PWE3 Working Group
Internet Draft
Expires: January 2006

David McDysan
MCI

Yuichiro Wada
NTT Communications

Mike Duckett
Bellsouth

Yeongil Seo
Korea Telecom

Chris Metz
Cisco Systems

Ping Pan
Hammerhead Systems

Florin Balus
Mike Loomis
Jeff Sugimoto
Nortel

Andy Malis
Tellabs

Paul Doolan
Mangrove Systems

Prayson Pate
Overture Networks

Vasile Radoaca
Consultant

July 2005

Multi-Segment Pseudowire Setup and Maintenance using LDP

draft-balus-mh-pw-control-protocol-02.txt

Status of this Memo

Abstract

[MS PWE3 Requirements] describes the requirements to allow a service provider to extend the reach of pseudo-wires across multiple domains. A Multi-Segment PW is defined as a set of two or more contiguous PW segments that behave and function as a single point-to-point PW.

The current specification of the PW Architecture [PW ARCH] defines the PW as a single Segment entity, connecting the Attachment Circuits between two Ultimate PEs (U-PE). The current procedures for establishing a single segment PW (SS-PW) is described in [PW Control], where typically an LDP session is established between the ultimate PEs handling the Pseudowire End Service (PWES). No intermediate nodes, between the PEs, are aware of the PW.

The purpose of this draft is to specify new LDP extensions, end to end signaling procedures to address the related requirements specified in [MS-PWE3 Requirements]. The proposed procedures follow the guidelines defined in [RFC3036bis] and enable the usage of addressing schemes (L2FECs) and other TLVs already defined for PWs in [PW Control].

The solution described in the draft provides a MS-PW Operational Model, Signaling Procedures consistent with the regular (SS-)PWs, in order to enable seamless implementation, deployment. The resulting MS-PW building blocks accommodate and enhance LDP-VPLS, VPWS solutions with minimal changes in the Information Models and Software Modules related to the L2VPN functionality.

Table of Contents

1. Terminology

The terminology used in this document is consistent with the terminology used in [MS PWE3 Requirements]:

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

. Single-Segment PW(SS-PW). A PW setup directly between two U-PE devices. Each LSP in one direction of a SS-PW traverses one PSN tunnel that connects the two U-PEs.

. Multi-Segment PW (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 an U-PE.

. 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. It is therefore a PW switching point for a MS-PW. A PW Switching Point is never both U-PE and S-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 ultimate PEs.

. PW Segment. A part of a Single-Segment or Multi-Segment PW, which is set up between two adjacent PE devices, U-PEs and/or S-PEs.

. Extended LDP session (E-LDP). An LDP session established using targeted discovery mode [RFC3036bis]

2. Introduction and Scope

[MS PWE3 Requirements] describes the requirements to allow a service provider to extend the reach of pseudo-wires across multiple domains. A MS-PW is defined as a set of two or more contiguous PW segments that behave and function as a single point-to-point PW.

The current specification of the PW Architecture [PW ARCH] defines the PW as a single Segment entity, connecting attachment circuits on exactly two PEs. The current procedures for establishing PWs are described in [PW Control], where typically an LDP session is established between the PEs handling the pseudowire end service (PWES). The LDP session is referred to as "targeted" because it uses a targeted discovery (via hello messages) to establish an LDP session between the two PEs exchanging the PW labels. The tandem nodes between the PEs are unaware of the PW and are only involved with establishing a PSN tunnel between the (U-)PEs.

The purpose of this draft is to specify new LDP extensions and end

to end signaling procedures to address the requirements specified in [MS PWE3 Requirements].

The proposed procedures follow the guidelines defined in [RFC3036bis] and enable the reuse of existing addressing schemes (L2FECs) and other TLVs already defined for SS-PWs in [PW Control].

3. Relevant SS and MS-PW Architectures

The following two figures describe the reference models [MS PWE3 Requirements] to support SS and MS-PW emulated services.

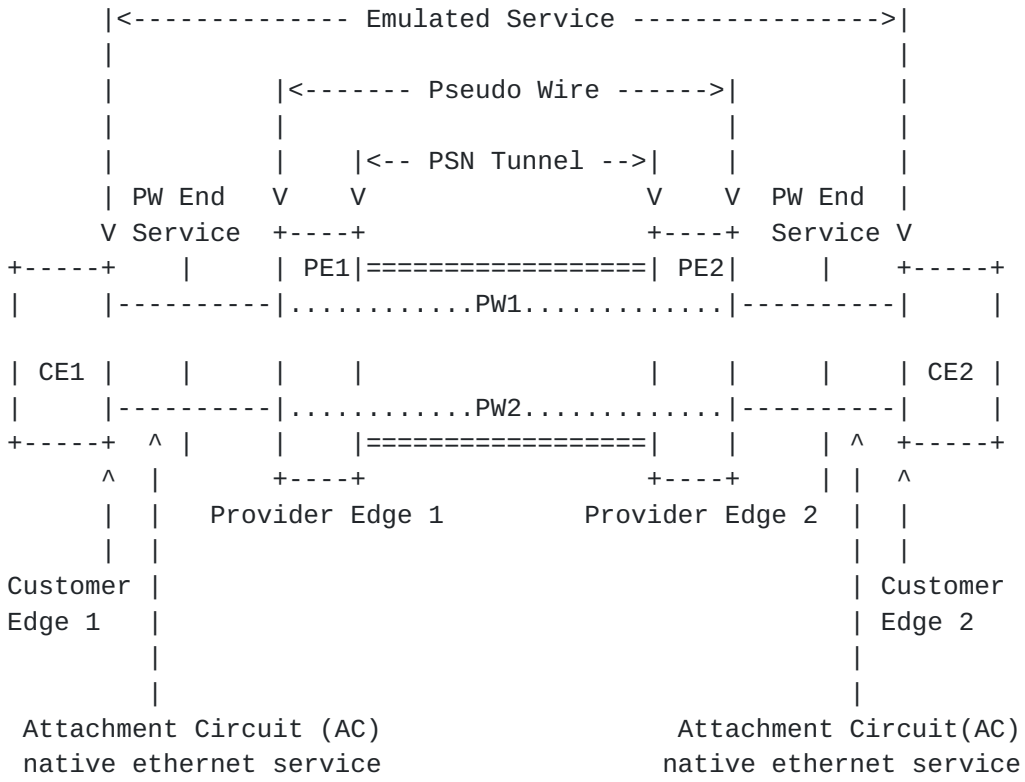
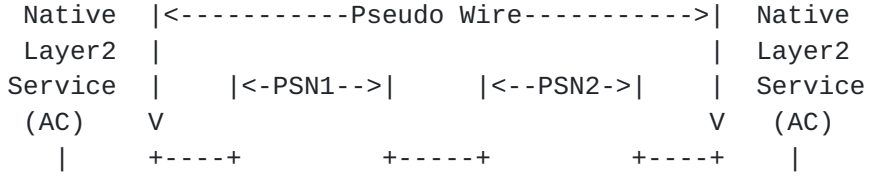


Figure 1: PWE3 Reference Configuration

Figure 1 shows the PWE3 reference architecture [PWE3-ARCH]. 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.



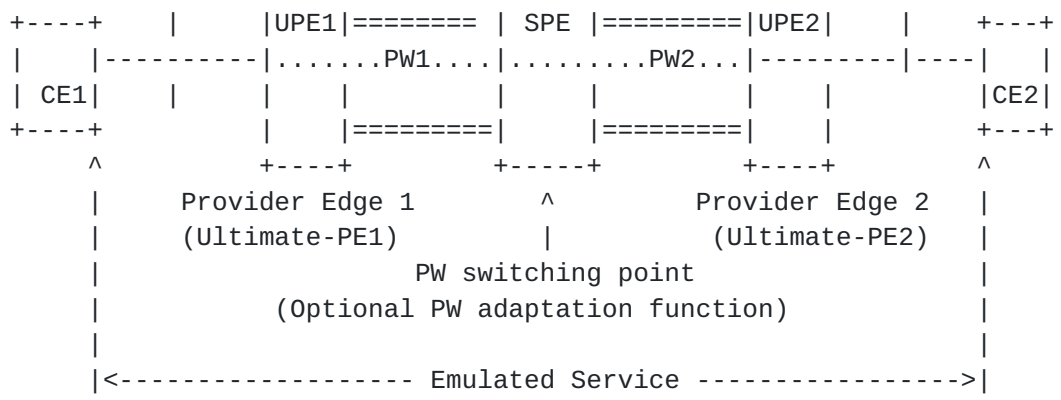


Figure 2: MS-PW Reference Model

Figure 2 extends this architecture to show a Multi-Segment case. UPE1 and UPE2 provide a Pseudowire from CE1 to CE2. Each UPE resides in a different PSN domain. A PSN domain may correspond to a single Provider's network or to a subset of nodes within a Provider network. A PSN tunnel extends from UPE1 to SPE across PSN1, and a second PSN tunnel extends from SPE to UPE2 across PSN2.

PWs are used to connect the Attachment circuits (ACs) attached to UPE1 to the corresponding ACs attached to UPE2. The PW segment on the tunnel across PSN1 is switched to a PW segment in the tunnel across PSN2 at SPE to complete the Multi-Segment PW (MS-PW) between UPE1 and UPE2. S-PE is therefore a PW switching point node and will be referred to as the PW switching provider edge (S-PE). PW segments of the same MS-PW (e.g., PW1 and PW2) MUST be of the same PW type, but PSN tunnels (e.g., PSN1 and PSN2) can be the same or different technology.

Note that although Figure 2 only shows a single S-PE, a PW may transit more than one S-PE along its path.

4. Motivations and Resulting Design Requirements

This section describes the motivations and highlights the architectural objectives of the proposal.

4.1 Satisfy the MS-PW requirements in [MS PWE3 Requirements]

4.1.1 Scalability and Inter-Domain Signaling and Routing

If a MS-PW deployment extends to large and far reaching portions of

one or more networks, mandating an E-LDP session between all switching points of a MS-PW may lead to a control plane scalability issue [MS PWE3 Requirements]. Some network topologies have a natural hierarchy, as described in the use cases section of [MS PWE3 Requirements]. For example, multiple providers who wish to provide PWs that span two or more networks will likely have a relatively small number of gateway nodes as switching points (S-PE) that provide access to a larger number of end nodes (U-PE) forming a hierarchy. As another example, in some MPLS access network topologies, it is foreseeable that thousands or even tens of thousands of U-PE nodes may specify a small number of gateway nodes as switching points (S-PE) for access to the MPLS backbone, breaking the overall MPLS network into a well established hierarchy of MPLS "domains".

In a more generic sense, [MS PWE3 Requirements] discusses a number of cases of a PW Service that has to span multiple domains: e.g. Inter-Provider, Inter-AS (same provider), MAN-WAN. In any of these

cases the interaction between domains is controlled by certain gateways with a specific set of requirements for each individual scenario.

This proposal eliminates the requirement for an E-LDP session between every pair of U-PE nodes for which a PW is required while at the same time preserving the necessary end to end signaling properties. In doing so it alleviates the control plane scalability requirements described in the previous paragraphs. Our proposal enables the end to end PW signaling through a "chain" of (E-)LDP sessions, using a dynamically determined set of S-PEs. If the S-PE and U-PE are identified by IP addresses, then IP routing protocols can distribute information to facilitate dynamic selection of a set of PEs between a Source U-PE and a Destination U-PE based upon parameters (e.g., metric, TE constraints, BGP attributes). U-PE reachability information could be reduced by assignment of IP address prefixes and/or prefix aggregation by a routing protocol.

There could also be some Inter-Provider scenarios where the U-PEs located in a certain Provider domain may not be permitted to communicate directly via an (E)-LDP session to a U-PE in a different domain for operational and security reasons. For other reasons (e.g., security, administrative, etc.) the local U-PE may have no knowledge of the IP address of the remote U-PE. The requirements for these valid scenarios are still being specified and it is not clear whether or not a solution for dynamic end to end signaling is

required or even allowed.

A solution for these scenarios is for further study.

4.1.2 Signaling Requirements

The signaling described in this proposal is based on extensions to [RFC3036bis] and [PW Control]. The new elements (section 6) provide a flexible model that permits interoperability with manual provisioning models, but also enable an end to end MS-PW to be established with minimal number of OSS touches, ideally only one as specified in [MS PWE3 Requirements]. Specifically, the proposal enables the dynamic creation of an end-to-end MS-PW that does not require any manual intervention at the S-PE nodes.

This draft allows for either the same set of S-PE nodes to be traversed in each direction of the MS-PW, or a different set.

[Segmented PW] specifies the case where the set of intermediate S-PEs is manually configured and the PW is stitched at these points by matching the L2FEC for each segment and associating this with the next segment. This case is not precluded by, and could interoperate

with, the method described in this document.

4.2 Operational Consistency with SS-PWs

In a Service Provider network it is understood that SS and MS-PWs will co-exist, possibly for an indefinite amount of time. Furthermore, it is foreseeable that existing SS-PWs may one day be forced to migrate to a MS-PW scenario for a number of reasons. In any case, it should be an advantage to vendors developing PW implementations as well as providers of PW services to minimize the differences between SS and MS-PWs. Operationally, the procedures for identifying (addressing), provisioning and troubleshooting a SS or a MS-PW should be similar.

4.2.1 Service Identification and Provisioning Models

[PW CONTROL] specifies that a PW is uniquely associated with a set of connection identifiers: i.e. PWID (& U-PE pair) for PWID FEC or AGI, AII1, AII2 for the Generalized ID FEC. This proposal reuses the same service identifiers as SS-PW (PWID and Generalized ID FEC) to identify MS-PWs.

From a provisioning perspective, this proposal is consistent with the existing models for SS-PWs. For MS-PWs, both a single ended and double ended model are possible as defined by [L2VPN SIGN], with no user intervention required at any S-PE node.

In a MS-PW scenario, the S-PE nodes are aware of the PW. In the case of PWID addressing, in order to reuse the service identifiers for SS-PWs, the unique association between the U-PE pair and the PWID FEC must be maintained when transiting through the S-PE nodes. In the Generalized ID case a PW is identified by <PE1, <AGI, AII1>, PE2, <AGI, AII2>> in one direction and by <PE2, <AGI, AII2>, PE1, <AGI, AII1>> in the reverse direction [L2VPN SIGN].

This document proposes some extensions to LDP to address the requirements described above for consistent operational model across different PW types. The proposed solution re-uses the same L2FEC definitions as in [PW CONTROL] for identifying the virtual connections and a similar service provisioning model.

The proposal does not preclude the use or support of existing Auto-discovery procedures (e.g. BGP-AD, RADIUS).

4.2.2 OAM

It is important to support the end to end PW OAM concepts already described in [VCCV] and [PW Control]. To meet this requirement, the

S-PE must participate in the negotiation of the PW OAM options and Status TLV.

The current definition of PW OAM functions (e.g. VCCV (LSP-Ping, BFD)) [VCCV] are specified only for operation on a U-PE to U-PE basis. This means that the concatenation of PW switching of S-PEs in MS-PW appears as a PSN tunnel to the PW OAM function.

Support for PW OAM on a U-PE to S-PE, or S-PE to S-PE segment basis, will require changes in the OAM messages and procedures to indicate whether the OAM message is intended for the destination U-PE, intermediate S-PEs, or both.

4.3 Service Resiliency

Several MPLS mechanisms exist today, including procedures defined in [RFC3036bis], [MPLS FRR], [Grace RS] etc. This draft does not

preclude the use of any of these mechanisms.
From a MS-PW perspective, Service Resiliency refers to the ability to choose a backup path in case of failure of the existing MS-PW path (including S-PE failure or any segment failure) [MS PWE3 Requirements].

5. Information Model for Dynamic Signaling of MS-PWs

In the current (SS) PW Architecture (see figure 1), the setup and maintenance of the PW connection is based on a direct, E-LDP Session between PE1 and PE2. As a result of the bidirectional nature of PWs, there is an association between the L2FEC, Source and Destination U-PEs. This association is derived from the information related to the (E-)LDP session between PEs and it is used as part of the end to end message exchange.

In the case of a MS-PW (see figure 2), there is not an E-LDP session between U-PE1 and U-PE2. Instead two LDP Sessions are to be used to establish the MS-PW connection: LDP1 between U-PE1 and S-PE, LDP2 between U-PE2 and S-PE.

The procedures defined in [PW Control] can not be applied to achieve the end to end signaling of the MS-PW. Specifically:

- the identity of the PW endpoints can no longer be derived from the attributes of the local LDP session
- the PWID U-PE pair association is lost. PWID becomes globally unique
- for the Generalized ID the direct association between PW and <<PE1, <AGI, AII1>, PE2, <AGI, AII2>> respectively <PE2, <AGI, AII2>, PE1, <AGI, AII1>> is lost.

- the forwarding of received Label Mapping (LM) messages is not allowed

In order to support dynamic end to end signaling [MS PWE3 Requirements], while maintaining a consistent operational model with SS-PW, there is a need to maintain the relationships between L2FEC and PW endpoints (as discussed above) that are lost when the direct LDP session is not available. This document proposes transporting the address of the Source and Destination U-PEs in the related LDP messages transiting through S-PE node(s). The L2FEC in combination with the source and destination U-PE information form unique PW endpoint identifiers; for example using the GID FEC, the TAI and destination U-PE information will be unique, similarly for the source U-PE and SAI information.

This information could be transported in a number of ways: via new "fields" inserted in the existing Generalized ID FEC or via a new LDP TLV. Choosing one vehicle versus the other is orthogonal to the concepts described in this document as long as the Source and Destination information together with the corresponding L2FEC is explicitly carried in the signaling message and used to identify, route the PW signaling message from source to destination U-PE.

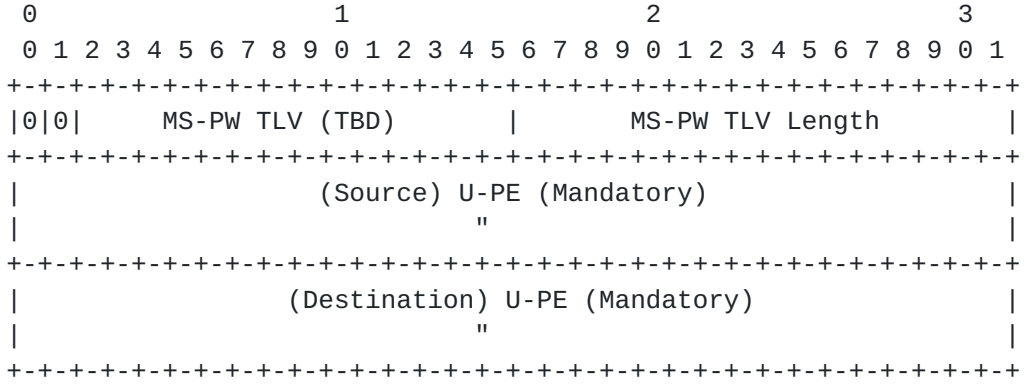
We describe, in section 5.1, the details of the LDP TLV approach as it ensures backwards compatibility with existing deployments, offering support for both PWID and Generalized ID FECs.

Details of the Generalized ID FEC usage is for further study

5.1 MS-PW TLV Design

We are introducing a new TLV, the Multi-Segment PW TLV, which is appended by the Source U-PE to the LDP messages related to a MS-PW.

The following format is being proposed:



- UF bits 00 - U equal 0 means that if the receiving PE does not understand the TLV, a notification must be returned to the message originator and the entire message must be ignored.

- MS-PW TLV (TBD) - To be assigned by IANA. Identifies this TLV as a MS-PW. The presence of this TLV in LDP messages indicates this is a MS-PW.

- MS-PW TLV Length - Specifies the total length in octets of the TLV.

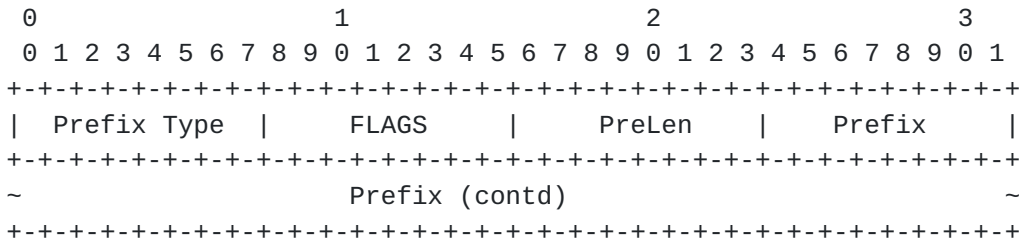
- Source U-PE (Mandatory) - The address of the originating U-PE (e.g. U-PE1). In most of the cases it carries the IP loopback address of the Source U-PE, although other address types - e.g. IPv6, NSAP - could be supported.
This field is used by a MS-PW Network Element for maintaining the uniqueness of PWID FECs and, optionally, in single sided provisioning the discovery of the remote U-PE by the Destination U-PE. When double sided provisioning is used, it is used to verify the remote U-PE against the provisioned value.

- Destination U-PE (Mandatory) - The address of the Destination U-PE (e.g. U-PE2)

Its value could be provisioned at the Source U-PE or is determined as part of the single-sided provisioning behavior [L2VPN SIGN].

The Destination U-PE address field is used to select the next hop through the MS-PW domains.

The basic construct used to carry the Address of the Source and Destination U-PEs is the Prefix Element which is defined below:



- Prefix Type - one octet quantity. It encodes the address type for the address prefix in the Prefix field. Initial formats supported are:
- IPv4 0x01
- IPv6 0x02
- NSAP 0x03

Addition of other formats (or combinations of existing ones) for

further study: for example, AS Numbers, URLs etc.

- FLAGS - One octet field. The field is reserved for future use: i.e. MUST be set to zero when transmitting a message and MUST be ignored at the receiving PE.

- PreLen
One octet unsigned integer containing the length in bits of the address prefix that follows.

- Prefix Value
An address prefix encoded according to the Prefix type field, whose length, in bits, was specified in the PreLen field, padded to a byte boundary.

6. Signaling Procedures

The following are generic procedures for signaling of an MS-PW.

Note that we are using throughout the next sections, examples based on existing IP Loopbacks (as U-PE addresses) and references to IP routing procedures.

According to section 6.1 of [MS PW Requirements]: "MS-PWs are composed of SS-PW, and SS-PW are bi-directional, therefore both directions of a PW segment MUST terminate on the same S-PE/U-PE". In other words both directions of a MS-PW should traverse the same set of S-PEs/U-PEs.

Next section introduces the concepts, procedures that ensure compliance of the solution described in this document with the above requirement. Note that should this requirement change (e.g. "MUST" to "MAY terminate [..]") to enable for diverse S-PEs paths, our solution could accommodate both options.

6.1 Ensuring both MS-PW directions traverse the same U/S-PEs

The proposed procedure is based on an "Ordered" establishment of the individual PW segments that belong to a certain MS-PW. In other words, the signaling is initiated only from the "originating" U-PE node (selected based on provisioned information at nodal/PW level). Any S-PEs or the other U-PE will not initiate a LM Message for the setup of a MS-PW until it receives an incoming LM message for that MS-PW.

We will refer to the direction from the originating U-PE node to the other U-PE node as the "Forward" direction of a MS-PW. The Forward

direction describes the direction of the LDP label mapping messages rather than the direction of the user dataplane. The Reverse

direction is the opposite logic, describing the direction towards the originating U-PE node.

The following section first discusses the signaling in the "Forward" direction followed by a brief description of the deltas in the "Reverse" direction. Note that the flags from the destination U-PE field (see section 5.1) may be used to indicate directionality/ behavior in determining the next-signaling-hop.

6.2 LDP Signaling Walkthrough

The following section focuses on the step by step, generic signaling procedures involved in the setup of a MS-PW. The procedures involved in discovery of Next Signaling Hop are referenced in section 6.4.

1. The PW FEC (PWID or Generalized ID) and Destination U-PE is provisioned on both U-PEs. If single sided provisioning or auto discovery is used, the Destination U-PE needs only to be configured on one of the U-PEs.

2. The originating U-PE builds the MS-PW TLV by inserting its local address in the Source U-PE field and the address of Destination U-PE in the Destination U-PE field. The MS-PW TLV and optional TLVs, (e.g. QOS TLV) are appended to the LM message which is sent to the Next Signaling Hop. The next signaling hop towards the dU-PE can be determined by referencing the PW end point information against the MS-PW information disseminated as per section 6.4.

3. When the next signaling hop receives the LM message, it verifies a PSN tunnel exists to the upstream MS-PW NE. If a PSN tunnel is not available a label release message is sent. However if the S-PE and the next signaling hop are directly connected, with no P device between them, the PSN tunnel may not be necessary [PW Control].

4. OAM parameters (VCCV, Status TLV support) are validated. If the request cannot be supported a label release message is sent to the upstream MS-PW NE.

5. If QoS information* was included in the LM message, the local NE performs a CAC against the selected PSN Tunnel to requesting NE. If the CAC fails a label release message is sent. Alternatively, based on Service Provider choice an increase in the capacity of a PSN tunnel may be tried to accommodate the bandwidth requirements of the MS-PW.

6. If the Destination U-PE address does not equal the MS-PW NE address, a new label mapping message is generated and sent to the Next Signaling Hop, with the original L2FEC and MS-PW TLV, replacing just the value of the service label in the Label TLV with one from its own label space. Note that the S-PE should not comply with the text of section 5.2.3 of [PW Control] i.e. should not initiate a LM message in the opposite direction towards U-PE1 ("Ordered" Mode). Go to step 3.

7. When the Destination U-PE receives the LM message containing the MS-PW TLV (the value from the destination U-PE field matches the address of the local Network Element), it attempts to match the L2 FEC with its local provisioning.
   a) If the L2 FEC and the Source U-PE address do not match the local provisioning, a label release message is sent.

   b) If the L2 FEC is not provisioned, the label maybe retained by virtue of liberal label retention

8. The remaining Destination U-PE processing of the PW label mapping message is as defined in PWE3 control signaling standard [PW Control] (see also tasks outlined in steps 3-5). This completes the Signaling in the Forward direction.

9. MS-PW Signaling in the Reverse direction starts. The (new) source U-PE and subsequently the S-PEs in the Reverse direction will perform the tasks described in steps 2-8.

   The next hop at any S-PE is determined by referencing the LDP sessions used to setup the LSP in the Forward direction. The particular LDP Session is determined using the index (dU-PE, TAI/PWID) information from the LM message received from the Reverse direction. The association between (L2FEC, sU-PE, dU-PE) and the incoming LDP session is stored as the PW Segments are established. This information is always required for further Control Plane Exchanges (e.g. Label Release, PW Status) but is used to also setup the MS-PW in the Reverse direction.

* The term "QoS Information" is used here to mean either one or both of Quantity (e.g. Bandwidth) and/or Quality (e.g. DiffServ) of Service. The detailed definition of the TLVs used to signal this information is outside the scope of this document. Description of possible TLV structures could be found in [TSPEC] and respectively [RFC3270].

6.3 Using common LDP Signaling procedures for MS and SS-PWs

In addition to the OSS and Operational consistency between SS [PW Control] and MS-PWs concepts described in this document, it would be preferable to have consistent procedures used in the Network Elements in order to minimize implementation deltas.

If the U/S-PEs support the signaling procedures described in the previous section for MS-PWs, then these Network Elements could use consistent procedures to establish also SS-PWs between them.

In this context it is important to note that steps 1-9 are the same for both SS/MS-PWs. The only difference between a SS/MS-PW is the amount of times the procedure cycles through steps 3-6: i.e. in the SS-PW case, the first receiving PE (see step 6) will determine that the destination U-PE is itself and the source U-PE is the same with the originator of the LDP session on which the LM message was received. As a result it will run right away through the remaining of the steps (7-9) instead of cycling 1/more times through steps 3-6 as for a MS-PW.

6.4 Determining the Next Signaling Hop

To support end-to-end dynamic signaling of MS-PWs, information must be present in MS-PW aware nodes to support the determination of the next signaling hop. Such information can be provisioned on each MS-PW system or disseminated via regular routing protocols (e.g. BGP).

The following section describes procedures that could be used to "discover" the next-signaling-hop in MS-PW aware systems.

6.4.1 Static Provisioning of the next-signaling-hop

The simplest way to build next-signaling-hop knowledge is by static provisioning. The provisioning of the next-signaling-hop (e.g. S-PE) is similar to the way IP static routes/default gateways are provisioned: e.g. in a U-PE at the nodal level, a default S-PE is provisioned manually when the MS-PW feature is enabled. This can be a simple and effective method, when the network topology is simple and well defined.

As long as the U-PE prefixes from one domain can be summarized the static method could be also expanded to entire domains: i.e. all the U-PEs in one domain being represented by 1/just a few static entries of this sort: U-PE Prefix (47.0.0.0/8), NH = S-PE1. At the other

extreme, when special treatment is required for a certain PW a "fully qualified" entry could be provisioned: e.g. AGI (40),U-PE1 (47.1.1.1), AII (200) -> NH (S-PE1).

Note that static provisioning may be used in combination with dynamic discovery. Indeed, some PW domains may use static provisioning while other PW domains along the multi-hop signaling path may use dynamic discovery within their domain. An example of this scenario is where many U-PEs in a given network will always use a well known primary and backup S-PE "gateways" as the next hop. This S-PE gateway may have many possible S-PE peers and may use a dynamic discovery mechanism to determine the next-signaling-hop of its S-PE peer for a given MS-PW.

6.4.2 "Discovery" Mechanisms for the next-signaling-hop

The next-signaling-hop selection can also be determined by dynamically learning, for each PW Domain, the association between the (Destination U-PE and optionally TAI/PWID) and the next-signaling-hop.

There could be several mechanisms that allow dynamic discovery, advertisement of the next-signaling-hop. The focus of this section is on how this can be accomplished with BGP-based procedures. Note that these procedures may have an end-to-end scope (e.g. Inter-AS Use Case) or may be limited just to the <Core> PW Domain (e.g. MAN-WAN Use Case), depending upon the availability of BGP in the related MS-PW capable nodes.

The signaling procedures described in this draft are compatible and make use of the L2VPN provisioning models and related AD procedures described in [L2VPN SIGN] and respectively [BGP AD].

If the Source U-PE knows apriori the address of the Destination U-PE, there is no need to advertise a "fully qualified" address on a per PW Attachment Circuit. The Destination U-PE may advertise only its Prefix address (and not the Attachment <Circuit> Identifier (AI)) as part of well known BGP auto-discovery procedures - see [BGP AD], [L2VPN SIGN].

As PW Endpoints are provisioned in the U-PEs, the Source U-PE will use this information to obtain the first S-PE hop (i.e., first BGP next hop) where the first PW segment will be established and subsequent S-PEs will use the same information (i.e. the next BGP next-hop(s)) to obtain the next-signaling-hop(s) on the path to the

Destination U-PE.

This is not an exhaustive list, merely examples of how discovery can be accomplished using BGP. It can also be envisioned, in some particular scenarios, that IGP with TE extensions could be used to control the selection of the next-signaling-hop, while avoiding non MS-PW aware devices (e.g. Ps, 2547 PEs).

7. Service Resiliency

With the introduction of dynamic determination of the intermediate S-PEs, this proposal introduces the possibility of end to end (as well as segment) connection resiliency for MS-PWs.

For failures between MS-PW elements, this document does not preclude any existing MPLS failure recovery mechanisms from being used (i.e. [MPLS FRR]).

For failures that prevent one MS-PW system from establishing a PW segment to the succeeding MS-PW system (e.g. U-PE to S-PE), this document adopts the procedures described in section 6 to allow for the dynamic selection of intermediate next hops for the purpose of service resiliency. For example, a source U-PE node can select a candidate S-PE next hop via local preference (or via any other metrics) for the purpose of service resiliency.

Several options are possible for service resiliency and a simple example is provided here, with further optimizations to be explored in future revisions of the document. Existing MPLS or PSN tunnel recovery mechanisms must be attempted before the procedures described below.

As a result of the MS-PW following the same forward and reverse path, we propose that only the upstream node from the failure in the forward path make the next hop selection. This provides consistency with the procedures used to establish the original MS-PW (described in section 6), where the forward path determines the backwards path as well. Recall that each MS-PW system is already aware of the direction of the MS-PW signaling, and its relation to that direction for any particular MS-PW L2 FEC, sU-PE, dU-PE triplet. The MS-PW segments downstream from the failure MUST be released as a new path may be selected that does not overlap with the previous path.

If alternate routing is not possible at the closest MS-PW node upstream from the failure, that node must release the PW segment to the next upstream MS-PW system to attempt additional rerouting.

8. OAM Considerations

This section deals with the Negotiation of the OAM Capabilities described in [VCCV], where the OAM functions (e.g. VCCV (LSP-Ping, BFD)) are specified only for operation on a U-PE to U-PE basis.

Support for PW OAM on a U-PE to S-PE, or S-PE to S-PE segment basis, require changes in the OAM messages and procedures to indicate whether the OAM message is intended for the destination U-PE,

intermediate S-PEs, or both. These changes are for further study.

8.1 MS-PW Capabilities

Common OAM capabilities should be supported on all U-PE and S-PE nodes in the MS-PW. MS-PW takes a least common denominator approach to OAM. The minimum OAM functionality supported on a MS-PW is label withdraw.

8.1.1 PW Status Capability Negotiation

PW Status capability is negotiated across the MS-PW when the MS-PW is first setup. Support for PW status notification is indicated by the presence of the status TLV in the label mapping message.

PW Status capability negotiation at the U-PE occurs as described in [PWE3 CNTL].

It is strongly recommended that MS-PW implement PW status TLV.

8.1.2 VCCV Capability Negotiation

VCCV capability is negotiated across the MS-PW when the MS-PW is first setup. Support for VCCV is indicated by the presence of the VCCV parameter in the interfaces parameter TLV. This parameter is included in the label mapping message within the parameter TLV as described in [VCCV]

VCCV capability negotiation at the U-PE occurs as described in [VCCV]

An S-PE successfully negotiates VCCV capability for the MS-PW when it support VCCV itself and the label mapping messages from its upstream and downstream neighbors indicate support for VCCV for a

given MS-PW FEC.

8.2 PW Status Notification Operation

PW Status notification at the U-PE occurs as described in [PWE3 CNTL].

When an S-PE receives a PW status notification message, the message is processed at the S-PE and propagated down stream along the control path.

8.3 VCCV Operation

VCCV operation at the MS-PW Network Element (NE) occurs as described in [VCCV], with the S-PEs transparently forwarding these messages

towards the destination U-PE.

Support for MS-PW segment OAM, trace-route is for further study.

9. Security Considerations

To be addressed later.

10. IANA Considerations

A new TLV code point needs to be allocated by IANA for MS-PW TLV.

11. Acknowledgements

The editors gratefully acknowledge the following contributors: Luca Martini, Nabil Bitar, Richard Spencer, Simon Delord, Bruce Davie, Elizabeth Hache, Hamid Ould-Brahim, Praveen Muley, Arashmid Akhavain.

12. Appendix: Example of Signaling Procedures

The following section discusses an example of an end to end signaling walkthrough for a MS-PW using the architecture depicted in Figure 2.

Let us assume that Double-sided provisioning and Generalized ID FEC are being used to set up the MS-PW built using segments PW1 and PW3

and using LDP1 and LDP2 sessions.

Here are the required steps:

1. Service Provisioning
   a) at U-PE1: AGI = 40, SAII=100, TAII=200, Remote PE = U-PE2 (IP2 loopback), Origin = Yes
   b) at U-PE2: AGI = 40, SAII=200, TAII=100, Remote PE = U-PE1 (IP1 loopback);

2. The originating U-PE (U-PE1 in our example) builds the MS-PW TLV by inserting its loopback address in the Source U-PE field and the address of U-PE2 in the Destination U-PE field. Next it appends the MS-PW TLV to the label mapping message associating the provisioned FEC information - i.e. (40,100,200) - with the corresponding PW service label.

3. Using the address of Destination U-PE (U-PE2), U-PE1 selects the next signaling hop (S-PE) determined by referencing the PW end point information - IP2,40,200 - against the MS-PW information disseminated as per section 6.4.

4. On receipt of the LM message, S-PE performs the following tasks: Verifies it has a PSN tunnel to U-PE1. If no tunnel is found a label release message is sent.

   a) Verifies it can support the requested OAM parameters (VCCV, Status TLV support). If the request cannot be supported a label release message is sent to U-PE1.

   b) If QoS information* was included in the LM message, it performs a CAC against the selected PSN Tunnel to U-PE1. If the CAC fails a label release message is sent to U-PE1. Alternatively, based on Service Provider choice, an increase in the capacity of the PSN tunnel may be tried to accommodate the bandwidth requirements of the MS-PW.

   c) Checks to see if it is the Destination U-PE by comparing the address within the MS-PW TLV d-UPE field with its own address. If the addresses are not the same, S-PE looks for a next signaling hop to get to U-PE2 - see step 3 above. Then it signals the final segment of the MS-PW by generating and forwarding a new label mapping message to U-PE2, with the original L2FEC (40,100,200) and MS-PW TLV

(IP1, IP2), replacing just the value of the service label in the Label TLV with one from its own label space.

5. When U-PE2 receives the LM message containing the MS-PW TLV, it performs tasks outlined in step 4.

6. U-PE2 then attempts to match the L2 FEC with its local provisioning.
   a) If the FEC information and the U-PE1 address do not match the local provisioning, a label release message is sent.
   b) If the FEC information (40,200,100) is not yet provisioned, the label may be retained by virtue of liberal label retention.

7. The remaining U-PE2 processing of the PW label mapping message is defined in PWE3 control signaling [PW Control].

8. MS-PW Signaling in the Reverse direction U-PE2 to U-PE1 starts. The U-PE2 and subsequently S-PE will perform the tasks described in steps 2-7. The next hop at U-PE2 and S-PE is determined by referencing the LDP sessions used to setup the LSP in the Forward direction. The particular LDP Session is determined in the U-PE2

and S-PE using the index (IP1,TAI(40,100)) from the LM message received from the Reverse direction.

13. Full Copyright Statement

14. Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in

this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an 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.

By submitting this Internet-Draft, I certify that any applicable patent or other IPR claims of which I am aware have been disclosed, or will be disclosed, and any of which I become aware will be disclosed, in accordance with RFC 3668.

15. References

[RFC3036bis] Andersson, Minei, Thomas. "LDP Specification" draft-

ietf-mpls-rfc3036bis-01.txt, IETF Work in Progress, November 2004

[MS PWE3 Requirements] Martini et al. "Requirements for inter domain Pseudo-Wires", draft-ietf-pwe3-ms-pw-requirements-00.txt, IETF Work in Progress, June 2005

[PW Control] Martini et.al. "Pseudowire Setup and Maintenance using LDP", draft-ietf-pwe3-control-protocol-17.txt, IETF Work in Progress, June 2005

[VCCV] Nadeau et.al., "Pseudo Wire (PW) Virtual Circuit Connection Verification (VCCV)", draft-ietf-pwe3-vccv-03.txt, June 2004

[L2VPN SIGN] Rosen et. al. "Provisioning Models and Endpoint Identifiers in L2VPN Signaling", draft-ietf-l2vpn-signaling-03.txt, IETF Work in Progress, February 2005

[Segmented PW] Martini et.al. " Segmented Pseudo Wire", draft-ietf-pwe3-segmented-pw-00.txt, IETF Work in Progress, July 2005

[RFC3270] Le Faucheur, et. al. "MPLS Support of Differentiated

Services", RFC 3270, May 2002

[TSPEC] Wroclawski, J. "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997

16. Authors' Information

Andrew G. Malis
Tellabs, Inc.
2730 Orchard Parkway
San Jose, CA, USA 95134
Email: Andy.Malis@tellabs.com

Chris Metz
Cisco Systems, Inc.
3700 Cisco Way
San Jose, Ca. 95134
Email: chmetz@cisco.com

David McDysan
MCI
22001 Loudoun County Pkwy
Ashburn, VA, USA 20147
dave.mcdysan@mci.com

Florin Balus
Nortel
3500 Carling Ave.
Ottawa, Ontario, CANADA
balus@nortel.com

Jeff Sugimoto
Nortel
3500 Carling Ave.
Ottawa, Ontario, CANADA
sugimoto@nortel.com

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

Mike Loomis
Nortel

600, Technology Park Dr
Billerica, MA, USA
mloomis@nortel.com

Paul Doolan
Mangrove Systems
IO Fairfield Blvd
Wallingford, CT, USA 06492
pdoolan@mangrovesystems.com

Ping Pan
Hammerhead Systems
640 Clyde Court
Mountain View, CA, USA 94043
e-mail: ppan@hammerheadsystems.com

Prayson Pate
Overture Networks, Inc.
507 Airport Blvd, Suite 111
Morrisville, NC, USA 27560
Email: prayson.pate@overturenetworks.com

Yuichiro Wada
NTT Communications
3-20-2 Nishi-Shinjuku, Shinjuke-ku
Tokyo 163-1421, Japan
yuichiro.wada@ntt.com

Yeongil Seo
Korea Telecom Corp.
463-1 Jeonmin-dong, Yusung-gu
Daejeon, Korea
syi1@kt.co.kr