing VCs over PWs" sub-section of the previous "Frame Relay" section of this document.

5.4.2.9. ATM Signaling Transparency

See the discussion in the "Signaling Transparency" sub-section of the previous "Frame Relay" section of this document.

5.4.2.10. Soft PVC Support

See the discussion and figures in the "Soft PVC Support" sub-section of the previous "Frame Relay" section of this document.

5.4.2.11. Segmentation and Reassembly (SAR)

The bandwidth overhead of the ATM cell is about 10% (= 5 overhead bytes out of 53 bytes total). For AAL5 traffic [I.363.5], it may be more efficient in terms of bandwidth to carry the re-assembled AAL5 frames instead of the individual ATM cells. This would involve some cost in terms of a SAR operation at each end of the PW. In some cases, especially if OAM is required to be supported over the PW, the PW may have no choice but to transport the ATM traffic in cell format.

Whether the ATM traffic is transported in a frame or cell format, it is the responsibility of the PE to emulate the OAM functions to the adjacent ATM interface at each end.

6. PW Maintenance

6.1. PW-PDU Validation

It is a common practice to use a checksum, CRC or FCS to assure end-to-end integrity of frames. The PW service-specific mechanisms must define whether the packet's checksum shall be preserved across the PWD or be removed at the ingress PE and then be re-calculated at the egress PE. The former approach saves work, while the later saves bandwidth.

For protocols like ATM and Frame Relay, the checksum is only applicable to a single link. This is because the circuit identifiers (e.g. Frame Relay DLCI or ATM VPI/VCI) have only local significance and are changed on each hop or span. If the circuit identifier (and thus checksum) is going to change as a part of the PW emulation, it would be more efficient to strip and re-calculate the checksum.

Other PDU headers (e.g. UDP in IP) do not change during transit. It would make sense to preserve these types of checksums.

The AS for each protocol must describe the validation scheme to be used.

6.2. PW-PDU Sequencing

One major consideration of PW design is to ensure in-sequence delivery of packets, if needed. The design of the PW for each protocol must consider the support of the PSN for in-order delivery as well as the requirements of the particular application. For example, MPLS supports connection-oriented transport with a guarantee of in-order delivery. A sequence number in the PW layer is not needed when used with MPLS. IP is connectionless and does not guarantee in-order delivery. When using IP, a PW sequence number may be needed for some applications (such as TDM).

6.3. Session Multiplexing

One way to facilitate scaling is to increase the number of PWs per underlying tunnel. There are two ways to achieve this:

- For a service like Relay or ATM, all of the VCs on a given port could be lumped together. VCs would not be distinguishable within the PWD.

- Service SDUs could be distinguished within a PW-PDU by port, channel or VC identifiers. This approach would allow for switching or grooming in the PWD.

6.4. Security

Each AS must specify a means to protect the control of the PWE and the PE/PW signaling. The security-related protection of the encapsulated content of the PW is outside of scope.

6.5. Encapsulation Control

6.5.1. Scalability

Different service types may be required between CEs, Support of multiple services implies that a range of PWD label spaces may be needed. If the PWD spans a PSN supporting traffic engineering, then the ability to supporting label stacking would be desirable.

6.5.2. Service Integration

It may be desirable to design a PW to transport a variety of services which have different transport characteristics. To achieve this integration it may be useful to allow the service requirements to be mapped to the tunneling label in such a way that the PWD can apply the appropriate service and transport management to the PW.

6.6. Statistics

The PE can tabulate statistics that help monitor the state of the network, and to help with measurement of SLAs. Typical counters include:

- Counts of PW-PDUs sent and received, with and without errors.

- Counts of PW-PDUs lost (TDM only).

- Counts of service PDUs sent and received, with and without errors (non-TDM).

- Service-specific interface counts.

These counters would be contained in a MIB, they should not replicate existing MIB counters.

6.7. Traceroute

Tracing functionality is desirable for emulated circuits and services, because it allows verification and remediation of the operation and configuration of the forwarding plane. [BONICA]

describes the requirements for a generic route tracing application. Applicability of these requirements to PWE3 is an interesting problem, as many of the emulated services have no equivalent function. In general, there is not a way to trace the forwarding plane of an TDM or Frame Relay PVC. ATM does provide an option in the loopback OAM cell to return each intermediate hop (see [I.610]).

There needs to be a mechanism through which upper layers can ask emulated services to reveal their internal forwarding details. A common mechanism for all emulated services over a particular PSN may be possible. For example, if MPLS is the PSN, the path for a VC LSP could be revealed via the signaling from the underlying TE tunnel LSP, or perhaps via the proposed MPLS OAM. However, when we are trying to trace the entire emulated service, starting from the CE (e.g. an ATM VCC), then a uniform approach probably will not work and different approaches would be required for different emulated services.

6.8. Congestion Considerations

[RFC2914] describes how devices connected to the Internet should handle congestion. The discussion of congestion with respect to PWE3 will be broken into two sections: CBR applications and VBR applications.

6.8.1. VBR Applications

VBR applications include Ethernet, Frame Relay, and ATM (other than AAL CES). During periods of congestion the PE may be able to take action to communicate to the CE the need to slow down.

6.8.1.1. Frame Relay

In the presence of congestion, the PE could perform several actions. These are the same actions that a Frame Relay switch might take. If available, a measure of the degree of congestion would be useful.

- While a service provide may define an SLA for a Frame Relay Service, Frame Relay itself does not have a guarantee of delivery. Given this fact, the PE could do nothing in the face of congestion. The Frame Relay application at the CE would then have to detect congestion and act appropriately.

- Frame Relay defines BECN as an indication to a Frame Relay device that traffic that it sent is experiencing congestion. See Figure

16 for an example of how BECN is sent. For mild congestion, the PE could send BECN back to the CE. The CE could then reduce the amount of traffic being sent. It is worth nothing that many Frame Relay devices ignore BECN.

- The CE could also send FECN in the direction in which congestion is occurring. See Figure 16 for an example of how FECN is sent.

- During congestion, the PE could discard all frames with DE set

- If the PE was aware of the CIR for the VCs, it could drop any traffic in excess of CIR.

- For severe congestion the PE could take the interface down. If the PE was generating the PVC status signaling then these messages could be used to convey the problem to the CE. This approach is not elegant and may not work well with existing Frame Relay applications.

6.8.1.2. ATM

ATM has a forward notification of congestion (EFCI), but unlike Frame Relay there is no backwards notification. This leaves the following choices of action. These are the same actions that an ATM switch might take.

- Do nothing and let the ATM application at the CE handle the problem. This may work for some applications, but it will make it difficult for service providers to guarantee a high QoS on the VC.

- If the PE was aware of the traffic parameters for the VCs, it could drop any traffic that was out of profile.

- For severe congestion the PE could take the interface down. This may be worse than doing nothing.

6.8.1.3. Ethernet

A PE providing a PW to an Ethernet CE could react to congestion in one of the following ways.

- A PE could use Ethernet flow control during congestion by sending a PAUSE frame as described in Annex 31B of [802.3].

- A PE could do nothing and let the Ethernet application at the CE handle the problem.

- For severe congestion the PE could take the interface down. This may be worse than doing nothing.

6.8.2. CBR Applications

CBR applications include layer 1 applications such as emulated TDM/SONET streams, as well as layer 2 applications such as ATM AAL1 CES. These applications present a constant load on the network at all times. They cannot slow down; they are either running at full speed, or they are impaired. If congestion causes an excessive number of packets to be lost, the PE could take down the interface and send AIS to the CE. There is probably not much point in doing this if the PE is operating in an unstructured mode, as the framer in the CE will probably declare a LOS condition anyway. A PE operating in a structured (framed) mode would always send a clean frame pattern to the CE, so it might be desirable to send AIS to notify the CE that there are problems with the PW.

7. Packet Switched Networks

This section discusses various types of PSNs for providing the PW transport. The areas of considerations are:

- Explicit Multi-protocol Encapsulation Identifier
- Transport Integrity
- Traffic Engineering Ability
- Session Multiplexing
- Flow and Congestion Control
- Packet Ordering
- Tunnel Maintenance
- Scalability
- Overhead
- QoS and Traffic Management

7.1. IP

Below is a description of the aspects of the Internet Protocol [IP].

Explicit MP Encap ID	No support for a full range of multi-service protocols e.g. there is no protocol type assigned for ATM or MPLS.
Transport Integrity	IP has a checksum over the header but not over the payload.

Traffic Engineering	The TOS bits may be used as DiffServ code points.
Session Multiplexing	No support for session multiplexing.
Packet Order	No support for preservation of order.
Tunnel Maintenance	Protocols such as L2TP may be used to establish tunnels using IP packets.
Flow/Congestion Control	No built-in flow control to manage congestion. IP relies on the upper layer protocol, e.g. TCP, to perform the congestion management.
Scalability	It would be hard to imagine a protocol more scalable than IP.
Transport Overhead	Minimum of 20-byte header.
QoS/Traffic Management	No built-in QoS and traffic management. However, one can apply DiffServ to select a per-hop behavior for a class of traffic.

7.2. L2TP

Layer 2 Tunneling (L2TP) [L2TP] provides a virtual extension of PPP across an IP PSN.

Explicit MP Encap ID	Supports any routed protocol, e.g. IP, IPX and AppleTalk that is supported by PPP.
Transport Integrity	Support a checksum for the entire encapsulated frame.
Traffic Engineering	No companion traffic engineering mechanism

	to support L2TP tunnel establishment.
Session Multiplexing	Supports two levels of session multiplexing via the use of the "tunnel-id" and "session-id" fields.
Packet Order	By supporting the optional sequence number, packet re-ordering can be done at the PWE
Tunnel Maintenance	L2TP uses control messages to establish, terminate and monitor the status of the logical PPP sessions. These are independent of the data messages. L2TP also provides an optional keep-alive mechanism to detect non-operational tunnel.

Flow/Congestion Control	L2TP defines flow and congestion control mechanisms for the control traffic only; no control for the data traffic. Even so, the PE could apply value-added functions such as admission control, policing and shaping of the L2TP tunnel at the aggregate flows level, e.g. DiffServ-TE.
Scalability	Lack of label stacking ability.
Transport Overhead	Minimum overhead is 44-byte (20-byte IP header + 12-byte UDP header + 8-byte minimum L2TP header + 4-byte PPP header) to support L2TP encapsulation
QoS/Traffic Management	No built-in QoS and traffic management. However, one can apply IntServ or DiffServ to select the preferred transport behavior for the entire tunnel, i.e. one traffic class per L2TP tunnel.

7.3. MPLS

Multi-protocol Label Switching [MPLS] is designed to combine Layer 2 switching and Layer 3 routing technology to provide efficient packet processing and forwarding over a variety of link layer and transport technologies e.g. ATM, Frame Relay and SONET.

Explicit MP Encap ID	No defined standard to identify the encapsulated multi-protocol PDU.
Transport Integrity	No checksum support.
Traffic Engineering	Designed with many signaling, routing and traffic management extensions to support traffic engineering.
Session Multiplexing	Supports session multiplexing via the MPLS label and the EXP field.
Packet Order	Connection-oriented transport with guaranteed in-sequence delivery.
Tunnel Maintenance	MPLS signaling provides the ability to establish, terminate and monitor the status of the LSP.
Flow and Congestion Control	MPLS-TE assumes external admission control, policy and shaping mechanism to provide flow and congestion control at the aggregate traffic level.

Scalability	Label stacking facilitates scalability.
Transport Overhead	Minimum overhead is 4-byte + any MPLS extension to support multi-protocol encapsulation and transport.
QoS/Traffic Management	MPLS-TE supports QoS and traffic management.

8. Acknowledgments

This document was created by the PWE3 Framework design team.

9. References

9.1. IETF RFCs

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

[RTP] H. Schulzrinne et al, "RTP: A Transport Protocol for Real-Time Applications", RFC1889, January 1996.

[NTP] D. Mills, "Network Time Protocol Version 3", RFC1305, March 1992.

[MPLS] E. Rosen, "Multiprotocol Label Switching Architecture", RFC3031, January 2001.

[IP] DARPA, "Internet Protocol", RFC791, September 1981.

9.2. IETF Drafts

[ANAVI] Anavi et al, "TDM over IP" draft-anavi-tdmoip-01.txt, work in progress, February 2001.

[MALIS] Malis et al, "SONET/SDH Circuit Emulation Service Over MPLS (CEM) Encapsulation" (draft-malis-sonet-ces-mpls-03.txt), work in progress, February 2001.

[XIAO] Xiao et al, "Requirements for Pseudo Wire Emulation Edge-to-Edge (PWE3)" (draft-pwe3-requirements-01.txt), work in progress, July 2001.

[MARTINI] Martini et al, "Transport of Layer 2 Frames Over MPLS" (draft-martini-l2circuit-trans-mpls-06.txt), work in progress, May 2001.

[BONICA] Bonica et al, "Tracing Requirements for Generic Tunnels" (draft-bonica-tunneltrace-01.txt), work in progress, February 2001.

[CALLON] Callon et al, "A Framework for Provider Provisioned Virtual Private Networks" (draft-ietf-ppvpn-framework-00.txt), work in progress, February 2001.

9.3. ATM Forum

[ATMCES] ATM Forum, "Circuit Emulation Service Interoperability Specification Version 2.0" (af-vtoa-0078-000), January 1997.

[TM4.0] ATM Forum, "Traffic Management Specification Version 4.0", (af-tm-0056.000), April, 1996.

[ILMI] ATM Forum, "Integrated Local Management Interface (ILMI) Specification Version 4.0", (af-ilmi-0065.000), September, 1996.

9.4. Frame Relay Forum

[FRF.1.2] D. Sinicrope, "PVC User-to-Network Interface (UNI) Implementation Agreement", Frame Relay Forum FRF.1.2, July 2000.

[FRF.5] O'Leary et al, "Frame Relay/ATM PVC Network Interworking Implementation Agreement", Frame Relay Forum FRF.5, December 20, 1994.

[FRF.13] K. Rehbehn, "Service Level Definitions Implementation Agreement", Frame Relay Forum FRF.13, August 4, 1998.

9.5. ITU

[Q.933A] ITU, "ISDN Signaling Specifications for Frame Mode Switched and Permanent Virtual Connections Control and Status Monitoring" ITU Recommendation Q.933, Annex A, Geneva, 1995.

[I.356] ITU, "B-ISDN ATM Layer Cell Transfer Performance", ITU Recommendation I.356, To Be Published.

[I.363.1] ITU, "B-ISDN ATM Adaptation Layer specification: Type 1 AAL", Recommendation I.363.1, August, 1996.

[I.363.2] ITU, "B-ISDN ATM Adaptation Layer (AAL) type 2 specification", Recommendation I.363.2, To Be Published.

[I.363.5] ITU, "B-ISDN ATM Adaptation Layer specification: Type 5 AAL", Recommendation I.363.5, August, 1996.

[I.371] ITU, "Traffic Control and Congestion Control in B-ISDN" ITU Recommendation I.371, To Be Published.

[I.610] ITU, "B-ISDN Operation and Maintenance Principles and Functions", ITU Recommendation I.610, February, 1999.

[G.811] ITU, "Timing Characteristics of Primary Reference Clocks", ITU Recommendation G.811, September 1997.

[G.823] "The Control of Jitter and Wander Within Digital Networks

Which Are Based on the 2048 kbit/s Hierarchy", ITU Recommendation G.823, March 2000.

[G.824] "The Control of Jitter and Wander Within Digital Networks Which Are Based on the 1544 kbit/s Hierarchy", ITU Recommendation G.824, March 2000.

9.6. 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.

9.7. ANSI

[T1.403] ANSI, "Network and Customer Installation Interfaces - DS1 Electrical Interfaces", T1.403-1999, May 24, 1999.

[T1.617D] ANSI, "Digital Subscriber System No. 1 DSS1 Signaling Specification for Frame Relay Bearer Service", ANSI T1.617-1991 (R1997), Annex D.

9.8. Telcordia

[GR253] Telcordia, "Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria" (GR253CORE), Issue 3, September 2000.

10. Security Considerations

It may be desirable to define methods for ensuring security during exchange of encapsulation control information at an administrative boundary of the PSN.

11. Authors' Addresses

Prayson Pate
Overture Networks
P. O. Box 14864
RTP, NC, USA 27709
Email: prayson.pate@overturenetworks.com

XiPeng Xiao
Photuris, Inc.
2025 Stierlin Court
Mountain View, CA 94043
Email: xxiao@photuris.com

Tricci So
Caspian Networks
170 Baytech Dr.
San Jose, CA 95134
E-Mail: tso@caspiannetworks.com

Craig White
Level 3 Communications, LLC.
1025 Eldorado Blvd.
Broomfield, CO, 80021
e-mail: Craig.White@Level3.com

Kireeti Kompella
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
Email: kireeti@juniper.net

Andrew G. Malis
Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134
Email: Andy.Malis@vivacenetworks.com

Thomas K. Johnson
Litchfield Communications
76 Westbury Park Rd.
Watertown, CT 06795
Email: tom_johnson@litchfieldcomm.com

12. Full Copyright Section