Network Working Group Luca Martini
Internet Draft Chris Metz
Expiration Date: April 2006 Thomas D. Nadeau
Cisco Systems Inc.

Vasile Radoaca Mike Duckett
Consultant Bellsouth

Matthew Bocci Florin Balus
Alcatel Nortel Networks

October 2005

Segmented Pseudo Wire

draft-ietf-pwe3-segmented-pw-01.txt

Status of this Memo

Abstract

This document describes how to connect pseudo wires (PW) between two distinct PW control planes or PSN domains. The PW control planes may belong to independent autonomous systems, or the PSN technology is heterogeneous, or a PW might need to be aggregated at a specific PSN point. The PW packet data units are simply switched from one PW to another without changing the PW payload.

Table of Contents

1. Specification of Requirements

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

2. Terminology

- Ultimate PE (U-PE). A PE where the customer-facing ACs are bound to a PW forwarder. An ultimate PE is present in the first and last segments of a MH-PW.

- Single-Hop PW (SH-PW). A PW set up between two PE devices using standard PWE3 signaling and encapsulation methods.

- Multi-hop PW (MH-PW). Static or dynamically configured set of two or more contiguous PW segments (SH-PW) that behave and function as a single point-to-point PW. Each PW segment is setup and managed using standard PWE3 encapsulation and signaling methods.

- PW Switching Point. (S-PE) A PE capable of switching the control and data planes of the preceding and succeeding PW segments (SH-PW) in a MH-PW. A PW Switching Point is never the ultimate PE in a MH-PW. A PW switching point runs standard PWE3 protocols to setup and manage PW segments with other PW switching points and ultimate PEs.

3. Introduction

PWE3 defines the signaling and encapsulation techniques for establishing SH-PWs between a pair of ultimate PEs and in the vast majority of cases this will be sufficient. MH-PWs come into play in two general cases:

-i. When it is not possible, desirable or feasible to establish a PW control channel between the ultimate source and destination PEs. At a minimum PW control channel establishment requires knowledge of and reachability to the remote (ultimate) PE IP address. The local (ultimate) PE may

not have access to this information related to topology, operational or security constraints.

An example is the inter-AS L2VPN scenario where the ultimate PEs reside in different provider networks (ASes) and it is the practice to MD5-key all control traffic exchanged between two networks. Technically a SH-PW could be used but this would require MD5-keying on ALL ultimate source and destination PE nodes. An MH-PW allows the providers to confine MD5 key administration to just the PW switching points connecting the two domains.

A second example might involve a single AS where the PW setup path between the ultimate PEs is computed by an external entity (i.e. client-layer routing protocol). Assume a full mesh of PWE3 control channels established between PE-A, PE-B and PE-C. A client-layer L2 connection tunneled through a PW is required between ultimate PE-A and PE-C. The external entity computes a PW setup path that passes through PE-B. This results in two discrete PW segments being built: one between PE-A and PE-B and one between PE-B and PE-C. The successful client-layer L2 connection between ultimate PE-A and ultimate PE-C requires that PE-B performs the PWE3 switching process.

A third example involves the use of PWs in hierarchical IP/MPLS networks. Access networks connected to a backbone use PWs to transport customer payloads between customer sites serviced by the same access network and up to the edge of the backbone where they can be terminated or switched onto a succeeding PW segment crossing the backbone. The use of PWE3 switching between the access and backbone networks can potentially reduce the PWE3 control channels and routing information processed by the access network U-PEs.

It should be noted that PWE3 switching does not help in any way to reduce the amount of PW state supported by each access network U-PE.

-ii. PWE3 signaling and encapsulation protocols are different. The ultimate PEs are connected to networks employing different PW signaling and encapsulation protocols. In this case it is not possible to use a SH-PW. A MH-PW with the appropriate interworking performed at the PW switching points can enable PW connectivity between the ultimate PEs in this scenario.

There are four different signaling protocols that are defined to signal PWs:

- -i. Static configuration of the PW (MPLS or L2TPv3).
- -ii. LDP using FEC 128
- -iii. LDP using the generalized FEC 129
- -iv. L2TPv3

4. General Description

A pseudo-wire (PW) is a tunnel established between two provider edge (PE) nodes to transport L2 PDUs across a packet switched network (PSN) as described in Figure 1 and in [PWE3-ARCH]. Many providers are looking at PWs as a means of migrating existing (or building new) L2VPN services (i.e. Frame-Relay, ATM, Ethernet) on top of a PSN by using PWs. PWs may span multiple autonomous systems of the same or different provider networks. In these scenarios PW control channels (i.e. targeted LDP, L2TPv3) and PWs will cross AS boundaries.

Inter-AS L2VPN functionality is currently supported and several techniques employing MPLS encapsulation and LDP signaling have been documented [2547BIS]. It is also straightforward to support the same inter-AS L2VPN functionality employing L2TPv3. In this document we define methodology to switch a PW between two PW control planes.

```
         |<-------------- Emulated Service ---------------->|
         |                                                  |
         |          |<------- Pseudo Wire ------>|          |
         |          |                            |          |
         |          |    |<-- PSN Tunnel -->|    |          |
         |          V    V                  V    V          |
         V    AC    +----+                  +----+     AC   V
   +-----+    |     | PE1|==================| PE2|     |    +-----+
   |     |----------|............PW1.............|----------|     |
   | CE1 |    |     |    |                  |    |     |    | CE2 |
   |     |----------|............PW2.............|----------|     |
   +-----+  ^ |     |    |==================|    |     | ^  +-----+
         ^  |       +----+                  +----+     | |  ^
         |  |   Provider Edge 1         Provider Edge 2  |  |
         |  |                                            |  |
   Customer |                                            | Customer
   Edge 1   |                                            | Edge 2
            |                                            |
      native service                               native service
```

Figure 1: PWE3 Reference Model

There are two methods for switching a PW between two PW control planes. In the first method (Figure 2), the two control planes terminate on different PEs.

```
                |<------------Emulated Service---------->|
                |      PSN                      PSN      |
            AC  |    |<-1->|                  |<-2->|    |  AC
            |   V    V     V                  V     V    V  |
            |   +----+     +-----+       +----+     +----+  |
+----+   |   |    |=====|     |       |    |=====|    |  |    +----+
|    |-------|......PW1.......|--AC1--|......PW2......|-------|    |
| CE1|   |   |    |     |     |       |    |     |    |  |    |CE2 |
|    |-------|......PW3.......|--AC2--|......PW4......|-------|    |
+----+   |   |    |=====|     |       |    |=====|    |  |    +----+
     ^       +----+     +-----+       +----+     +----+       ^
     |         PE1        PE2          PE3         PE4        |
     |                     ^            ^                     |
     |                     |            |                     |
     |                  PW stitching points                   |
     |                                                        |
     |                                                        |
     |<-------------------- Emulated Service ---------------->|
```

Figure 2: PW Switching using ACs Reference Model

In Figure 2, pseudo wires in two separate PSNs are stitched together using native service attachment circuits. PE2 and PE3 only run the control plane for the PSN to which they are directly attached. At PE2 and PE3, PW1 and PW2 are connected using attachment circuit AC1, while PW3 and PW4 are connected using attachment circuit AC2.

```
           Native  |<-----------Pseudo Wire----------->|  Native
           Layer2  |                                   |  Layer2
          Service  |    |<-PSN1-->|     |<--PSN2->|    |  Service
           (AC)    V    V         V     V         V    V   (AC)
             |     +----+         +-----+         +----+     |
   +----+    |     | PE1|=========| PE2 |=========| PE3|     |    +----+
   |    |----------|........PW1.........|...PW3........|----------|    |
   | CE1|    |     |    |         |     |         |    |     |    |CE2 |
   |    |----------|........PW2.........|...PW4........|----------|    |
   +----+    |     |    |=========|     |=========|    |     |    +----+
        ^          +----+         +-----+         +----+     |    ^
        |      Provider Edge 1       ^        Provider Edge 3     |
        |       (Ultimate PE)        |         (Ultimate PE)      |
        |                            |                            |
        |                    PW switching point                   |
        |            (Optional PW adaptation function)            |
        |                                                         |
        |<------------------- Emulated Service ------------------>|
```

Figure 3: PW Control Plane Switching Reference Model

In Figure 3 PE2 runs two separate control planes: one toward PE1, and one Toward PE3. The PW switching point is at PE2 which is configured to connect PW1 and PW3 together to complete the multi-hop PW between PE1 and PE3. PW1 and PW3 MUST be of the same PW type, but PSN1 and PSN2 need not be the same technology. In the latter case, if the PW is switched to a different technology, the PEs must adapt the PDU encapsulation between the different PSN technologies. In the case where PSN1 and PSN2 are the same technology the PW PDU does not need to be modified, and PDUs are then switched between the pseudo-wires at the PW label level.

It should be noted that it is possible to adapt one PSN technology to a different one, for example MPLS over an IP or GRE [MPLS-IP-GRE] encapsulation, but this is outside the scope of this document. Further, one could perform an interworking function on the PWs themselves at the PW switching point, allowing conversion from one PW type to another, but this is also outside the scope of this document.

The pseudowire switching methodology described in this document assumes manual configuration of the switching point at PE2. It should also be noted that a PW can traverse multiple PW switching points along it's path, and the edge PEs will not require any specific knowledge of how many PW switching points the PW has traversed (though this may be reported for troubleshooting purposes).

In general the approach taken is to connect the individual control planes by passing along any signaling information immediately upon

reception. First the PW switching point is configured to switch a SH-PW from a specific peer to another SH-PW destined for a different peer. No control messages are exchanged yet as the PW switching point PE does not have enough information to actually initiate the PW setup messages. However, if a session does not already exist, a control protocol (LDP/L2TP) session is setup. In this model the MH-PW setup is starting from the U-PE devices. Next once the U-PE is configured it sends the PW control setup messages. These messages are received, and immediately used to form the PW setup messages for the next SH-PW of the MH-PW. If one of the Switching PEs doesn't accept an LDP Setup message then a Label release message is sent back to the originator U-PE. A MH-PW is declared UP when all the constituent SH-PWs are UP.

5. PW Switching with Attachment Circuits

In PW switching with attachment circuits, each PSN may be of a different type e.g. MPLS, L2TPv3. However the details of this are outside the scope of this document. The PWs and the attachment circuits MUST be of the same type e.g. ATM, Ethernet, etc.

The PWs in each PSN are established independently, with each PSN being treated as a separate PWE3 domain. For example, in Figure 2 for case of MPLS PSNs, PW1 is setup between PE1 and PE2 using the LDP targeted session as described in [PWE3-MPLS], and at the same time a separate pseudo wire, PW2, is setup between PE3 and PE4. The ACs for PW1 and PW2 at PE2 and PE3 MUST be configured such that they are the same native circuit e.g. ATM VCC, Ethernet VLAN, etc. These ACs thus connect the PWs in PSN1 and PSN2 together.

6. PW-MPLS to PW-MPLS Control Plane Switching

Referencing Figure 3, PE2 sets up a PW1 using the LDP targeted session as described in [PWE3-MPLS], at the same time a separate pseudo wire PW3 is setup to PE3. Each PW is configured independently on the PEs, but on PE2 pseudo wire PW1 is connected to pseudo wire PW3. PDUs are then switched between the pseudo-wires at the PW label level. Hence the data plane does not need any special knowledge of the specific pseudo wire type. A simple standard MPLS label swap operation is sufficient to connect the two PWs, and in this case the PW adaptation function is not used.

This process can be repeated as many times as necessary, the only limitation to the number of PW switching points traversed is imposed by the TTL field of the PW MPLS Label. The setting of the TTL is a matter of local policy on the originating PE, but SHOULD be set to

255.

6.1. MPLS PW Control Plane Switching

There are three MPLS to MPLS PW control planes:
- -i. Static configuration of the PW.
- -ii. LDP using FEC 128
- -iii. LDP using the generalized FEC 129

This results in four distinct PW switching situations that are significantly different, and must be considered in detail:
- -i. PW Switching between two static control planes.
- -ii. Static Control plane switching to LDP dynamic control plane.
- -iii. Two LDP control planes using the same FEC type
- -iv. LDP using FEC 128, to LDP using the generalized FEC 129

6.1.1. Static Control plane switching

In the case of two static control planes the PW switching point MUST be configured to direct the MPLS packets from one PW into the other. There is no control protocol involved in this case. When one of the control planes is a simple static PW configuration and the other control plane is a dynamic LDP FEC 128 or generalized PW FEC, then the static control plane should be considered identical to an attachment circuit (AC) in the reference model of Figure 1. The switching point PE SHOULD signal the proper PW status if it detects a failure in sending or receiving packets over the static PW. Because the PW is statically configured, the status communicated to the dynamic LDP PW will be limited to local interface failures. In this case, the PW switching point PE behaves in a very similar manner to a U-PE, assuming an active role.

6.1.2. Two LDP control planes using the same FEC type

As stated in a section above, the PW switching point PE should assume an initial passive role. This means that once independent PWs are configured on the switching point, the LSR does not advertise the LDP PW FEC mapping until

it has received at least one of the two PW LDP FECs from a remote PE. This is necessary because the switching point LSR does not know a priory what the interface parameter field in the initial FEC advertisement will contain.

The PWID is a unique number between each pair of PEs. Hence Each SH-PW that forms an MH-PW may have a different PWID. In the case of The

Generalized PW FEC, the AGI/SAI/TAI may have to also be different for some, or sometimes all, SH-PWs.

6.1.3. LDP using FEC 128 to LDP using the generalized FEC 129

When a PE is using the generalized FEC 129, there are two distinct roles that a PE can assume: active and passive. A PE that assumes the active role will send the LDP PW setup message, while a passive role PE will simply reply to an incoming LDP PW setup message. The PW switching point PE, will always remain passive until a PWID FEC 128 LDP message is received, which will cause the corresponding generalized PW FEC LDP message to be formed and sent. If a generalized FEC PW LDP message is received while the switching point PE is in a passive role, the corresponding PW FEC 128 LDP message will be formed and sent.

PWIDs need to be mapped to the corresponding AGI/TAI/SAI and vice versa. This can be accomplished by local PW switching point configuration, or by some other means, such as some form of auto discovery. Such other means are outside the scope of this document.

6.1.4. PW switching point TLV

The edge to edge PW might traverse several switching points, in separate administrative domains. For management and troubleshooting reasons it is useful to record all the switching points that the PW traverses. This is accomplished by using a PW switching point TLV.

The OPTIONAL PW switching point TLV is appended to the PW FEC at each switching point and is encoded as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|0|     PW sw TLV  (0x096D)   |     PW sw TLV  Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |    Length     |    Variable Length Value      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Variable Length Value                 |
|                             "                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

[note] LDP TLV type is pending IANA approval.

- PW sw TLV Length

  Specifies the total length of all the following PW switching point TLV fields in octets

- Type

  Encodes how the Value field is to be interpreted.

- Length

  Specifies the length of the Value field in octets.

- Value

  Octet string of Length octets that encodes information to be interpreted as specified by the Type field.

PW Switching point TLV Types are assigned by IANA according the the process defined in the "IANA Allocations" section below.

The PW switching Point TLV is an OPTIONAL TLV that can appear once for each switching point traversed.

6.1.5. Adaptation of Interface Parameters

[PWE3-MPLS] defines several interface parameters, which are used by the Network Service Processing (NSP) to adapt the PW to the Attachment Circuit (AC). The interface parameters are only used at the end points, and MUST be passed unchanged across the PW switching point. However the following interface parameters MAY be modified as follows:

- 0x03 Optional Interface Description string This Interface parameter MAY be modified, or altogether removed from the FEC element depending on local configuration policies.

- 0x09 Fragmentation indicator This parameter MAY be inserted in the FEC by the switching point if it is capable of re-assembly of fragmented PW frames according to [PWE3-FRAG].

- 0x0C VCCV parameter The switching point MAY not be able to inspect the VCCV control channel. In this case the Control Channel Type indicator MUST be modified to reflect the capability of the PE acting as the PW switching point.

6.1.6. Group ID

The Group ID (GR ID) is used to reduce the number of status messages that need to be sent by the PE advertising the PW FEC. The GR ID has local significance only, and therefore MUST be mapped to a unique GR ID allocated by the PW switching point PE.

6.1.7. PW Loop Detection

A switching point PE SHOULD check the OPTIONAL PW switching Point TLV, to verify if it's own IP address appears in it. If it's IP address appears in a received PW switching Point TLV, the PE SHOULD break the loop, and send a label release message with the following error code:

```
Assignment E Description
0x0000003A 0 "PW Loop Detected"
```

[note: error code pending IANA allocation]

7. PW-MPLS to PW-L2TPv3 Control Plane Switching

Both MPLS and L2TPv3 PWs may be static or dynamic. This results in four possibilities when switching between L2TPv3 and MPLS.

- -i. Switching between MPLS and L2TPv3 static control planes.
- -ii. Switching between a static MPLS PW and a dynamic L2TPv3 PW.
- -iii. Switching between a static L2TPv3 PW and a dynamic MPLS PW.
- -iv. Switching between a dynamic MPLS PW and a dynamic L2TPv3 PW.

7.1. Static MPLS PW and Dynamic L2TPv3 PW

When a statically configured MPLS PW is switched to a dynamic L2TPv3 PW, the static control plane should be considered identical to an attachment circuit (AC) in the reference model of Figure 1. The switching point PE SHOULD signal the proper PW status if it detects a failure in

sending or receiving packets over the static PW. Because the PW is statically configured, the status communicated to the dynamic L2TPv3 PW will be limited to local interface failures. In this case, the PW switching point PE behaves in a very similar manner to a U-PE, assuming an active role.

7.2. Static L2TPv3 PW and Dynamic LDP/MPLS PW

When a statically configured L2TPv3 PW is switched to a dynamic LDP/MPLS PW, then the static control plane should be considered identical to an attachment circuit (AC) in the reference model of Figure 1. The switching point PE SHOULD signal the proper PW status (via an L2TPv3 SLI message) if it detects a failure in sending or receiving packets over the static PW. Because the PW is statically configured, the status communicated to the dynamic LDP/MPLS PW will be limited to local interface failures. In this case, the PW switching point PE behaves in a very similar manner to a U-PE, assuming an active role.

7.3. Dynamic LDP/MPLS and L2TPv3 PWs

When switching between dynamic PWs, the switching point always assumes an initial passive role. Thus, it does not initiate an LDP/MPLS or L2TPv3 PW until it has received a connection request (Label Mapping or ICRQ) from one side of the node. Note that while MPLS PWs are made up of two unidirectional LSPs bonded together by FEC identifiers, L2TPv3 PWs are bidirectional in nature, setup via a 3-message exchange (ICRQ, ICRP and ICCN). Details of Session Establishment, Teardown, and PW Status signaling are detailed below.

7.3.1. Session Establishment

When the PW switching point receives an L2TPv3 ICRQ message, the identifying AVPs included in the message are mapped to FEC identifiers and sent in an LDP label mapping message. Conversely, if an LDP Label Mapping message is received, it is either mapped to an ICRP message or causes an L2TPv3 session to be initiated by sending an ICRQ.

Following are two example exchanges of messages between LDP and L2TPv3. The first is a case where an L2TPv3 U-PE initiates an MH-PW, the second is a case where an MPLS U-PE initiates an MH-PW.

```
      PE 1 (L2TPv3)      PW Switching Node       PE3 (MPLS/LDP)

        AC "Up"
        L2TPv3 ICRQ --->
                         LDP Label Mapping  --->
                                                    AC "UP"
                                             <--- LDP Label Mapping
                   <--- L2TPv3 ICRP
        L2TPv3 ICCN  --->
      <-------------------- MH PW Established ------------------>


      PE 1 (MPLS/LDP)      PW Switching Node       PE3 (L2TPv3)

        AC "Up"
        LDP Label Mapping --->
                              L2TPv3 ICRQ  --->
                                                <--- L2TPv3 ICRP
                         <--- LDP Label Mapping
                              L2TPv3 ICCN --->
                                                   AC "Up"
      <-------------------- MH PW Established ------------------>
```

7.3.2. Adaptation of PW Status message

L2TPv3 uses the SLI message to indicate a interface status change (such as the interface transitioning from "Up" or "Down"). MPLS/LDP PWs either signal this via an LDP Label Withdraw or the PW Status message defined in section 5.3.2 of [PWE3-MPLS].

7.3.3. Session Teardown

L2TPv3 uses a single message, CDN, to tear down a pseudowire. The CDN message translates to a Label Withdraw message in LDP. Following are two example exchanges of messages between LDP and L2TPv3. The first is a case where an L2TPv3 U-PE initiates the termination of an MH-PW, the second is a case where an MPLS U-PE initiates the termination of an MH-PW.

```
   PE 1 (L2TPv3)      PW Switching Node       PE3 (MPLS/LDP)

   AC "Down"
     L2TPv3 CDN --->
                      LDP Label Withdraw  --->
                                                 AC "Down"

   <--------------- MH PW Data Path Down ------------------>

                                         <--- LDP Label Withdraw *

     * This LDP Label Withdraw is not necessary to break the end-to-end
       MH PW data path.

   PE 1 (MPLS LDP)     PW Switching Node       PE3 (L2TPv3)

   AC "Down"
   LDP Label Withdraw  --->
                            L2TPv3 CDN -->
                                                 AC "Down"

   <---------------- MH PW Data Path Down ------------------>

                <---   LDP Label Withdraw *
     * This LDP Label Withdraw is not necessary to break the end-to-end
       MH PW data path.
```

7.4. Adaptation of LDP/L2TPv3 AVPs and Interface Parameters

[PWE3-MPLS] defines several interface parameters which must be mapped to similar AVPs in L2TPv3 setup messages.

* Interface MTU

  The Interface MTU parameter is mapped directly to the L2TP Interface MTU AVP defined in [L2TP-L2VPN]

* Max Number of Concatenated ATM cells

  This interface parameter is mapped directly to the L2TP "ATM Maximum Concatenated Cells AVP" described in section 6 of [L2TP-ATM].

* Optional Interface Description String

  This string may be carried as the "Call-Information AVP" described in section 2.2 of [L2TP-INFOMSG]

* PW Type

  The PW Type defined in [PWE3-IANA] is mapped to the L2TPv3 "PW Type" AVP defined in [L2TPv3].

* PW ID (FEC 128)

  For FEC 128, the PW ID is mapped directly to the L2TPv3 "Remote End ID" AVP defined in [L2TPv3].

* Generalized FEC 129 SAI/TAI

  Section 4.3 of [L2TP-L2VPN] defines how to encode the SAI and TAI parameters. These can be mapped directly.

Other interface parameter mappings will either be defined in a future version of this document, or are unsupported when switching between LDP/MPLS and L2TPv3 PWs.

7.5. Switching Point TLV in L2TPv3

When translating between LDP and L2TPv3 control messages, the PW Switching Point TLV described earlier in this document is carried in a single variable length L2TP AVP present in the ICRQ, ICRP messages, and optionally in the ICCN message.

The L2TP "Switching Point AVP" is Attribute Type TBA-L2TP-AVP-1. The AVP MAY be hidden (the L2TP AVP H-bit may be 0 or 1), the length of the AVP is 6 plus the length of the series of Switching Point sub-TLVs included in the AVP, and the AVP MUST NOT be marked Mandatory (the L2TP AVP M-bit MUST be 0).

7.6. L2TPv3 and MPLS PW Data Plane

When switching between an MPLS and L2TP PW, packets are sent in their entirety from one PW to the other, replacing the MPLS label stack with the L2TPv3 and IP header or vice versa. There are some situations where an additional amount of interworking must be provided between the two data planes at the PW switching node.

7.6.1. PWE3 Payload Convergence and Sequencing

Section 5.4 of [PWE3-ARCH] discusses the purpose of the various shim headers necessary for enabling a pseudowire over an IP or MPLS PSN. For L2TPv3, the Payload Convergence and Sequencing function is carried out via the Default L2-Specific Sublayer defined in [L2TPv3]. For MPLS, these two functions (together with PSN Convergence) are carried out via the MPLS Control Word. Since these functions are different between MPLS and L2TPv3, interworking between the two may be necessary.

The L2TP L2-Specific Sublayer and MPLS Control Word are shim headers which in some cases are not necessary to be present at all. For example, an Ethernet PW with sequencing disabled will generally not require an MPLS Control Word or L2TP Default L2-Specific Sublayer to be present at all. In this case, Ethernet frames are simply sent from one PW to the other without any modification beyond the MPLS and L2TP/IP encapsulation and decapsulation.

The following section offers guidelines for how to interwork between L2TP and MPLS for those cases where the Payload Convergence, Sequencing, or PSN Convergence functions are necessary on one or both sides of the switching node.

7.6.2. Mapping

The MPLS Control Word consists of (from left to right):

-i. These bits are always zero in MPLS are not necessary to be mapped to L2TP.

-ii. These six bits may be used for Payload Convergence depending on the PW type. For ATM, the first four of these bits are defined in [PWE3-ATM]. These map directly to the bits defined in [L2TP-ATM]. For Frame Relay, these bits indicate how to set the bits in the Frame Relay header which must be regenerated for L2TP as it carries the Frame Relay header intact.

-iii. L2TP determines its payload length from IP. Thus, this Length field need not be carried directly to L2TP. This Length field will have to be calculated and inserted for MPLS when necessary.

-iv. The Default L2-Specific Sublayer has a sequence number with different semantics than that of the MPLS Control Word. This difference eliminates the possibility of supporting sequencing across the MH-PW by simply carrying the sequence number through the switching point transparently. As such, sequence numbers MAY be supported by checking the sequence numbers of packets arriving at the switching point and regenerating a new sequence number in the proper format for the PW on egress. If this type of sequence interworking at the switching node is not supported, and a U-PE requests sequencing of all packets via the L2TP control channel during session setup, the switching node SHOULD NOT allow the session to be established by sending a CDN message with Result Code set to 17 "sequencing not supported" (subject to IANA Assignment).

8. Operation And Management

8.1. Pseudo Wire Status

In the PW switching with attachment circuits case (Figure 2), PW status messages indicating PW or attachment circuit faults SHOULD be mapped to fault indications or OAM messages on the connecting AC as defined in [PW-MSG-MAP]. If the AC connecting two PWs crosses an administrative boundary, then the manner in which those OAM messages are treated at the boundary is out of scope of this draft.

In the PW control plane switching case (Figure 3), there is no attachment circuit at the PW switching point, but the two PWs are connected together. Similarly, the status of the PWs are forwarded unchanged from one PW to the other by the control plane switching function. However, it may sometimes be necessary to communicate status of one of the locally attached SH-PW at a PW switching point. For LDP this can be accomplished by sending an LDP status notification message containing the PW status TLV, as well as an OPTIONAL PW switching point TLV.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |0|   Notification   (0x0001)   |      Message Length           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                           Message ID                          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |0|1| Status (0x0300)           |      Length                   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |0|1|                Status Code=0x00000028                     |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                     Message ID=0                              |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |      Message Type=0           |      PW Status TLV            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                         PW Status TLV                         |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |         PW Status TLV         |            PWId FEC           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    |                                                               |
    |                 PWId FEC or Generalized ID FEC                |
    |                                                               |
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |1|0|     PW sw TLV  (0x096D)   |     PW sw TLV  Length         |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |     Type      |    Length     |    Variable Length Value      |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Only one PW switching point TLV can be present in this message. This message is then relayed by each PW switching point unchanged. The U-PE decodes the status message and the included PW switching point TLV to detect where exactly the fault occurred. At the U-PE if there is no PW switching point TLV included in the LDP status notification then the status message can be assumed to have originated at the remote U-PE. It should also be noted that once a PW status notification message is initiated at a PW switching point, any further status message received from an upstream neighbor is processed locally and not forwarded until the PW switching point original status message is cleared. When a PW status event, for a particular PW, occurs at the S-PE, the appropriate PW status notification MUST be send toward both remote S-PEs or U-PEs attached to the PW.

8.2. Pseudowire Status Negotiation Procedures

Pseudowire Status signaling methodology, defined in [PWE3-MPLS], SHOULD be transparent to the switching point.

8.3. Status Dampening

When the PW control plane switching methodology is used to cross an administrative boundary it might be necessary to prevent excessive status signaling changes from being propagated across the administrative boundary. This can be achieved by using a similar method as commonly employed for the BGP protocol route advertisement dampening. The details of this OPTIONAL algorithm are a matter of implementation, and are outside the scope of this document.

9. Peering Between Autonomous Systems

The procedures outlined in this document can be employed to provision and manage MH-PWs crossing AS boundaries.

The use of more advanced mechanisms involving auto-discovery and ordered PWE3 MH-PW signaling will be covered in a separate document.

10. Security Considerations

This document specifies the LDP and L2TPv3 extensions that are needed for setting up and maintaining Pseudowires. The purpose of setting up Pseudowires is to enable layer 2 frames to be encapsulated and transmitted from one end of a Pseudowire to the other. Therefore we treat the security considerations for both the data plane and the control plane.

10.1. Data Plane Security

Data plane security consideration as discussed in [PWE3-MPLS], [L2TPv3], and [PWE3-ARCH] apply to this extension without any changes.

10.2. Control Protocol Security

General security considerations with regard to the use of LDP are specified in section 5 of RFC 3036. Security considerations with regard to the L2TPv3 control plane are specified in [L2TPv3]. These considerations apply as well to the case where LDP or L2TPv3 is used to set up PWs.

A Pseudowire connects two attachment circuits. It is important to make sure that LDP connections are not arbitrarily accepted from anywhere, or else a local attachment circuit might get connected to an arbitrary remote attachment circuit. Therefore an incoming session request MUST NOT be accepted unless its IP source address is known to be the source of an "eligible" peer. The set of eligible peers could be pre-configured (either as a list of IP addresses, or as a list of address/mask combinations), or it could be discovered dynamically via an auto-discovery protocol which is itself trusted. (Obviously if the auto-discovery protocol were not trusted, the set of "eligible peers" it produces could not be trusted.)

Even if a connection request appears to come from an eligible peer, its source address may have been spoofed. So some means of preventing source address spoofing must be in place. For example, if all the eligible peers are in the same network, source address filtering at the border routers of that network could eliminate the possibility of source address spoofing.

For a greater degree of security, the LDP MD5 authentication key option, as described in section 2.9 of RFC 3036, or the Control Message Authentication option of [L2TPv3] MAY be used. This provides integrity and authentication for the control messages, and eliminates the possibility of source address spoofing. Use of the message authentication option does not provide privacy, but privacy of control messages are not usually considered to be highly urgent. Both the LDP and L2TPv3 message authentication options rely on the configuration of pre-shared keys, making it difficult to deploy when the set of eligible neighbors is determined by an auto-configuration protocol.

When the Generalized ID FEC Element is used, it is possible that a particular peer may be one of the eligible peers, but may not be the right one to connect to the particular attachment circuit identified by the particular instance of the Generalized ID FEC element. However, given that the peer is known to be one of the eligible peers (as discussed above), this would be the result of a configuration error, rather than a security problem. Nevertheless, it may be advisable for a PE to associate each of its local attachment circuits with a set of eligible peers, rather than having just a single set of

eligible peers associated with the PE as a whole.

11. IANA Considerations

11.1. L2TPv3 AVP

This document uses a ne L2TP parameter, IANA already maintains a registry of name "Control Message Attribute Value Pair" defined by [RFC3438]. The following new values are required:

TBA-L2TP-AVP-1 - PW Switching Point AVP

11.2. LDP TLV TYPE

This document uses several new LDP TLV types, IANA already maintains a registry of name "TLV TYPE NAME SPACE" defined by RFC3036. The following value is suggested for assignment:

```
TLV type  Description
 0x096D   Pseudo Wire Switching TLV
```

11.3. LDP Status Codes

This document uses several new LDP status codes, IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by RFC3036. The following value is suggested for assignment:

```
Assignment E Description
0x0000003A 0 "PW Loop Detected"
```

11.4. L2TPv3 Result Codes

This document uses several new LDP status codes, IANA already maintains a registry of name "L2TPv3 Result Codes" defined by RFC3438. The following value is suggested for assignment:

```
Assignment  Description
    17      "sequencing not supported"
```

11.5. New IANA Registries

IANA needs to set up a registry of "PW Switching Point TLV Type". These are 8-bit values. Types value 1 through 3 are defined in this document. Type values 4 through 64 are to be assigned by IANA using the "Expert Review" policy defined in RFC2434. AGI Type values 65 through 127, 0 and 255 are to be allocated using the IETF consensus policy defined in [RFC2434]. AGI types values 128 through 254 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC2434.

The Type Values are assigned as follows:

Type	Length	Description
0x01	4	PW ID of last PW traversed
0x02	variable	PW Switching Point description string
0x03	4	IP address of PW Switching Point (Optional)

12. Intellectual Property Statement

13. Full Copyright Statement

14. Acknowledgments

The authors wish to acknowledge the contributions of Wei Luo, and Skip Booth.

15. Normative References

[PWE3-MPLS] "Transport of Layer 2 Frames Over MPLS", Martini, L., et al.,draft-ietf-pwe3-control-protocol-16.txt, (work in progress), May 2005.

[2547BIS] "BGP/MPLS IP VPNs", Rosen, E, Rekhter, Y. draft-ietf-l3vpn-rfc2547bis-03.txt (work in progress), October 2004.

[L2TPv3] "Layer Two Tunneling Protocol (Version 3)", J. Lau, M. Townsley, I. Goyret, RFC3931

[RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations section in RFCs", BCP 26, RFC 2434, October 1998.

[RFC2277] Alvestrand, H., "IETF Policy on Character Sets and Languages", BCP 18, RFC 2277, January 1998.

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[BCP79] S. Bradner, Ed., "Intellectual Property Rights in IETF Technology", BCP 79, RFC 3979, March 2005.

[BCP78] S. Bradner, Ed., "IETF Rights in Contributions", BCP 78, RFC 3978, March 2005.

16. Informative References

[MPLS-IP-GRE] "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", Rosen, E, Rekhter, Y. draft-ietf-mpls-in-ip-or-gre-08.txt, (work in progress), December 2004.

[PWE3-ARCH] "PWE3 Architecture" Bryant, et al., draft-ietf-pwe3-arch-07.txt (work in progress), June 2003.

[PWE3-FRAG] "PWE3 Fragmentation and Reassembly", A. Malis, W. M. Townsley, draft-ietf-pwe3-fragmentation-05.txt (work in progress) February 2004

[PWE-IANA] "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)" Martini, Townsley, draft-ietf-pwe3-iana-allocation-04.txt (work in progress), April 2004

[L2TP-L2VPN] "L2VPN Extensions for L2TP", Luo, Wei, draft-ietf-l2tpext-l2vpn-00.txt, (work in progress), Jan 2004

[L2TP-INFOMSG] "L2TP Call Information Messages", Mistretta, Goyret, McGill, Townsley, draft-mistretta-l2tp-infomsg-02.txt, (work in progress), July 2004

[L2TP-ATM] "ATM Pseudo-Wire Extensions for L2TP", Singh, Townsley, Lau, draft-ietf-l2tpext-pwe3-atm-00.txt, (work in progress), March 2004.

[PWE3-ATM] "Encapsulation Methods for Transport of ATM Over IP and MPLS Networks", Martini, Rosen, Bocci, "draft-ietf-pwe3-atm-encap-05.txt", (work in progress), April 2004.

[RFC3438] W. M. Townsley, "Layer Two Tunneling Protocol (L2TP) Internet"

[BCP68] Assigned Numbers Authority (IANA) Considerations Update", RFC 3438, BCP 68, November 2002.

[PW-MSG-MAP] "Pseudo Wire (PW) OAM Message Mapping", Nadeau et al, draft-ietf-pwe3-oam-msg-map-02.txt, (work in progress), February 2005

17. Author Information

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO, 80112
e-mail: lmartini@cisco.com

Thomas D. Nadeau
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough, MA 01719
e-mail: tnadeau@cisco.com

Chris Metz
Cisco Systems, Inc.
e-mail: chmetz@cisco.com

Mike Duckett
Bellsouth
Lindbergh Center
D481
575 Morosgo Dr
Atlanta, GA 30324
e-mail: mduckett@bellsouth.net

Vasile Radoaca
e-mail: radoaca@hotmail.com

Matthew Bocci
Alcatel
Grove House, Waltham Road Rd
White Waltham, Berks, UK. SL6 3TN
e-mail: matthew.bocci@alcatel.co.uk

Florin Balus
Nortel Networks
3500 Carling Ave.
Ottawa, Ontario, CANADA
e-mail: balus@nortel.com