Network Working Group
Internet Draft

M Bocci
Alcatel

S.Bryant
Cisco Systems

Expires: July 2006

January 11, 2006

An Architecture for Multi-Segment Pseudo Wire Emulation Edge-to-Edge

draft-ietf-pwe3-ms-pw-arch-00.txt

Status of this Memo

Abstract

This document describes an architecture for extending pseudo wire emulation across multiple packet switched network segments. Scenarios are discussed where each segment of a given edge-to-edge emulated service spans a different provider's PSN, and where the emulated service originates and terminates on the same providers PSN, but may pass through several PSN tunnel segments in that PSN. It presents an architectural framework for such multi-segment pseudo wires, defines terminology, and specifies the various protocol elements and their functions.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

Table of Contents

1. Introduction

RFC 3985 [2] defines the architecture for pseudo wires, where a pseudo wire (PW) both originates and terminates on the edge of the same packet switched network (PSN). The PW passes through a maximum of one PSN tunnel between the originating and terminating PEs.

This document extends the architecture in RFC 3985 to enable pseudo wires to be extended through multiple PSN tunnels. Use cases for multi-segment pseudo wires, and the consequent requirements, are defined in [3].

1.1. Motivation

PWE3 aims to provide point-to-point connectivity between two edges of a provider network. Requirements for Multi-Segment Pseudo-Wires for this are specified in [3]. These requirements address three main problems:

- How to constrain the density of the mesh of PSN tunnels when the number of PEs grows to many hundreds or thousands, while minimizing the complexity of the PEs and P routers.

- How to provide PWE3 across multiple PSN routing domains or areas in the same provider.

- How to provide PWE3 across multiple provider domains, and different PSN types.

Consider a single PWE3 domain, such as that shown in Figure 1. There are 4 PEs, and PWE3 must be provided from any PE to any other PE. Traditionally, this would be achieved by establishing a full mesh of PSN tunnels between the PEs. This would also require a full mesh of LDP signaling adjacencies between the PEs. Pseudo wires could then be established between any PE and any other PE via a single, direct tunnel. PEs must terminate all pseudo wires that are carried on PSN tunnels that terminate on that PE according to the architecture of RFC 3985. This solution is adequate for small numbers of PEs, but the number of PEs and signaling adjacencies will grow in proportion to the square of the number of PEs.

A more efficient solution for large numbers of PEs would be to support a partial mesh of PSN tunnels between the PEs, as shown in Figure 1. For example, consider a PWE3 service whose endpoints are PE1 and PE4. Pseudo wires for this can take the path PE1->PE2->PE3, and rather than terminating at PE2, be switched between ingress and egress PSN tunnels on that PE. This requires a capability in PE2 that can concatenate PW segments PE1-PE2 to PW segments PE2-PE3. The end-to-end PW is known as a multi-segment PW.

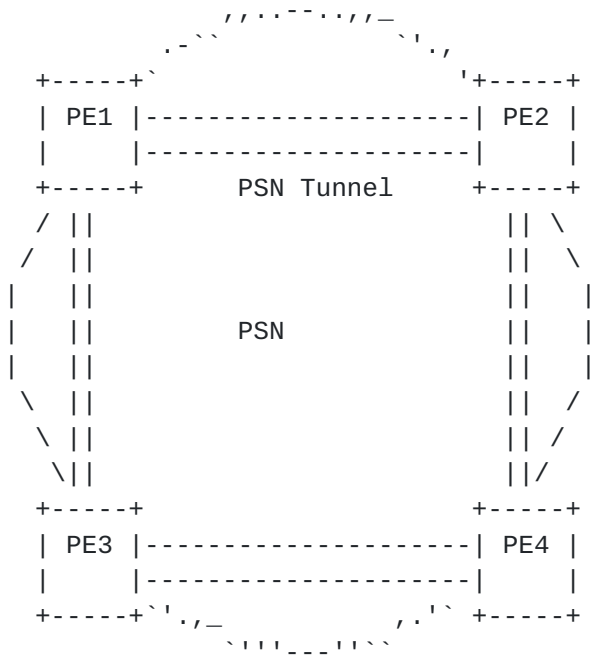

Figure 1 Single PSN PWE3 Scaling

Figure 1 shows a simple flat PSN topology. However, large provider networks are typically not flat, consisting of many domains that are connected together to provide edge-to-edge services. The elements in each domain are specialized for a particular role.

An example application is shown in Figure 2. Here, the providers network is divided into three domains: Two access domains and the core domain. The access domains represent the edge of the provider's network at which services are delivered. In the access domain, simplicity is required in order to minimize the cost of the network. The core domain must support all of the aggregated services from the access domains, and the design requirements here are for scalability, performance, and information hiding (i.e. minimal state). The core must not be exposed to the state associated with large numbers of individual edge-to-edge flows. That is, the core must be simple and fast.

In a traditional layer 2 network, the interconnection points between the domains are where services in the access domains are aggregated for transport across the core to other access domains. In an IP network, the interconnection points would also represent interworking points between different types of IP networks e.g. those with MPLS and those without, and also points where network policies can be applied.

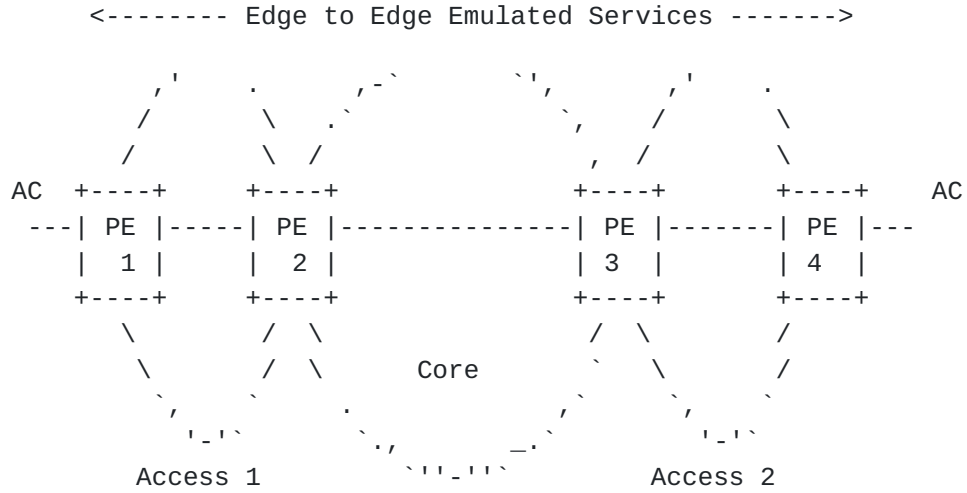

Figure 2 Multi-Domain Network Model

This model can also be applied to inter-provider services, where they also rely on a number of separate provider networks to be connected together.

Consider the application of this model to PWE3. PWE3 uses tunneling mechanisms such as MPLS to enable the underlying IP PSN to emulate characteristics of the native service. One solution to the multi-domain network model above is to extend PSN tunnels edge-to-edge between all of the PEs in access domain 1 and all of the PEs in access domain 2, but this requires a large number of PSN tunnels as

described above, and also exposes the access and the core of the network to undesirable complexity. An alternative is to constrain the complexity to the network domain interconnection points (PE2 and PE3 in the example above). Pseudo-wires between PE1 and PE4 would then be switched between PSN tunnels at the interconnection points, enabling PWs from many PEs in the access domains to be aggregated across only a few PSN tunnels in the core of the network. PEs in the access domains would only need to maintain direct signaling sessions, and PSN tunnels, with other PEs in their own domain, thus minimizing complexity of the access domains.

1.2. Non-Goals of this Document

The following are non-goals for this document:

- The on-the-wire specification of PW encapsulations
- The detailed specification of mechanisms for establishing and maintaining multi-segment pseudo-wires.

1.3. Terminology

The terminology specified in RFC 3985 applies. In addition, we define the following terms:

- PW Terminating Provider Edge (T-PE). A PE where the customer-facing attachment circuits (ACs) are bound to a PW forwarder. A Terminating PE is present in the first and last segments of a MS-PW. This incorporates the functionality of a PE as defined in RFC 3985.
- Single-Segment Pseudo Wire (SS-PW). A PW setup directly between two T-PE devices. Each PW in one direction of a SS-PW traverses one PSN tunnel that connects the two T-PEs.
- Multi-Segment Pseudo Wire (MS-PW). A static or dynamically configured set of two or more contiguous PW segments that behave and function as a single point-to-point PW. Each end of a MS-PW by definition MUST terminate on a T-PE.
- PW Segment. A part of a single-segment or multi-segment PW, which is set up between two PE devices, T-PEs and/or S-PEs.

o PW Switching Provider Edge (S-PE). A PE capable of switching the control and data planes of the preceding and succeeding PW segments in a MS-PW. The S-PE terminates the PSN tunnels of the preceding and succeeding segments of the MS-PW. It is therefore a PW switching point for a MS-PW. A PW Switching Point is never the S-PE and the T-PE for the same MS-PW. A PW switching point runs necessary protocols to setup and manage PW segments with other PW switching points and terminating PEs.

2. Applicability

A MS-PW is a single PW that for technical or administrative reasons is segmented into a number of concatenated hops. From the perspective of a T-PE, a MS-PW is indistinguishable from a SS-PW. Thus, the following are equivalent from the perspective of the T-PE

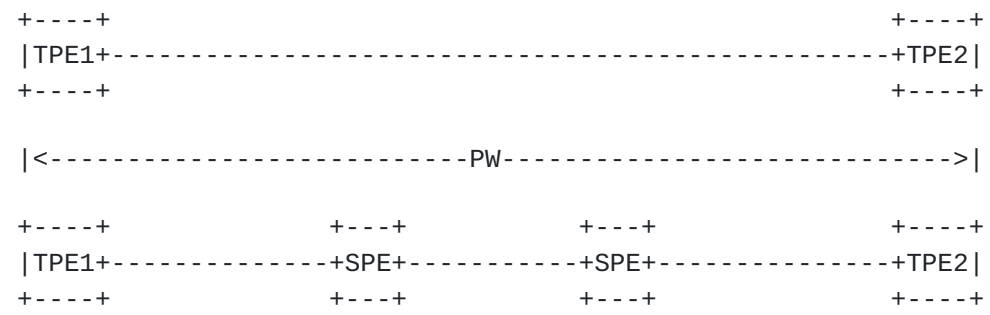

Figure 3 MS-PW Equivalence

Although a MS-PW may require services such as node discovery and path signaling to construct the PW, it should not be confused with a L2VPN system, which also requires these services. A VPWS connects its endpoints via a set of PWs. MS-PW is a mechanism that abstracts the construction of complex PWs from the construction of a L2VPN. Thus a T-PE might be an edge device optimized for simplicity and an S-PE might be an aggregation device designed to absorb the complexity of continuing the PW across the core of one or more service provider networks to another T-PE located at the edge of the network.

3. Protocol Layering model

The protocol-layering model specified in RFC 3985 applies to multi-segment PWE3 with the following clarification: the pseudo-wires may be considered to be a separate layer to the PSN tunnel. That is, they are independent of the PSN tunnel routing, operations, signaling and maintenance. The design of PW routing domains should not imply that the underlying PSN routing domains are the same. However, MS-PW will

reuse the protocols of the PSN and may use information that is extracted from the PSN e.g. reachability.

3.1. Domain of Multi-Segment PWE3

PWE3 defines the Encapsulation Layer, i.e. the method of carrying various payload types, and the interface to the PW Demultiplexer Layer. It is expected that other layers will provide the following:

. PSN tunnel setup, maintenance and routing

. T-PE discovery

It is assumed that any node that is reachable via a PSN tunnel from an S-PE or T-PE is a PE, a subset of which may be capable of behaving as an S-PE. The selection of which S-PEs to use to reach a T-PE is considered to be within the domain of PWE3.

3.2. Payload Types

Multi-segment PWE3 is applicable to all PWE3 payload types. Encapsulations defined for SS-PWs are also used for MS-PW without change. If different segments run over different PSN types, the encapsulation may change but the S-PE must not need an NSP. It is recommended that a list of compatible PWE3 encapsulations that do not need an NSP be published. Translations between segments must not require processing of the underlying payload.

4. Multi-Segment PWE3 Reference Model

The PWE3 reference architecture for the single segment case is shown in [2]. This architecture applies to the case where a PSN tunnel extends between two edges of a single PSN domain to transport a PW with endpoints at these edges.

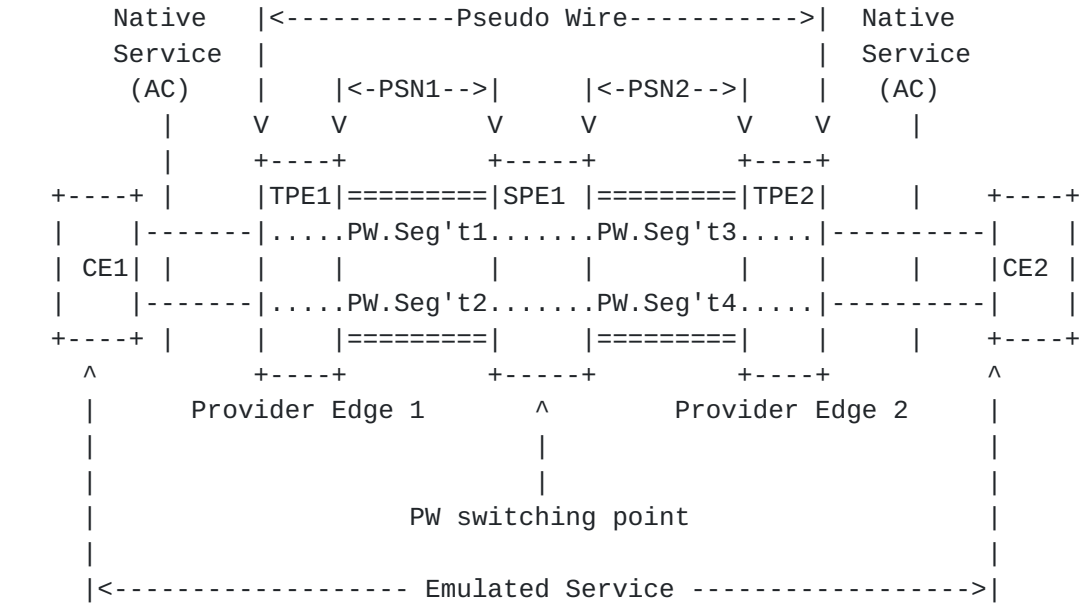

Figure 4 PW switching Reference Model

Figure 4 extends this architecture to show a multi-segment case. The PEs that provide PWE3 to CE1 and CE2 are Terminating-PE1 (T-PE1) and Terminating-PE2 (T-PE2) respectively. A PSN tunnel extends from T-PE1 to switching-PE1 (S-PE1) across PSN1, and a second PSN tunnel extends from S-PE1 to S-PE2 across PSN2. PWs are used to connect the attachment circuits (ACs) attached to PE1 to the corresponding ACs attached to PE3. Each PW segment on the tunnel across PSN1 is switched to a PW segment in the tunnel across PSN2 at S-PE1 to complete the multi-segment PW (MS-PW) between T-PE1 and T-PE2. S-PE1 is therefore the PW switching point. PW segment 1 and PW segment 3 are segments of the same MS-PW while PW segment 2 and PW segment 4 are segments of another MS-PW. PW segments of the same MS-PW (e.g., PW1 and PW3) MAY be of the same PW type or different type, and PSN tunnels (e.g., PSN1 and PSN2) can be the same or different technology. This document requires support for MS-PWs with segments of the same type. An S-PE switches an MS-PW from one segment to another based on the PW identifiers (e.g., PW label in case of MPLS PWs).

Note that although Figure 4 only shows a single S-PE, a PW may transit more one S-PE along its path. This architecture is applicable when the S-PEs are statically chosen, or when they are chosen using a dynamic path selection mechanism.

4.1. Intra-Provider Architecture

There is a requirement to deploy PWs edge to edge in large service provider networks [3]. Such networks typically encompass hundreds or thousands of aggregation devices at the edge, each of which would be a PE. These networks may be partitioned into separate metro and core PWE3 domains, where the PEs are interconnected by a sparse mesh of tunnels.

Whether or not the network is partitioned into separate PWE3 domains, there is a also a requirement to support a partial mesh of traffic engineered PSN tunnels.

The architecture shown in Figure 4 can be used to support such cases. PSN1 and PSN2 may be in different administrative domains or access, core or metro regions within the same providers network. Alternatively, TPE1, SPE1 and TPE2 may reside at the edges of the same PSN.

4.1.1. Intra-Provider Switching Using ACs

In this model, the PW reverts to the native service AC at the PE. This AC is then connected to a separate PW on the same PE. In this case, the reference models of RFC 3985 apply to each segment and to the PEs. The remaining PE architectural considerations in this document do not apply to this case.

4.1.2. Intra-Provider Switching Using PWs

In this model, PW segments are switched between PSN tunnels that span portions of a provider's network, without reverting to the native service at the boundary. For example, in Figure 4, PSN 1 and PSN 2 would be portions of the same provider's network.

4.2. Inter-Provider Architecture

Intra-provider PWs may need to be switched between PSN tunnels at the provider boundary in order to minimize the number of tunnels required to provide PWE3 services to CEs attached to each providers network. In addition, AAA and security and mechanisms may need to be implemented on a per-PW basis at the provider boundary.

4.2.1. Inter-Provider Switching Using ACs.

In this model, the PW reverts to the native service at the provider boundary PE. This AC is then connected to a separate PW at the peer provider boundary PE. In this case, the reference models of RFC 3985 apply to each segment and to the PEs. The remaining PE architectural considerations in this document do not apply to this case.

4.2.2. Inter-Provider Switching Using PWs.

In this model, PW segments are switched between PSN tunnels in each provider's network, without reverting to the native service at the boundary. For example, in Figure 4, PSN 1 and PSN 2 would be different provider's networks. However, this would require that S-PE1 be a member of both provider networks.

An alternative network architecture is shown in Figure 5. Here, S-PE1 and S-PE2 are provider border routers. PW segment 1 is switched to PW segment 2 at S-PE1. PW segment 2 is then carried across an inter-provider PSN tunnel to S-PE2, where it is switched to PW segment 3 in PSN 2.

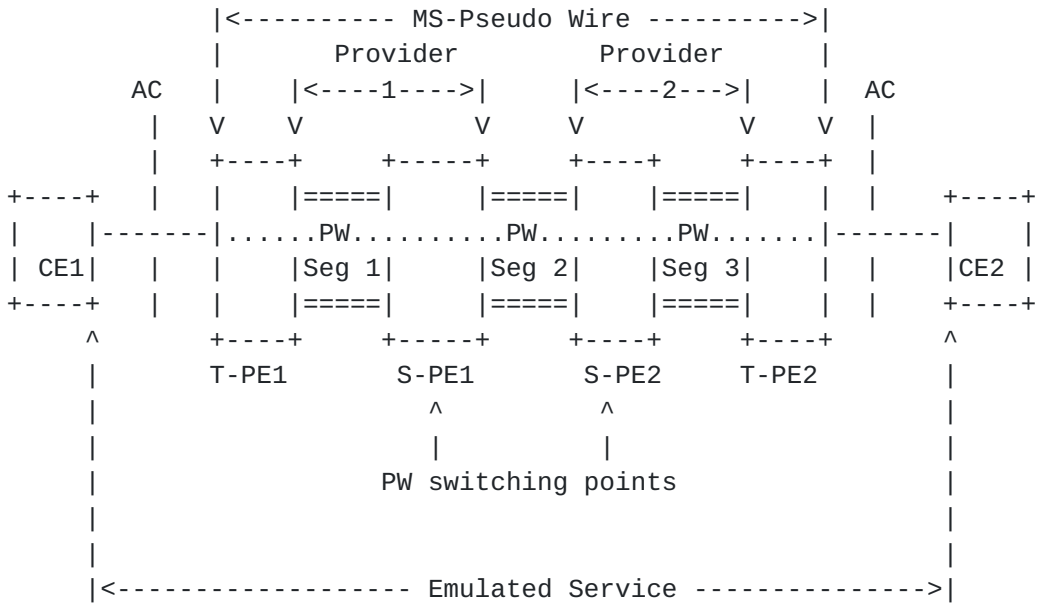

Figure 5 Inter-Provider Reference Model

5. PE Reference Model

5.1. PWE3 Pre-processing

PWE3 preprocessing is applied in the T-PEs as specified in RFC 3985. Processing at the S-PEs is specified in the following sections.

5.1.1. Forwarding

Each forwarder in the S-PE forwards packets from one PW segment on the ingress PSN facing interface of the S-PE to one PW segment on the egress PSN facing interface of the S-PE.

The forwarder selects the egress segment PW based on the ingress PW label. The mapping of ingress to egress PW label may be statically or dynamically configured. Figure 6 shows how a single forwarder is associated with each PW segment at the S-PE.

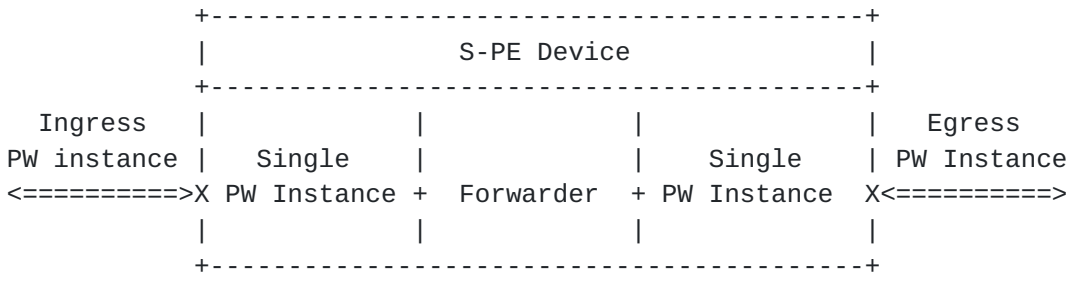

Figure 6 Point-to-Point Service

Other mappings of PW to forwarder are for further study.

5.1.2. Native Service Processing

There is no native service processing in the S-PEs.

6. Protocol Stack reference Model

Figure 7 illustrates the protocol stack reference model for multi-segment PWs.

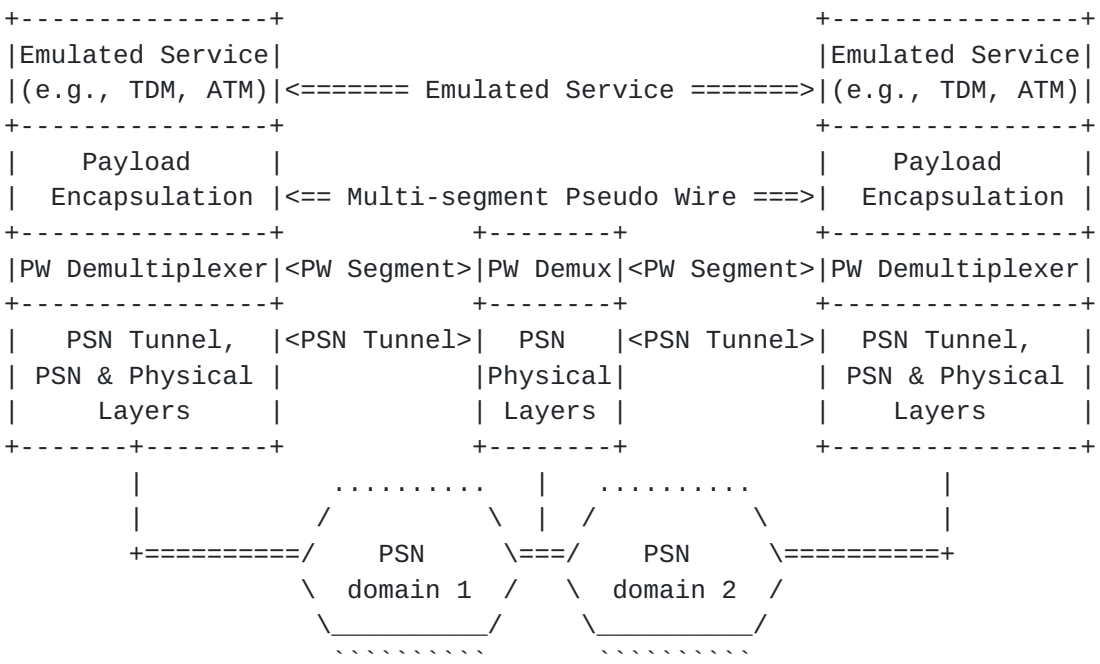

Figure 7 Multi-Segment PW Protocol Stack

The MS-PW provides the CE with an emulated physical or virtual connection to its peer at the far end. Native service PDUs from the CE are passed through an Encapsulation Layer and a PW demultiplexer is added at the sending T-PE. The PDU is sent over PSN domain 1. The receiving S-PE removes the existing PW demultiplexer, adds a new demultiplexer, and then sends the PDU over PSN2. Policies may also be applied to the PW at this point. Examples of such policies include: admission control, rate control, QoS mappings, and security. The receiving T-PE removes the PW demultiplexer and restores the payload to its native format for transmission to the destination CE.

Where the encapsulation format is different e.g. MPLS and L2TPv3, the payload encapsulation may be transparently translated at the S-PE.

7. Maintenance Reference Model

Figure 8 shows the maintenance reference model for multi-segment PWE3.

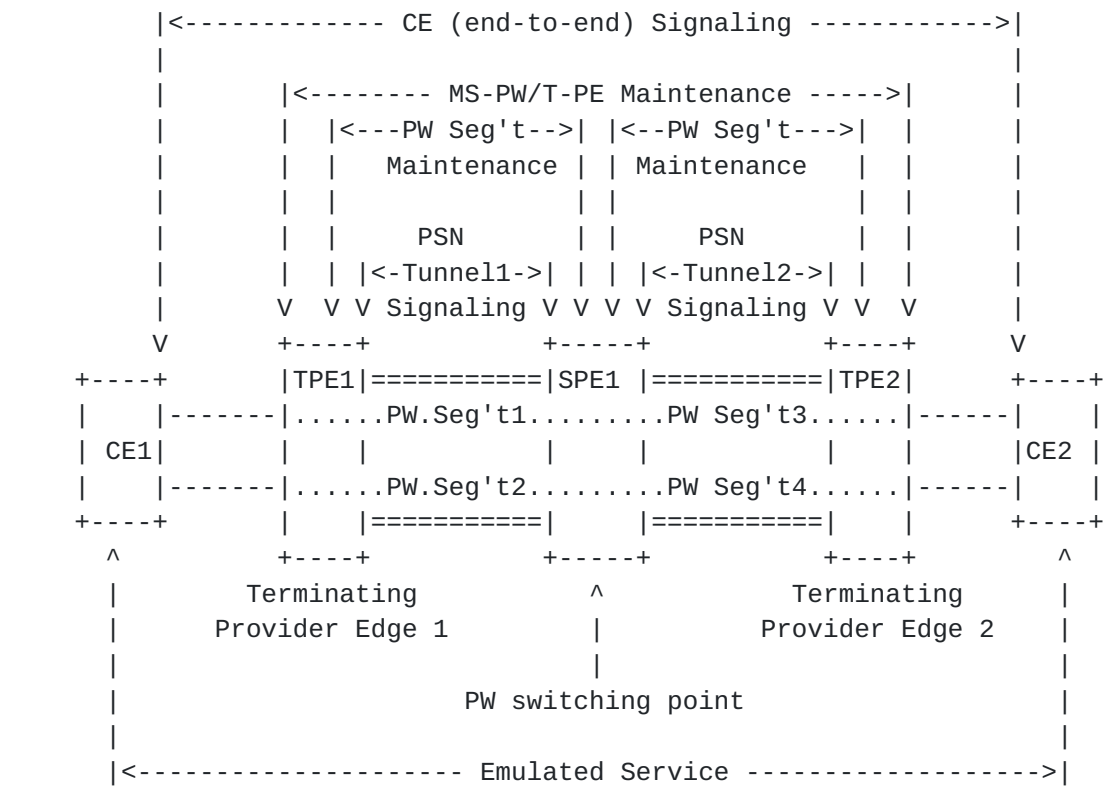

Figure 8 MS-PWE3 Maintenance Reference Model

RFC 3985 specifies the use of CE (end-to-end) and PSN tunnel signaling, and PW/PE maintenance. CE and PSN tunnel signaling is as specified in RFC 3985. However, in the case of MS-PWE3, signaling between the PEs now has both an edge-to-edge and a hop-by-hop context. That is, signaling and maintenance between T-PEs and S-PEs and between adjacent S-PEs is used to set up, maintain, and tear down the MS-PW segments, which including the coordination of parameters related to each switching point, as well as the MS-PW end points.

8. PW Demultiplexer Layer and PSN Requirements

8.1. Multiplexing

The purpose of the PW demultiplexer layer at the S-PE is to demultiplex PWs from ingress PSN tunnels and to multiplex them into egress PSN tunnels. Although each PW may contain multiple native service circuits, e.g. multiple ATM VCs, the S-PEs do not have visibility of, and hence do not change, this level of multiplexing because they contain no NSP.

8.2. Fragmentation

An S-PE is not required to make any attempt to reassemble a fragmented PW payload. An S-PE may fragment a PW payload.

9. Control Plane

9.1. Setup or Teardown of Pseudo Wires

For multi-segment pseudo wires, the intermediate PW switching points may be statically provisioned, or they may be dynamically signaled. For the dynamic case, there are two options for selecting the path of the PW:

- o T-PEs determine the full path of the PW through intermediate switching points. This may be either static or based on a dynamic PW path selection mechanism.

- o Each segment of the PW path is determined locally by each T-PE or S-PE, either through static configuration or based on a dynamic PW path selection mechanism.

Further details of the impact of these on the control plane architecture will be provided in a future revision.

9.2. Pseudo-Wire Up/Down Notification

Since a multi-segment PW consists of a number of concatenated PW segments, the emulated service can only be considered as being up when all of the PW segments and PSN tunnels (if used) are functional along the entire path of the MS-PW.

If a native service requires bi-directional connectivity, the corresponding emulated service can only be signaled as being up when the PW segments and PSN tunnels (if used), are functional in both directions.

RFC 3985 describes the need for failure and other status notification mechanisms for PWs. These considerations also apply to multi-segment. In addition, the S-PE must be able to propagate such notifications between concatenated segments of the same PW.

9.3. Misconnection and Payload Type Mismatch

With PWE3, misconnection and payload type mismatch can occur. Misconnection can breach the integrity of the system. Payload mismatch can disrupt the customer network. In both instances, there are security and operational concerns.

The services of the underlying tunneling mechanism and its associated control protocol can be used to ensure that the identity of the PW next hop is as expected. As part of the PW setup, a PW-TYPE identifier is exchanged. This is then used by the forwarder and the NSP of the T-PEs to verify the compatibility of the ACs. This can also be used by S-PEs to ensure that concatenated segments of a given MS-PW are compatible, or that a MS-PW is not misconnected into a local AC. In addition, it is advisable to do an end to end connection verification to check the integrity of the PW and to verify the identity of the T-PE.

10. Management and Monitoring

The management and monitoring as described in RFC 3985 apply here.

The need for an S-PE ping and PW trace route, and the mechanisms to provide these, are for further study.

11. Congestion Considerations

The control plane and the data plane fate-share in traditional IP networks. The implication of this is that congestion in the data plane can cause degradation of the operation of the control plane. Under quiescent operating conditions it is expected that the network will be designed to avoid such problems. However, MS-PW mechanisms should also consider what happens when congestion does occur, when the network is stretched beyond its design limits, for example during unexpected network failure conditions.

In addition to protecting the operation of the underlying PSN, consistent QoS and traffic engineering mechanisms should be used on each segment of a MS-PW to support the requirements of the emulated service.

12. IANA Considerations

This document does not contain any IANA actions.

13. Security Considerations

The security considerations described in RFC-3985 apply here.

Additional consideration needs to be given to the security of the S-PEs, particularly when these are dynamically selected and/or when the MS-PW transits the networks of multiple operators.

When the MS-PW is dynamically created by the use of a signaling protocol, an S-PE SHOULD determine the authenticity of the peer entity from which it receives the request, and its compliance with policy.

Particular consideration needs to be given to Quality of Service requests because the inappropriate use of priority may impact other service guarantees.

Where an S-PE provides interconnection between different providers, similar considerations to those applied to ASBRs apply. In particular peer entity authentication SHOULD be used.

Where an S-PE also supports T-PE functionality, mechanisms should be provided to ensure that MS-PWs to switched correctly to the appropriate outgoing PW segment, rather than a local AC. Other mechanisms for PW end point verification may also be used to confirm the correct PW connection prior to enabling the attachment circuits.

14. Acknowledgments

The authors gratefully acknowledge the input of Mustapha Aissaoui, Dimitri Papadimitrou, and Luca Martini.

15. References

15.1. Normative References

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

[2] Bryant, S. and Pate, P. (Editors), "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005

[3] Martini, S. Bitar, N. and Bocci, M (Editors), "Requirements for inter domain Pseudo-Wires", draft-ietf-pwe3-ms-pw-requirements-01.txt, Internet Draft, October 2005

Author's Addresses

Matthew Bocci
Alcatel
Voyager Place,
Shoppenhangers Rd,
Maidenhead, Berks, UK Email: matthew.bocci@alcatel.co.uk

Stewart Bryant
Cisco Systems,
250, Longwater,
Green Park,
Reading, RG2 6GB,
United Kingdom. Email: stbryant@cisco.com

Intellectual Property Statement

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

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

Disclaimer of Validity

Copyright Statement

Acknowledgment

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