Network Working Group
Request for Comments: 5659
Category: Informational

M. Bocci
Alcatel-Lucent
S. Bryant
Cisco Systems

October 2009

An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge

Abstract

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

Status of This Memo

Copyright and License Notice

Table of Contents

1. Introduction

RFC 3985 [1] defines the architecture for pseudowires, where a pseudowire (PW) both originates and terminates on the edge of the same packet switched network (PSN). The PW label is unchanged between the originating and terminating provider edges (PEs). This is now known as a single-segment pseudowire (SS-PW).

This document extends the architecture in RFC 3985 to enable point-to-point pseudowires to be extended through multiple PSN tunnels. These are known as multi-segment pseudowires (MS-PWs). Use cases for multi-segment pseudowires (MS-PWs), and the consequent requirements, are defined in RFC 5254 [5].

1.1. Motivation and Context

RFC 3985 addresses the case where a PW spans a single segment between two PEs. Such PWs are termed single-segment pseudowires (SS-PWs) and provide point-to-point connectivity between two edges of a provider network. However, there is now a requirement to be able to construct multi-segment pseudowires. These requirements are specified in RFC 5254 [5] and address three main problems:

i. 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.

ii. How to provide PWs across multiple PSN routing domains or areas in the same provider.

iii. How to provide PWs across multiple provider domains and different PSN types.

Consider a single PW domain, such as that shown in Figure 1. There are 4 PEs, and PWs must be provided from any PE to any other PE. PWs can be supported by establishing a full mesh of PSN tunnels between the PEs, requiring a full mesh of LDP signaling adjacencies between the PEs. PWs can therefore be established between any PE and any other PE via a single, direct PSN tunnel that is switched only by intermediate P-routers (not shown in the figure). In this case, each PW is an SS-PW. A PE must terminate all the pseudowires that are carried on the 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, PSN tunnels, and signaling adjacencies will grow in proportion to the square of the number of PEs.

For reasons of economy, the edge PEs that terminate the attachment circuits (ACs) are often small devices built to very low cost with limited processing power. Consider an example where a particular PE, residing at the edge of a provider network, terminates N PWs to/from N different remote PEs. This needs N PW signaling adjacencies to be set up and maintained. If the edge PE attaches to a single intermediate PE that is able to switch the PW, that edge PE only needs a single adjacency to signal and maintain all N PWs. The intermediate switching PE (which is a larger device) needs M signaling adjacencies, but statistically this is less than tN, where t is the number of edge PEs that it is serving. Similarly, if the PWs are running over TE PSN tunnels, there is a statistical reduction in the number of TE PSN tunnels that need to be set up and maintained between the various PEs.

One possible solution that is more efficient for large numbers of PEs, in particular for the control plane, is therefore to support a partial mesh of PSN tunnels between the PEs, as shown in Figure 1. For example, consider a PW service whose endpoints are PE1 and PE4. Pseudowires for this can take the path PE1->PE2->PE4 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-PE4. The end-to-end PW is known as a multi-segment PW.

```
                                   ,,'..--..,,_,
                               .-''            '',.
                       +-----+'                    '+-----+
                       | PE1 |---------------------| PE2 |
                       |     |---------------------|     |
                       +-----+      PSN Tunnel     +-----+
                        / ||                          || \
                       /  ||                          ||  \
                      |   ||                          ||   |
                      |   ||          PSN             ||   |
                      |   ||                          ||   |
                       \  ||                          ||  /
                        \ ||                          || /
                         \||                          ||/
                       +-----+                     +-----+
                       | PE3 |---------------------| PE4 |
                       |     |---------------------|     |
                       +-----+''.,_           ,.'' +-----+
                                   ''''___''''
```

Figure 1: PWs Spanning a Single PSN with Partial Mesh of PSN Tunnels

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, for example, supporting different PSN types or using different routing protocols.

An example application is shown in Figure 2. Here, the provider's 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 could also represent interworking points between different types of IP networks, e.g., those with MPLS and those without, and points where network policies can be applied.

```
            <-------- Edge to Edge Emulated Services ------->

                ,'    .      ,-'       '',       ,'    .
               /       \   .'             ',    /       \
              /        \  /                 ,  /        \
       AC  +----+     +----+               +----+       +----+    AC
        ---| PE |-----| PE |---------------| PE |-------| PE |---
           |  1 |     |  2 |               | 3  |       | 4  |
           +----+     +----+               +----+       +----+
              \        /  \                 /  \        /
               \       /  \      Core       '   \       /
                ',    '     .              ,'    ',    '
                 ,'-''       '.,       _.'        ,'-''
               Access 1         '''-'''         Access 2
```

Figure 2: Multi-Domain Network Model

A similar model can also be applied to inter-provider services, where a single PW spans a number of separate provider networks in order to connect ACs residing on PEs in disparate provider networks. In this case, each provider will typically maintain their own PE at the border of their network in order to apply policies such as security and Quality of Service (QoS) to PWs entering their network. Thus, the connection between the domains will normally be a link between two PEs on the border of each provider's network.

Consider the application of this model to PWs. PWs use tunneling mechanisms such as MPLS to enable the underlying 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). Pseudowires 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:

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

1.3. Terminology

The terminology specified in RFC 3985 [1] and RFC 4026 [2] applies. In addition, we define the following terms:

- o 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 an MS-PW. This incorporates the functionality of a PE as defined in RFC 3985.
- o Single-Segment Pseudowire (SS-PW). A PW set up directly between two T-PE devices. The PW label is unchanged between the originating and terminating T-PEs.

- o Multi-Segment Pseudowire (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 an MS-PW, by definition, terminates on a T-PE.

- o PW Segment. A part of a single-segment or multi-segment PW, which traverses one PSN tunnel in each direction between two PE devices, T-PEs, and/or S-PEs (switching PE).

- 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 an MS-PW. The S-PE terminates the PSN tunnels of the preceding and succeeding segments of the MS-PW. It therefore includes a PW switching point for an 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 set up and manage PW segments with other PW switching points and terminating PEs. An S-PE can exist anywhere a PW must be processed or policy applied. It is therefore not limited to the edge of a provider network.

  Note that it was originally anticipated that S-PEs would only be deployed at the edge of a provider network where they would be used to switch the PWs of different service providers. However, as the design of MS-PW progressed, other applications for MS-PW were recognized. By this time S-PE had become the accepted term for the equipment, even though they were no longer universally deployed at the provider edge.

- o PW Switching. The process of switching the control and data planes of the preceding and succeeding PW segments in a MS-PW.

- o PW Switching Point. The reference point in an S-PE where the switching takes place, e.g., where PW label swap is executed.

- o Eligible S-PE or T-PE. An eligible S-PE or T-PE is a PE that meets the security and privacy requirements of the MS-PW, according to the network operator’s policy.

- o Trusted S-PE or T-PE. A trusted S-PE or T-PE is a PE that is understood to be eligible by its next-hop S-PE or T-PE, while a trust relationship exists between two S-PEs or T-PEs if they mutually consider each other to be eligible.

2. Applicability

An 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 Layer 2 Virtual Private Network (L2VPN), an MS-PW is indistinguishable from an SS-PW. Thus, the following are equivalent from the perspective of the T-PE:

```
+----+                                                  +----+
|TPE1+--------------------------------------------------+TPE2|
+----+                                                  +----+

|<--------------------------PW---------------------------->|

+----+              +---+          +---+                +----+
|TPE1+--------------+SPE+----------+SPE+----------------+TPE2|
+----+              +---+          +---+                +----+
```

Figure 3: MS-PW Equivalence

Although an MS-PW may require services such as node discovery and path signaling to construct the PW, it should not be confused with an L2VPN system, which also requires these services. A Virtual Private Wire Service (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.

As well as supporting traditional L2VPNs, an MS-PW is applicable to providing connectivity across a transport network based on packet switching technology, e.g., the MPLS Transport Profile (MPLS-TP) [6], [8]. Such a network uses pseudowires to support the transport and aggregation of all services. This application requires deterministic characteristics and behavior from the network. The operational requirements of such networks may need pseudowire segments that can be established and maintained in the absence of a control plane, and may also need the operational independence of PW maintenance from the underlying PSN.

3. Protocol Layering Model

The protocol layering model specified in RFC 3985 applies to MS-PWs with the following clarification: the pseudowires may be considered to be a separate layer to the PSN tunnel. That is, although a PW segment will follow the path of the PSN tunnel between S-PEs, the

MS-PW is 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-PWs will reuse the protocols of the PSN and may, if applicable, use information that is extracted from the PSN, e.g., reachability.

3.1. Domain of MS-PW Solutions

PWs provide the Encapsulation Layer, i.e., the method of carrying various payload types, and the interface to the PW Demultiplexer Layer. Other layers provide the following:

- o PSN tunnel setup, maintenance, and routing
- o T-PE discovery

Not all PEs may be capable of providing S-PE functionality. Connectivity to the next-hop S-PE or T-PE must be provided by a PSN tunnel, according to [1]. The selection of which set of S-PEs to use to reach a given T-PE is considered to be within the scope of MS-PW solutions.

3.2. Payload Types

MS-PWs are applicable to all PW payload types. Encapsulations defined for SS-PWs are also used for MS-PW without change. Where the PSN types for each segment of an MS-PW are identical, the PW types of each segment must also be identical. However, if different segments run over different PSN types, the encapsulation may change but the PW segments must be of an equivalent PW type, i.e., the S-PE must not need to process the PW payload to provide translation.

4. Multi-Segment Pseudowire Reference Model

The pseudowire emulation edge-to-edge (PWE3) reference architecture for the single-segment case is shown in [1]. 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.

```
       Native  |<------Multi-Segment Pseudowire------>|  Native
       Service |         PSN              PSN         |  Service
        (AC)   |     |<-Tunnel->|     |<-Tunnel->|    |   (AC)
          |    V     V     1    V     V    2     V    V     |
          |    +----+           +-----+          +----+     |
   +----+ |    |TPE1|===========|SPE1 |==========|TPE2|     | +----+
   |    |------|..... PW.Seg't1....X....PW.Seg't3.....|-------|    |
   | CE1| |    |    |           |     |          |    |     | |CE2 |
   |    |------|..... PW.Seg't2....X....PW.Seg't4.....|-------|    |
   +----+ |    |    |===========|     |==========|    |     | +----+
        ^      +----+           +-----+          +----+      ^
        |   Provider Edge 1        ^        Provider Edge 2  |
        |                          |                         |
        |                          |                         |
        |                  PW switching point                |
        |                                                    |
        |<------------------ Emulated Service --------------->|
```

Figure 4: MS-PW Reference Model

Figure 4 extends this architecture to show a multi-segment case. The PEs that provide services 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 T-PE2 across PSN2. PWs are used to connect the attachment circuits (ACs) attached to PE1 to the corresponding ACs attached to T-PE2.

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., PW segment 1 and PW segment 3) must be of equivalent PW types, as described in Section 3.2, while PSN tunnels (e.g., PSN1 and PSN2) may be of the same or different PSN types. An S-PE switches an MS-PW from one segment to another based on the PW demultiplexer, i.e., a PW label that may take one of the forms defined in Section 5.4.1 of RFC 3985 [1].

Note that although Figure 4 only shows a single S-PE, a PW may transit more than 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. Both directions of an MS-PW must traverse the same set of S-PEs on a reciprocal path. Note that although the S-PE path is therefore reciprocal, the path taken by the PSN tunnels between the T-PEs and S-PEs might not be reciprocal due to choices made by the PSN routing protocol.

4.1. Intra-Provider Connectivity Architecture

There is a requirement to deploy PWs edge-to-edge in large service provider networks (RFC 5254 [5]). 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 PW domains, where the PEs are interconnected by a sparse mesh of tunnels.

Whether or not the network is partitioned into separate PW domains, there is 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 regions, core regions, or metro regions within the same provider's network. PSN1 and PSN2 may also be of different types. For example, S-PEs may be used to connect PW segments traversing metro networks of one technology, e.g., statically allocated labels, with segments traversing an MPLS core network.

Alternatively, T-PE1, S-PE1, and T-PE2 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 domain boundary 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, PSN1 and PSN2 would be portions of the same provider's network.

4.2. Inter-Provider Connectivity Architecture

Inter-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 PW-based services to CEs attached to each provider's network. In addition, the following may need to be implemented on a per-PW basis at the provider boundary:

- o Operations, Administration, and Maintenance (OAM). Note that this is synonymous with 'Operations and Maintenance' referred to in RFC 5254 [5].

- o Authentication, Authorization, and Accounting (AAA)

- o Security mechanisms

Further security-related architectural considerations are described in Section 12.

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. This is similar to the case in Section 4.1.1, except that additional security and policy enforcement measures will be required. 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. This 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 PSN2.

```
                |<------Multi-Segment Pseudowire------>|
                |       Provider         Provider      |
           AC   |    |<----1---->|     |<----2--->|    |  AC
            |   V    V           V     V          V    V  |
            |   +----+     +-----+     +----+     +----+  |
   +----+   |   |    |=====|     |=====|    |=====|    |  |    +----+
   |    |-------|......PW.....X....PW.....X...PW.......|-------|    |
   | CE1|   |   |    |Seg 1|     |Seg 2|    |Seg 3|    |  |    |CE2 |
   +----+   |   |    |=====|     |=====|    |=====|    |  |    +----+
        ^       +----+     +-----+     +----+     +----+       ^
        |       T-PE1       S-PE1       S-PE2     T-PE2        |
        |                     ^          ^                     |
        |                     |          |                     |
        |                  PW switching points                 |
        |                                                      |
        |                                                      |
        |<------------------- Emulated Service --------------->|
```

Figure 5: Inter-Provider Reference Model

5. PE Reference Model

5.1. Pseudowire Pre-Processing

Pseudowire pre-processing 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.

```
               +------------------------------------------+
               |                S-PE Device               |
               +------------------------------------------+
     Ingress   |             |             |              |   Egress
   PW instance |   Single    |             |    Single    | PW Instance
   <==========>X PW Instance +  Forwarder  + PW Instance  X<==========>
               |             |             |              |
               +------------------------------------------+
```

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.

```
   +-----------+                                  +-----------+
   |  Emulated |                                  |  Emulated |
   |  Service  |                                  |  Service  |
   |(e.g., ATM)|<======= Emulated Service =======>|(e.g., ATM)|
   +-----------+                                  +-----------+
   | Payload   |                                  | Payload   |
   |  Encap.   |<=== Multi-segment Pseudowire ===>|  Encap.   |
   +-----------+            +--------+            +-----------+
   | PW Demux  |<PW Segment>|PW Demux|<PW Segment>| PW Demux  |
   +-----------+            +--------+            +-----------+
   |PSN Tunnel,|<PSN Tunnel>|  PSN   |<PSN Tunnel>|PSN Tunnel,|
   | PSN & PHY |            |Physical|            | PSN & PHY |
   | Layers    |            | Layers |            |  Layers   |
   +----+------+            +--------+            +-----+-----+
        |            ..........   |   ..........        |
        |           /          \  |  /          \       |
        +==========/    PSN     \===/    PSN     \======+
                   \  domain 1  /   \  domain 2  /
                    \__________/     \__________/
                     ''''''''''       ''''''''''
```

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 via the PSN transport tunnel. The receiving S-PE swaps the existing PW demultiplexer for the demultiplexer of the next segment and then sends the PDU over transport tunnel in PSN2. Where the ingress and egress PSN domains of the S-PE are of the same type, e.g., they are both MPLS PSNs, a simple label swap operation is performed, as described in Section 3.13 of RFC 3031 [3]. However, where the ingress and egress PSNs are of different types, e.g., MPLS and L2TPv3, the ingress PW demultiplexer is removed (or popped), and a mapping to the egress PW demultiplexer is performed and then inserted (or pushed).

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 translated at the S-PE.

7. Maintenance Reference Model

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

```
        |<------------- CE (end-to-end) Signaling ------------>|
        |                                                      |
        |       |<-------- MS-PW/T-PE Maintenance ----->|      |
        |       |  |<---PW Seg't-->| |<--PW Seg't--->|  |      |
        |       |  |   Maintenance | | Maintenance   |  |      |
        |       |  |               | |               |  |      |
        |       |  |     PSN       | |     PSN       |  |      |
        |       |  | |<-Tunnel1->| | | |<-Tunnel2->| |  |      |
        |       V  V V Signaling V V V V Signaling V V  V      |
        V       +----+           +-----+           +----+      V
   +----+       |TPE1|===========|SPE1 |===========|TPE2|      +----+
   |    |-------|......PW.Seg't1....X....PW Seg't3......|------|    |
   | CE1|       |    |           |     |           |    |      |CE2 |
   |    |-------|......PW.Seg't2....X....PW Seg't4......|------|    |
   +----+       |    |===========|     |===========|    |      +----+
     ^          +----+           +-----+           +----+         ^
     |        Terminating          ^           Terminating        |
     |      Provider Edge 1        |         Provider Edge 2      |
     |                             |                              |
     |                     PW switching point                     |
     |                                                            |
     |<--------------------- Emulated Service ------------------->|
```

Figure 8: MS-PW Maintenance Reference Model

RFC 3985 specifies the use of CE (end-to-end) and PSN tunnel signaling as well as PW/PE maintenance. CE and PSN tunnel signaling is as specified in RFC 3985. However, in the case of MS-PWs, 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 includes the coordination of parameters related to each switching point as well as to the MS-PW endpoints.

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 virtual circuits (VCs), the S-PEs do not have visibility of, and hence do not change, this level of multiplexing because they contain no Native Service Processor (NSP).

8.2. Fragmentation

If fragmentation is to be used in an MS-PW, T-PEs and S-PEs must satisfy themselves that fragmented PW payloads can be correctly reassembled for delivery to the destination attachment circuit.

An S-PE is not required to make any attempt to reassemble a fragmented PW payload. However, it may choose to do so if, for example, it knows that a downstream PW segment does not support reassembly.

An S-PE may fragment a PW payload using [4].

9. Control Plane

9.1. Setup and Placement of MS-PWs

For multi-segment pseudowires, the intermediate PW switching points may be statically provisioned or chosen dynamically.

For the static case, there are two options for exchanging the PW labels:

o By configuration at the T-PEs or S-PEs.

o By signaling across each segment using a dynamic maintenance protocol.

A multi-segment pseudowire may thus consist of segments where the labels are statically configured and segments where the labels are signaled.

For the case of dynamic choice of the PW switching points, there are two options for selecting the path of the MS-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 T-PE and S-PE makes a local decision as to which next-hop S-PE to choose to reach the target T-PE. This choice is made either using locally configured information or by using a dynamic PW path-selection mechanism.

9.2. Pseudowire 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 constituting PW segments and PSN tunnels are functional and operational 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 and operational in both directions.

RFC 3985 describes the architecture of failure and other status notification mechanisms for PWs. These mechanisms are also needed in multi-segment pseudowires. In addition, if a failure notification mechanism is provided for consecutive segments of the same PW, the S-PE must propagate such notifications between the consecutive concatenated segments.

9.3. Misconnection and Payload Type Mismatch

Misconnection and payload type mismatch can occur with PWs. 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 or the PW control and OAM protocols 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 an MS-PW is not misconnected into a local AC. In addition, it is possible to perform an end-to-end connection verification to check the integrity of the PW, to verify the identity of S-PEs and check the correct connectivity at S-PEs, and to verify the identity of the T-PE.

10. Management and Monitoring

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

The MS-PW architecture introduces additional considerations related to management and monitoring, which need to be reflected in the design of maintenance tools and additional management objects for MS-PWs.

The first is that each S-PE is a new point at which defects may occur along the path of the PW. In order to troubleshoot MS-PWs, management and monitoring should be able to operate on a subset of the segments of an MS-PW, as well as edge-to-edge. That is, connectivity verification mechanisms should be able to troubleshoot and differentiate the connectivity between T-PEs and intermediate S-PEs, as well as the connectivity between T-PE and T-PE.

The second is that the set of S-PEs and P-routers along the MS-PW path may be less optimal than a path between the T-PEs chosen solely by the underlying PSN routing protocols. This is because the S-PEs are chosen by the MS-PW path selection mechanism and not by the PSN routing protocols. Troubleshooting mechanisms should therefore be provided to verify the set of S-PEs that are traversed by an MS-PW to reach a T-PE.

Some of the S-PEs and the T-PEs for an MS-PW may reside in a different service provider's PSN domain from that of the operator who initiated the establishment of the MS-PW. These situations may necessitate the use of remote management of the MS-PW, which is able to securely operate across provider boundaries.

11. Congestion Considerations

The following congestion considerations apply to MS-PWs. These are in addition to the considerations for PWs described in RFC 3985 [1], [7], and the respective RFCs specifying each PW type.

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.

Although congestion within a single provider's network can be mitigated by suitable engineering of the network so that the traffic imposed by PWs can never cause congestion in the underlying PSN, a significant number of MS-PWs are expected to be deployed for inter-provider services. In this case, there may be no way of a provider who initiates the establishment of an MS-PW at a T-PE guaranteeing that it will not cause congestion in a downstream PSN. A specific PSN may be able to protect itself from excess PW traffic by policing all PWs at the S-PE at the provider border. However, this may not be effective when the PSN tunnel across a provider utilizes the transit services of another provider that cannot distinguish PW traffic from ordinary, TCP-controlled IP traffic.

Each segment of an MS-PW therefore needs to implement congestion detection and congestion control mechanisms where it is not possible to explicitly provision sufficient capacity to avoid congestion.

In many cases, only the T-PEs may have sufficient information about each PW to fairly apply congestion control. Therefore, T-PEs need to be aware of which of their PWs are causing congestion in a downstream PSN and of their native service characteristics, and to apply congestion control accordingly. S-PEs therefore need to propagate PSN congestion state information between their downstream and upstream directions. If the MS-PW transits many S-PEs, it may take some time for congestion state information to propagate from the congested PSN segment to the source T-PE, thus delaying the application of congestion control. Congestion control in the S-PE at the border of the congested PSN can enable a more rapid response and thus potentially reduce the duration of congestion.

In addition to protecting the operation of the underlying PSN, consistent QoS and traffic engineering mechanisms should be used on each segment of an MS-PW to support the requirements of the emulated service. The QoS treatment given to a PW packet at an S-PE may be derived from context information of the PW (e.g., traffic or QoS parameters signaled to the S-PE by an MS-PW control protocol) or from PSN-specific QoS flags in the PSN tunnel label or PW demultiplexer, e.g., TC bits in either the label switched path (LSP) or PW label for an MPLS PSN or the DS field of the outer IP header for L2TPv3.

12. Security Considerations

The security considerations described in RFC 3985 [1] apply here. Detailed security requirements for MS-PWs are specified in RFC 5254 [5]. This section describes the architectural implications of those requirements.

The security implications for T-PEs are similar to those for PEs in single-segment pseudowires. However, S-PEs represent a point in the network where the PW label is exposed to additional processing. An S-PE or T-PE must trust that the context of the MS-PW is maintained by a downstream S-PE. OAM tools must be able to verify the identity of the far end T-PE to the satisfaction of the network operator. Additional consideration needs to be given to the security of the S-PEs, both at the data plane and the control plane, particularly when these are dynamically selected and/or when the MS-PW transits the networks of multiple operators.

An implicit trust relationship exists between the initiator of an MS-PW, the T-PEs, and the S-PEs along the MS-PW's path. That is, the T-PE trusts the S-PEs to process and switch PWs without compromising the security or privacy of the PW service. An S-PE should not select a next-hop S-PE or T-PE unless it knows it would be considered eligible, as defined in Section 1.3, by the originator of the MS-PW. For dynamically placed MS-PWs, this can be achieved by allowing the T-PE to explicitly specify the path of the MS-PW. When the MS-PW is dynamically created by the use of a signaling protocol, an S-PE or T-PE should determine the authenticity of the peer entity from which it receives the request and the compliance of that request with policy.

Where an MS-PW crosses a border between one provider and another provider, the MS-PW segment endpoints (S-PEs or T-PEs) or, for the PSN tunnel, P-routers typically reside on the same nodes as the Autonomous System Border Router (ASBRs) interconnecting the two providers. In either case, an S-PE in one provider is connected to a limited number of trusted T-PEs or S-PEs in the other provider. The number of such trusted T-PEs or S-PEs is bounded and not anticipated to create a scaling issue for the control plane authentication mechanisms.

Directly interconnecting the S-PEs/T-PEs using a physically secure link and enabling signaling and routing authentication between the S-PEs/T-PEs eliminates the possibility of receiving an MS-PW signaling message or packet from an untrusted peer. The S-PEs/T-PEs represent security policy enforcement points for the MS-PW, while the ASBRs represent security policy enforcement points for the provider's PSNs. This architecture is illustrated in Figure 9.

```
                  |<------------- MS-PW ---------------->|
                  |       Provider         Provider      |
             AC   |    |<----1---->|     |<----2--->|    |  AC
              |   V    V           V     V          V    V  |
              |   +----+     +-----+     +----+     +----+  |
      +---+   |   |    |=====|     |=====|    |=====|    |  |    +---+
      |   |-------|......PW.....X....PW.....X...PW.......|-------|   |
      |CE1|   |   |    |Seg 1|     |Seg 2|    |Seg 3|    |  |    |CE2|
      +---+   |   |    |=====|     |=====|    |=====|    |  |    +---+
          ^       +----+     +-----+  ^  +----+     +----+       ^
          |       T-PE1       S-PE1   |   S-PE2     T-PE2        |
          |                    ASBR   |    ASBR                  |
          |                           |                          |
          |                  Physically secure link              |
          |                                                      |
          |                                                      |
          |<------------------- Emulated Service --------------->|
```

Figure 9: Directly Connected Inter-Provider Reference Model

Alternatively, the P-routers for the PSN tunnel may reside on the ASBRs, while the S-PEs or T-PEs reside behind the ASBRs within each provider's network. A limited number of trusted inter-provider PSN tunnels interconnect the provider networks. This is illustrated in Figure 10.

```
                |<-------------- MS-PW -------------------->|
                |          Provider          Provider       |
            AC  |    |<------1----->|   |<-----2------->|   |  AC
             |  V    V              V   V               V   V  |
             |  +---+     +---+  +--+   +--+  +---+     +---+  |
      +---+  |  |   |=====|   |===============|   |=====|   |  |   +---+
      |   |-----|.....PW....X.......PW..............PW....X.|------|   |
      |CE1|  |  |   |Seg 1|   |    Seg 2      |   |Seg 3|   |  |   |CE2|
      +---+  |  |   |=====|   |===============|   |=====|   |  |   +---+
          ^     +---+     +---+  +--+ ^ +--+  +---+     +---+      ^
          |      T-PE1    S-PE1  ASBR | ASBR  S-PE2     T-PE2      |
          |                           |                            |
          |                           |                            |
          |                Trusted Inter-AS PSN Tunnel             |
          |                                                        |
          |                                                        |
          |<------------------- Emulated Service ----------------->|
```

Figure 10: Indirectly Connected Inter-Provider Reference Model

Particular consideration needs to be given to Quality of Service requests because the inappropriate use of priority may impact any service guarantees given to other PWs. Consideration also needs to be given to the avoidance of spoofing the PW demultiplexer.

Where an S-PE provides interconnection between different providers, security considerations that are similar to the security considerations for 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 are switched correctly to the appropriate outgoing PW segment, rather than to a local AC. Other mechanisms for PW endpoint verification may also be used to confirm the correct PW connection prior to enabling the attachment circuits.

13. Acknowledgments

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

14. References

14.1. Normative References

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

[2] Andersson, L. and T. Madsen, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC 4026, March 2005.

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

[4] Malis, A. and M. Townsley, "Pseudowire Emulation Edge-to-Edge (PWE3) Fragmentation and Reassembly", RFC 4623, August 2006.

14.2. Informative References

[5] Bitar, N., Ed., Bocci, M., Ed., and L. Martini, Ed., "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3)", RFC 5254, October 2008.

[6] Niven-Jenkins, B., Ed., Brungard, D., Ed., Betts, M., Ed., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[7] Bryant, S., Davie, B., Martini, L., and E. Rosen, "Pseudowire Congestion Control Framework", Work in Progress, June 2009.

[8] Bocci, M., Bryant, S., and L. Levrau, "A Framework for MPLS in Transport Networks", Work in Progress, August 2009.

Authors' Addresses

Matthew Bocci
Alcatel-Lucent
Voyager Place, Shoppenhangers Road,
Maidenhead, Berks, UK
Phone: +44 1633 413600
EMail: matthew.bocci@alcatel-lucent.com

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