Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 12, 2014

S. Sivabalan
S. Boutros
Cisco Systems, Inc.
H. Shah
Ciena Corp.
S. Aldrin
Huawei Technologies.
February 08, 2014

MAC Address Withdrawal over Static Pseudowire
draft-boutros-pwe3-mpls-tp-mac-wd-03.txt

Abstract

This document specifies a mechanism to signal MAC address withdrawal notification using PW Associated Channel (ACH). Such notification is useful when statically provisioned PWs are deployed in VPLS/H-VPLS environment.

Requirements Language

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 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 12, 2014.

Copyright Notice

Table of Contents

1. Introduction

An LDP-based MAC Address Withdrawal Mechanism is specified in [RFC4762] to remove dynamically learned MAC addresses when the source of those addresses can no longer forward traffic. This is accomplished by sending an LDP Address Withdraw Message with a MAC List TLV containing the MAC addressed to be removed to all other PEs over LDP sessions. When the number of MAC addresses to be removed is large, empty MAC List TLV may be used. [MAC-OPT] describes an optimized MAC withdrawal mechanism which can be used to remove only the set of MAC addresses that need to be re-learned in H-VPLS networks. The solution also provides optimized MAC Withdrawal operations in PBB-VPLS networks.

A PW can be signaled via LDP or can be statically provisioned. In the case of static PW, LDP based MAC withdrawal mechanism cannot be used. This is analogous to the problem and solution described in [RFC4762] where PW OAM message has been introduced to carry PW status TLV using in-band PW Associated Channel. In this document, we propose to use PW OAM message to withdraw MAC address(es) learned via static PW.

2. Terminology

The following terminologies are used in this document:

ACK: Acknowledgement for MAC withdraw message.

LDP: Label Distribution Protocol.

MAC: Media Access Control.

PE: Provide Edge Node.

MPLS: Multi Protocol Label Switching.

PW: PseudoWire.

PW OAM: PW Operations, Administration and Maintenance.

TLV: Type, Length, and Value.

VPLS: Virtual Private LAN Services.

3. MAC Withdraw OAM Message

LDP provides a reliable packet transport for control plackets for dynamic PWs. This can be contrasted with static PWs which rely on re-transmission and acknowledgments (ACK) for reliable OAM packet delivery as described in [RFC6478]. The proposed solution for MAC withdrawal over static PW also relies on re-transmissions and ACKs. However, ACK is mandatory. A given MAC withdrawal notification is sent as a PW OAM message, and the sender keeps re-transmitting the message until it receives an ACK for that message. Once a receiver successfully remove MAC address(es) in response to a MAC address withdraw OAM message, it should not unnecessarily remove MAC address(es) upon getting refresh message(s). To facilitate this, the proposed mechanism uses sequence number, and defines a new TLV to carry the sequence number.

The format of the MAC address withdraw OAM message is shown in Figure 1. The PW OAM message header is exactly the same as what is defined in [RFC6478]. Since the MAC withdrawal PW OAM message is not refreshed forever. A MAC address withdraw OAM message MUST contain a "Sequence Number TLV" otherwise the entire message is dropped. It MAY contain MAC Flush Parameter TLVs defined in [MAC-OPT] when static PWs are deployed in H-VPLS and PBB-VPLS scenarios. The first 2 bits of the sequence number TLV are reserved and MUST be set to 0 on transmit and ignored on receipt.

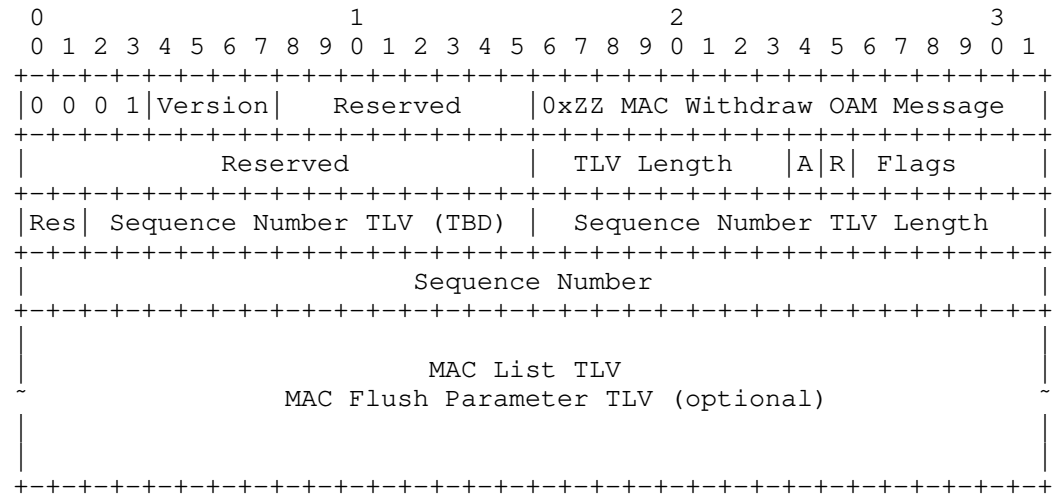


Figure 1: MAC Address Withdraw PW OAM Packet Format

In this section, MAC List TLV and MAC Flush Parameter TLV are collectively referred to as "MAC TLV(s)". The processing rules of MAC List TLV are governed by [RFC4762], and the corresponding rules of MAC Flush Parameter TLV are governed by [MAC-OPT].

"TLV Length" is the total length of all TLVs in the message, and "Sequence Number TLV Length" is the length of the sequence number field.

A single bit (called A-bit) is set to indicate if a MAC withdraw message is for ACK. Also, ACK does not include MAC TLV(s).

Only half of the sequence number space is used. Modular arithmetic is used to detect wrapping of sequence number. When sequence number wraps, all MAC addresses are flushed and the sequence number is reset.

A single bit (called R-bit) is set to indicate if the sender is requesting reset of the sequence numbers. The sender sets this bit when the Pseudowire is restarted and has no local record of send and expected receive sequence number.

4. Operation

This section describes how the initial MAC withdraw OAM messages are sent and retransmitted, as well as how the messages are processed and retransmitted messages are identified.

4.1. Operation of Sender

Each PW is associated with a counter to keep track of the sequence number of the transmitted MAC withdrawal messages. Whenever a node sends a new set of MAC TLVs, it increments the transmitted sequence number counter, and include the new sequence number in the message. The transmit sequence number is initialized to 1 at the onset.

The sender expects an ACK from the receiver within a time interval which we call "Retransmit Time" which can be either a default (1 second) or configured value. If the ACK does not arrive within the Retransmit Time, the sender retransmits the message with the same sequence number as the original message. The retransmission is ceased anytime when ACK is received or after three retries. This avoids unended retransmissions in the absence of acknowledgements. In addition, if during the period of retransmission, if a need to send a new MAC withdraw message with updated sequence number arises then retransmission of the older unacknowledged withdraw message is suspended and retransmit time for the new sequence number is initiated. In essence, sender engages in retrasmission logic only for the latest send withdraw message for a given PW.

In the event that a Pseudowire was deleted and re-added or the router is restarted with configuration, the local node may lose information about the send sequence number of previous incarnation. This becomes problematic for the remote peer as it will continue to ignore the received MAC withdraw messages with lower sequence numbers. In such cases, it is desirable to reset the sequence numbers at both ends of the Pseudowire. The 'R' reset bit is set in the first MAC withdraw to notify the remote peer to reset the send and receive sequence numbers. The 'R' bit must be cleared in subsequent MAC withdraw messages after the acknowledgement is received

4.2. Operation of Receiver

Each PW is associated with a register to keep track of the sequence number of the MAC withdrawal message received last. Whenever a MAC withdrawal message is received, and if the sequence number on the message is greater than the value in the register, the MAC address(es) contained in the MAC TLV(s) is/are removed, and the register is updated with the received sequence number. The receiver sends an ACK whose sequence number is the same as that in the received message.

If the sequence number in the received message is smaller than or equal to the value in the register, the MAC TLV(s) is/are not processed. However, an ACK with the received sequence number MUST be sent as a response. The receiver processes the ACK message as an

acknowledgement for all the MAC withdraw messages sent up to the sequence number present in the ACK message and terminates retransmission.

As mentioned above, since only half of the sequence number space is used, the receiver MUST use modular arithmetic to detect wrapping of the sequence number.

A MAC withdraw message with 'R' bit set MUST be processed by resetting the send and receive sequence number first. The rest of MAC withdraw message processing is performed as described above. The acknowledgement is sent with 'R' bit cleared.

5. IANA Considerations

The proposed mechanism requests IANA to a assign new channel type (recommended value 0x0028) from the registry named "Pseudowire Associated Channel Types". The description of the new channel type is "Pseudowire MAC Withdraw OAM Channel".

IANA needs to create a new registry for Pseudowire Associated Channel TLVs, and create an entry for "Sequence Number TLV". The recommended value is 0x0001.

6. References

6.1. Normative References

[MAC-OPT] Dutta, P., Balus, F., Stokes, O., and G. Calvinac, "LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS", draft-ietf-l2vpn-vpls-ldp-mac-opt-10.txt (work in progress), January 2014.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6478] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", RFC 6478, May 2012.

6.2. Informative References

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

Authors' Addresses

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

Email: msiva@cisco.com

Sami Boutros
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
US

Email: sboutros@cisco.com

Himanshu Shah
Ciena Corp.
3939 North First Street
San Jose, CA 95134
US

Email: hshah@ciena.com

Sam Aldrin
Huawei Technologies.
2330 Central Express Way
Santa Clara, CA 95051
US

Email: aldrin.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2013

M. Chen
W. Cao
Huawei Technologies Co., Ltd
A. Takacs
Ericsson
P. Pan
Infinera

October 21, 2012

LDP extensions for Pseudowire Binding to LSP Tunnels
draft-cao-pwe3-mpls-tp-pw-over-bidir-lsp-07.txt

Abstract

Many transport services require that user traffic, in the forms of Pseudowires (PW), to be delivered on a single co-routed bidirectional LSP or two LSPs that share the same routes. In addition, the user traffic may traverse through multiple transport networks.

This document specifies an optional extension in LDP that enable the binding between PWs and the underlying LSPs.

Requirements Language

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 [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 24, 2013.

Copyright Notice

Table of Contents

1. Introduction

Pseudo Wire (PW) Emulation Edge-to-Edge (PWE3) [RFC3985] is a mechanism to emulate layer 2 services, such as Ethernet p2p circuits. Such services are emulated between two Attachment Circuits (ACs) and the PW encapsulated layer 2 service payload is carried through Packet Switching Network (PSN) tunnels between Provider Edges (PEs). PWE3 typically uses Label Distribution Protocol (LDP) [RFC5036] or Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) [RFC3209] LSPs as PSN tunnels. The PEs select and bind the Pseudowires to PSN tunnels independently. Today, there is no protocol-based provisioning mechanism to associate PW's to PSN tunnels.

PW-to-PSN Tunnel binding has become increasingly common and important in many deployment scenarios. For instance, when connecting two remotely located sites, such as data centers, over the backbone, each site may deploy a high-performance router or switch to aggregate thousands of Ethernet VLAN flows. The aggregating routers and switches are interconnected via one or multiple MPLS/GMPLS LSP's, which may traverse through different routes or networks. Further, each Ethernet flow is offered to the customers as a bidirectional circuits with certain SLA attributes, such as bandwidth and latency. Hence, it's important for the operators to map the forwarding and reverse-direction traffic from an Ethernet circuit to the LSP's that are either bidirectional (e.g. GMPLS-initiated optical path) or co-routed.

The requirement for explicit control of PW-to-LSP mapping has been described in Section 5.3.2 ("Support for Explicit Control of PW-to-LSP Binding") of [RFC6373]. The following figure (Figure 1) provides the illustration.

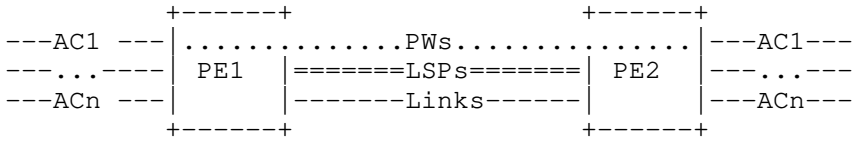


Figure 1: Explicit PW-to-LSP binding scenario

There are two PEs (PE1 and PE2) connected through multiple parallel links that may be on different fibers. Each link is managed and controlled as a bi-directional LSP. At each PE, there are a large number of bi-directional user flows from multiple Ethernet interfaces. Each user flow uses PW's to carry traffic on forwarding and reverse direction. The operators need to make sure that the user

flows (that is, the PW-pairs) to be carried on the same fiber (or, bidirectional LSP).

As mentioned above, there are a number of reasons behind this requirement. First, due to delay and latency constraints, traffic going over different fibers may require large amount of expensive buffer memory to compensate for the differential delay at the headend nodes. Further, the operators may apply different protection mechanisms on different parts of the network. As such, for optimal traffic management, traffic belongs to a particular user should traverse over the same fiber. That implies that both forwarding and reserve direction PW's that belong to the same user flow need to be mapped to the same co-routed bi-directional LSP or two LSPs with the same route.

Figure 2 illustrates a scenario where PW-LSP binding is not applied.

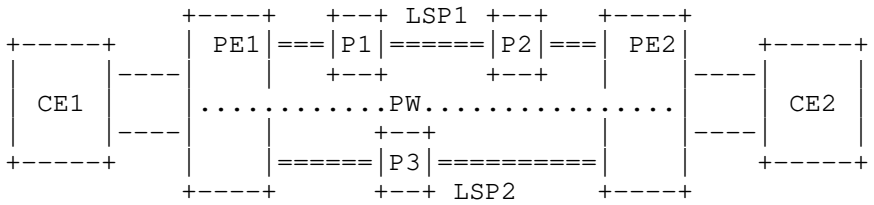


Figure 2: Inconsistent SS-PW to LSP binding scenario

LSP1 and LSP2 are two bidirectional connections on diverse paths. The operator is to deliver a bi-directional flow between PE1 and PE2. Using the existing mechanisms, it's possible that PE1 may select LSP1 (PE1-P1-P2-PE2) as the PSN tunnel for traffic from PE1 to PE2, while selecting LSP2 (PE1-P3-PE2) as the PSN tunnel for traffic from PE2 to PE1.

Consequently, the user traffic is delivered over two disjoint LSPs that may have very different service attributes in terms of latency and protection. This may not be acceptable as a reliable and effective transport service to the customers.

The similar problems may also exist in multi-segment PWs (MS-PWs), where user traffic on a particular PW may hop over different networks on forward and reverse directions.

One way to solve this problem is by introducing manual configuration at each PE to bind the PWs to the underlying PSN tunnels. However, this is prone to configuration errors and won't scale.

In this documentation, we will introduce an automatic solution by extending FEC 128/129 PW based on [RFC4447].

2. LDP Extensions

This document defines a new TLV, PSN Tunnel Binding TLV, to communicate tunnel/LSPs selection and binding requests between PEs. The TLV carries PW's binding profile and provides explicit or implicit information for the underlying PSN tunnel binding operation.

The binding TLV is optional, and MUST NOT affect the existing PW operation when not present in the messages.

The binding operation applies in both single-segment (SS) and multi-segment (MS) scenarios.

The extension supports two types of binding requests:

1. Strict binding: the requesting PE will choose and explicitly indicate the LSP information in the requests.

2. Congruent binding: a requesting PE will suggest an underlying LSP to a remote PE. On receive, the remote PE has the option to use the suggested LSP, or reply the information for an alternative.

In this document, the terminology of "tunnel" is identical to the "TE Tunnel" defined in Section 2.1 of [RFC3209], which is uniquely identified by a SESSION object that includes Tunnel end point address, Tunnel ID and Extended Tunnel ID. The terminology "LSP" is identical to the "LSP tunnel" defined in Section 2.1 of [RFC3209], which is uniquely identified by the SESSION object together with SENDER_TEMPLATE (or FILTER_SPEC) object that consists of LSP ID and Tunnel endpoint address.

2.1. PSN Tunnel Binding TLV

PSN Tunnel Binding TLV is an optional TLV and MUST be carried in the LDP Label Mapping message if PW to LSP binding is required. The format is as follows:

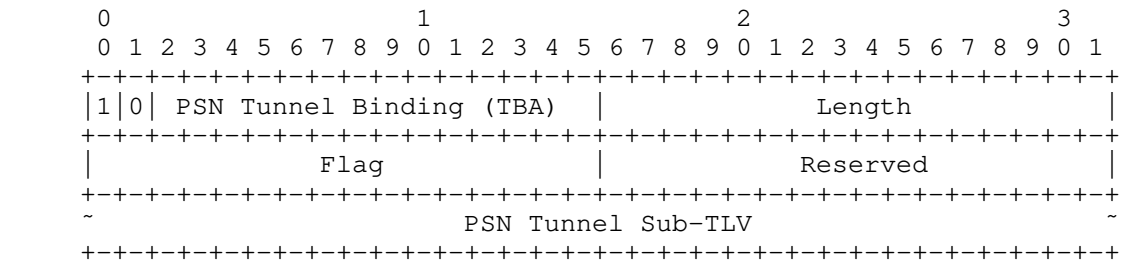
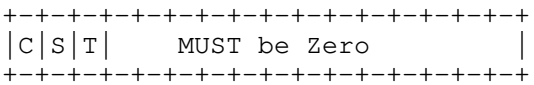


Figure 3: PSN Tunnel Binding TLV

The PSN Tunnel Binding TLV type is to be allocated by IANA

The Length field is 2 octets in length. It defines the length in octets of the entire TLV

The Flag field describes the binding requests, and has following format:



The flags are defined as the following:

C (Congruent path) bit: This informs the remote T-PE/S-PEs about the properties of the underlying LSPs. When set, the remote T-PE/S-PEs need to select LSPs with routes with the similar characiteristics (that is, bidirectional or co-routed path). If there is no such tunnel available, the node may trigger the remote T-PE/S-PEs to establish a new LSP.

S (Strict) bit: This instructs the PEs with respect to the handling of the underlying LSPs. When set, the remote PE MUST use the tunnel/ LSPs specified in the PSN Tunnel Sub-TLV as the PSN tunnel on the reverse direction of the PW, or the PW will fail to be established.

T (Tunnel Representation) bit: This indicates the format of the LSP tunnels. When the bit is set, the tunnel uses the tunnel information to identify itself, and the LSP Number fields in the PSN Tunnel sub-TLV (Section 2.1.1) MUST be set to zero. Otherwise, both tunnel and LSP information of the PSN tunnel are required. The default is set. The motivation for the T-bit is to support the MPLS protection operation where the LSP Number fields may be ignored.

C-bit and S-bit are mutually exclusive from each other, and cannot be set in the same message.

2.1.1. PSN Tunnel Sub-TLV

PSN Tunnel Sub-TLVs are designed for inclusion in the PSN Tunnel Binding TLV to specify the tunnel/LSPs to which a PW is required to bind.

Two sub-TLVs are defined: the IPv4 and IPv6 Tunnel sub-TLVs.

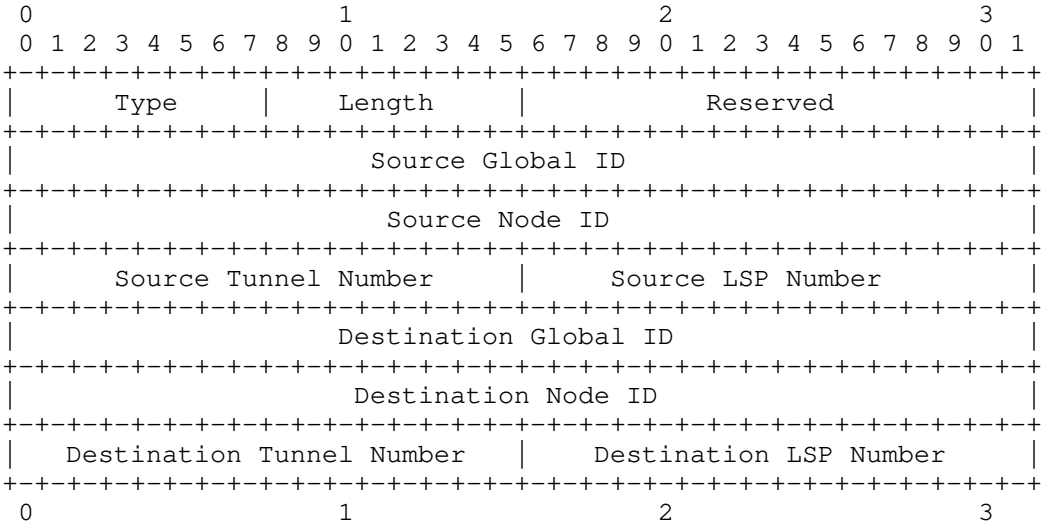


Figure 4: IPv4 PSN Tunnel sub-TLV format

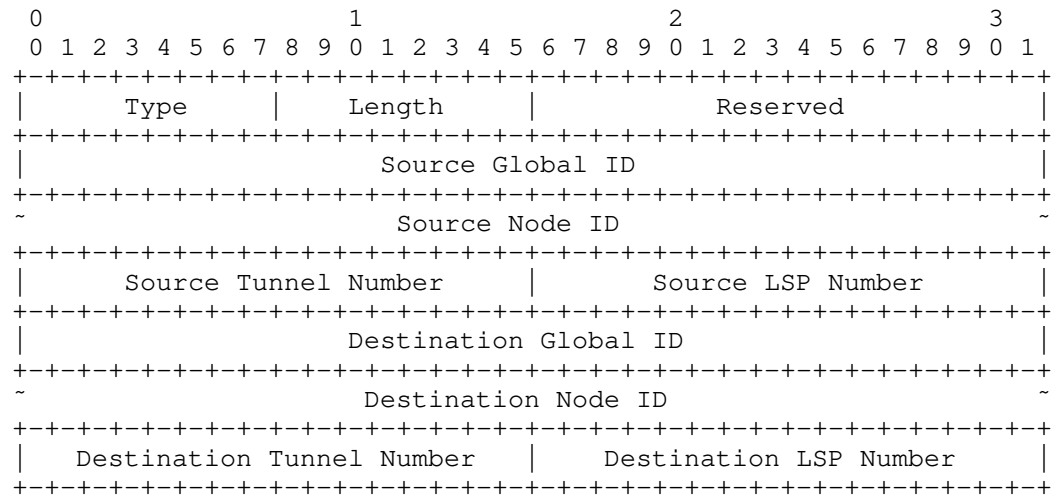


Figure 5: IPv6 PSN Tunnel sub-TLV format

The definition of Source and Destination Global/Node IDs and Tunnel/ LSP Numbers are derived from [RFC6370]. This is to describe the underlying LSP's. Note that the LSP's in this notation is globally unique.

As defined in Section 4.6.1.2 and Section 4.6.2.2 of [RFC3209], the "Tunnel endpoint address" is mapped to Destination Node ID, and "Extended Tunnel ID" is mapped to Source Node ID. Both IDs can be IPv6 addresses.

A PSN Tunnel sub-TLV could be used to either identify a tunnel or a specific LSP. The T-bit in the Flag field defines the distinction as such that, when the T-bit is set, the Source/Destination LSP Number fields MUST be zero and ignored during processing. Otherwise, both Source/Destination LSP Number fields MUST have the actual LSP IDs of specific LSPs.

Each PSN Tunnel Binding TLV can only have one such sub-TLV.

3. Theory of Operation

During PW setup, the PEs may select desired forwarding tunnels/LSPs, and inform the remote T-PE/S-PEs about the desired reverse tunnels/ LSPs.

Specifically, to set up a PW (or PW Segment), a PE may select a

candidate tunnel/LSP to act as the PSN tunnel. If none is available or satisfies the constraints, the PE will trigger and establish a new tunnel/LSP. The selected tunnel/LSP information is carried in the PSN Tunnel Binding TLV and sent with the Label Mapping message to the target PE.

Upon the reception of the Label Mapping message, the receiving PE will process the PSN Tunnel Binding TLV, determine whether it can accept the suggested tunnel/LSP or to find the reverse tunnel/LSP that meets the request, and respond with a Label Mapping message, which contains the corresponding PSN Tunnel Binding TLV.

It is possible that two PEs may request PSN binding to the same PW or PW segment over different tunnels/LSPs at the same time. There may cause collisions of tunnel/LSPs selection as both PEs assume the active role.

As defined in (Section 7.2.1, [RFC6073]), each PE may be generally categorized into active and passive roles:

1. Active PE: the PE which initiates the selection of the tunnel/ LSPs and informs the remote PE;

2. Passive PE: the PE which obeys the active PE's suggestion.

In the remaining of this document, we will elaborate the operation for SS-PW and MS-PW:

1. SS-PW: In this scenario, both PE's for a particular PE may assume the active roles

2. MS-PW: One PE is active, while the other is passive. The PW's are setup using FEC 129

4. PSN Binding Operation for SS-PW

As illustrated in Figure-5, both PEs (say, PE1 and PE2) of a PW may independently initiate the setup. To perform PSN binding, the Label Mapping messages MUST carry a PSN Tunnel Binding TLV, and the PSN Tunnel sub-TLV MUST contains the desired tunnel/LSPs of the sender.

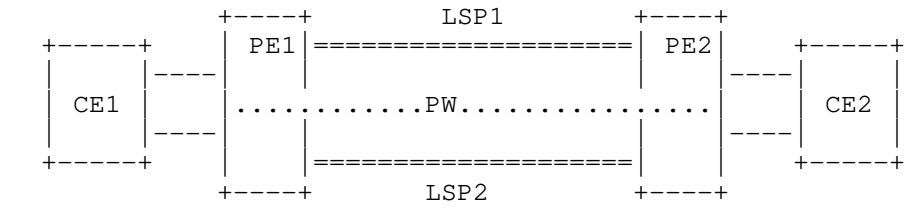


Figure 6: PSN binding operation in SS-PW environment

As outlined previously, there are two types of binding request: congruent and strict.

In strict binding, a PE (e.g., PE1) will mandate the other PE (e.g., PE2) to use a specified tunnel/LSP (e.g. LSP1) as the PSN tunnel on the reverse direction. In the PSN Tunnel Binding TLV, the S-bit MUST be set, the C-bit MUST be reset, and the Source and Destination IDs/ Numbers MUST be filled.

On receive, if the S-bit is set, other than following the processing procedure defined in Section 5.3.3 of [RFC4447], the receiving PE (i.e. PE2) needs to determine whether to accept the indicated tunnel/LSP in PSN Tunnel Sub-TLV.

If the receiving PE (PE2) is also an active PE, and may have initiated the PSN binding requests to the other PE (PE1), if the received PSN tunnel/LSP is the same as it has been sent in the Label Mapping message by PE2, then the signaling has converged on a mutually agreed Tunnel/LSP. The binding operation is completed.

Otherwise, the receiving PE (PE2) MUST compare its own Node ID against the received Source Node ID. If it is numerically lower, the PE (PE2) will reply a Label Mapping message to complete the PW setup and confirm the binding request. The PSN Tunnel Binding TLV in the message MUST contain the same Source and Destination IDs/Numbers as in the received binding request, in the appropriate order. On the other hand, if the receiving PE (PE2) has a Node ID that is numerically higher than the Source Node ID carried in the PSN Tunnel Binding TLV, it MUST reply a Label Release message with status code set to "Reject to use the suggested tunnel/LSPs" and the received PSN Tunnel Binding TLV, and the PW will not be established.

To support congruent binding, the receiving PE can select the appropriated PSN tunnel/LSP for the reverse direction of the PW, so long as the forwarding and reverse PSNs share the same route.

Initially, a PE (PE1) sends a Label Mapping message to the remote PE (PE2) with the PSN Tunnel Binding TLV, with C-bit set, S-bit reset, and the appropriate Source and Destination IDs/Numbers. In case of

unidirectional LSPs, the PSN Tunnel Binding TLV may only contain the Source IDs/Numbers, the Destination IDs/Numbers are set to zero and left for PE2 to fill when responding the Label Mapping message.

On receive, since PE2 is also an active PE, and may have initiated the PSN binding requests to the other PE (PE1), if the received PSN tunnel/LSP has the same route as the one that has been sent in the Label Mapping message to PE1, then the signaling has converged. The binding operation is completed.

Otherwise, it needs to compare its own Node ID against the received Source Node ID. If it's numerically lower, PE2 needs to find/ establish a tunnel/LSP that meets the congruent constraint, and reply a Label Mapping message with a PSN Binding TLV that contains the Source and Destination IDs/Numbers in the appropriate order. On the other hand, if the receiving PE (PE2) has a Node ID that is numerically higher than the Source Node ID carried in the PSN Tunnel Binding TLV, it MUST reply a Label Release message with status code set to "Reject to use the suggested tunnel/LSPs" and the received PSN Tunnel Binding TLV.

In both strict and congruent bindings, if T-bit is set, the LSP Number field MUST be set to zero. Otherwise, the field MUST contain the actual LSP number for the associated PSN LSP.

After a PW established, the operators may choose to move the PW's from the current tunnel/LSPs. Or, the underlying PSN is broken due to network failure. In this scenario, a new Label Mapping message MUST be sent to update the changes. Note that when T-bit is set, the working LSP broken will not trigger to update the changes if there are protection LSP's.

The message may carry a new PSN Tunnel Binding TLV, which contains the new Source and Destination Numbers/IDs. The handling of the new message should be identical to what has been described in this section.

However, if the new Label Binding message does not contain the PSN Tunnel Binding TLV, it declares the removal of any congruent/strict constraints. The PEs may not map the PW to the underlying PSN on purpose, the current independent PW to PSN binding will be used.

Further, as an implementation option, the PEs should not remove the traffic from an operational PW, until the completion of the underlying PSN tunnel/LSP changes.

5. PSN Binding Operation for MS-PW

MS-PW uses FEC 129 for PW setup. We refer the operation to Figure-6.

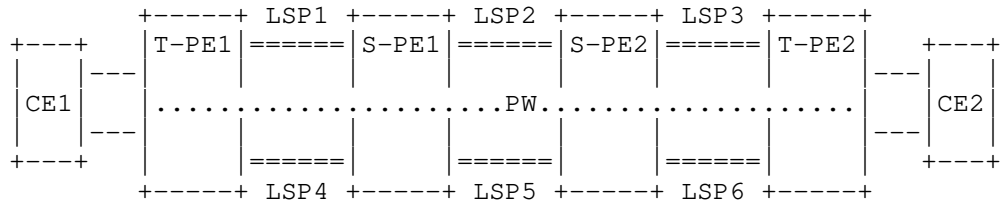


Figure 7: PSN binding operation in MS-PW environment

When an active PE (that is, T-PE1) starts to signal for a MS-PW, a PSN Tunnel Binding TLV MUST be carried in the Label Mapping message and sent to the adjacent S-PE (that is, S-PE1). The PSN Tunnel Binding TLV includes the PSN Tunnel sub-TLV that carries the desired tunnel/LSP of T-PE1's.

For strict binding, the initiating PE MUST set the S-bit, reset the C-bit and indicates the binding tunnel/LSP to the next-hop S-PE.

When S-PE1 receives the Label Mapping message, S-PE1 needs to determine if the signaling is for forward or reverse direction, as defined in Section 6.2.3 of [I-D.ietf-pwe3-dynamic-ms-pw].

If the Label Mapping message is for forward direction, and S-PE1 accepts the requested tunnel/LSPs from T-PE1, S-PE1 must save the tunnel/LSP information for reverse-direction processing later on. If the PSN binding request is not acceptable, S-PE1 MUST reply a Label Release Message to the upstream PE (T-PE1) with Status Code set to "Reject to use the suggested tunnel/LSPs".

Otherwise, S-PE1 relays the Label Mapping message to the next S-PE (that is, S-PE2), with the PSN Tunnel sub-TLV carrying the information of the new PSN tunnel/LSPs selected by S-PE1. S-PE2 and subsequent S-PEs will repeat the same operation until the Label Mapping message reaches to the remote T-PE (that is, T-PE2).

If T-PE2 agrees with the requested tunnel/LSPs, it will reply a Label Mapping message to initiate to the binding process on the reverse direction. The Label Mapping message contains the received PSN Tunnel Binding TLV for confirmation purposes.

When its upstream S-PE (S-PE2) receives the Label Mapping message, the S-PE relays the Label Mapping message to its upstream adjacent S-PE (S-PE1), with the previously saved PSN tunnel/LSP information in the PSN Tunnel sub-TLV. The same procedure will be applied on subsequent S-PEs, until the message reaches to T-PE1 to complete the PSN binding setup.

During the binding process, if any PE does not agree to the requested tunnel/LSPs, it can send a Label Release Message to its upstream adjacent PE with Status Code set to "Reject to use the suggested tunnel/LSPs".

For congruent binding, the initiating PE (T-PE1) MUST set the C-bit, reset the S-bit and indicates the suggested tunnel/LSP in PSN Tunnel sub-TLV to the next-hop S-PE (S-PE1).

During the MS-PW setup, the PEs have the option to ignore the suggested tunnel/LSP, and select another tunnel/LSP for the segment PW between itself and its upstream PE on reverse direction only if the tunnel/LSP is congruent with the forwarding one. Otherwise, the procedure is the same as the strict binding.

The tunnel/LSPs may change after a MS-PW being established. When a tunnel/LSP has changed, the PE that detects the change SHOULD select an alternative tunnel/LSP for temporary use while negotiating with other PEs following the procedure described in this section.

6. Security Considerations

The ability to control which LSP to carry traffic from a PW can be a potential security risk both for denial of service and traffic interception. It is RECOMMENDED that PEs do not accept the use of LSPs identified in the PSN Tunnel Binding TLV unless the LSP end points match the PW or PW segment end points. Furthermore, where security of the network is believed to be at risk, it is RECOMMENDED that PEs implement the LDP security mechanisms described in [RFC5036] and [RFC5920].

7. IANA Considerations

7.1. LDP TLV Types

This document defines new TLV [Section 2.1 of this document] for inclusion in LDP Label Mapping message. IANA is required to assign TLV type value to the new defined TLVs from LDP "TLV Type Name Space" registry.

7.1.1. PSN Tunnel Sub-TLVs

This document defines two sub-TLVs [Section 2.1.1 of this document] for PSN Tunnel Binding TLV. IANA is required to create a new registry ("PSN Tunnel Sub-TLV Name Space") for PSN Tunnel sub-TLVs and to assign Sub-TLV type values to the following sub-TLVs.

IPv4 PSN Tunnel sub-TLV - 0x01 (to be confirmed by IANA)

IPv6 PSN Tunnel sub-TLV - 0x02 (to be confirmed by IANA)

7.2. LDP Status Codes

This document defines a new LDP status codes, IANA is required to assigned status codes to these new defined codes from LDP "STATUS CODE NAME SPACE" registry.

"Reject to use the suggested tunnel/LSPs" - 0x0000003B (to be confirmed by IANA)

8. Acknowledgements

The authors would like to thank Adrian Farrel, Kamran Raza, Xinchun Guo, Mingming Zhu and Li Xue for their comments and help in preparing this document. Also this draft benefits from the discussions with Nabil Bitar, Paul Doolan, Frederic Journay, Andy Malis, Curtis Villamizar and Luca Martini.

9. References

9.1. Normative References

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

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.

9.2. Informative References

[I-D.ietf-pwe3-dynamic-ms-pw]
Martini, L., Bocci, M., and F. Balus, "Dynamic Placement

of Multi Segment Pseudowires",
draft-ietf-pwe3-dynamic-ms-pw-15 (work in progress),
June 2012.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[RFC6073] Martini, L., Metz, C., Nadeau, T., Bocci, M., and M. Aissaoui, "Segmented Pseudowire", RFC 6073, January 2011.

[RFC6373] Andersson, L., Berger, L., Fang, L., Bitar, N., and E. Gray, "MPLS Transport Profile (MPLS-TP) Control Plane Framework", RFC 6373, September 2011.

Authors' Addresses

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd
Q14 Huawei Campus, No. 156 Beiqing Road, Hai-dian District
Beijing 100095
China

Email: mach@huawei.com

Wei Cao
Huawei Technologies Co., Ltd
Q14 Huawei Campus, No. 156 Beiqing Road, Hai-dian District
Beijing 100095
China

Email: wayne.caowei@huawei.com

Attila Takacs
Ericsson
Laborc u. 1.
Budapest 1037
Hungary

Email: attila.takacs@ericsson.com

Ping Pan
Infinera
169 West Java Drive, Sunnyvale, CA 94089
US

Email: ppan@infinera.com

Network Working Group
Internet-Draft
Updates: 4379 (if approved)
Intended status: Standards Track
Expires: June 7, 2012

M. Chen
Huawei Technologies Co., Ltd
P. Pan
Infinera
C. Pignataro
R. Asati
Cisco

December 5, 2011

Label Switched Path (LSP) Ping for IPv6 Pseudowire FECs
draft-chen-mpls-ipv6-pw-lsp-ping-03

Abstract

Multi-Protocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute mechanisms are commonly used to detect and isolate data plane failures in all MPLS LSPs including Pseudowire (PW) LSPs. The PW LSP Ping and Traceroute elements, however, are not specified for IPv6 address usage.

This document extends the PW LSP Ping and Traceroute mechanisms so they can be used with IPv6 PWs, and updates RFC 4379.

Requirements Language

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 [RFC2119].

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Multi-Protocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute are defined in [RFC4379]. These mechanisms can be used to detect and isolate data plane failures in all MPLS Label Switched Paths (LSPs) including Pseudowires (PWs). The PW LSP Ping and Traceroute elements, however, are not specified for IPv6 address usage.

Specifically, the PW FEC sub-TLVs for the Target FEC Stack in the LSP Ping and Traceroute mechanism are defined only for IPv4 Provider Edge (PEs) routers, and are not applicable for the case where PEs use IPv6 addresses. Three PW related Target Forwarding Equivalence Class (FEC) sub-TLVs are currently defined (FEC 128 Pseudowire-Deprecated, FEC 128 Pseudowire-Current, and FEC 129 Pseudowire, see Sections 3.2.8 through 3.2.10 of [RFC4379]). These sub-TLVs contain the source and destination addresses of the target LDP session, and currently only IPv4 target LDP session is covered. Despite the fact that the PE IP address family is not explicit in the sub-TLV definition, this can be inferred indirectly by examining the lengths of the Sender's/Remote PE Address fields, or calculating the Length of the sub-TLVs (see Section 3.2 of [RFC4379]). When an IPv6 target LDP session is used, these existing sub-TLVs can not therefore be used since the addresses will not fit. Additionally, all other sub-TLVs are defined in pairs, one for IPv4 and another for IPv6, but not the PW sub-TLVs.

This document updates [RFC4379] to explicitly constraint the existing PW FEC sub-TLVs for IPv4 LDP sessions, and extends the PW LSP Ping to IPv6 LDP sessions (i.e., when IPv6 LDP sessions are used to signal the PW, the Sender's and Receiver's IP addresses are IPv6 addresses). This is done by renaming the existing PW sub-TLVs to say "IPv4", and also by defining two new Target FEC sub-TLVs (IPv6 FEC 128 Pseudowire sub-TLV and IPv6 FEC 129 Pseudowire sub-TLV) to extend the application of PW LSP Ping and Traceroute to the IPv6 usage when an IPv6 LDP session [I-D.ietf-mpls-ldp-ipv6] is used to signal the Pseudowire. Note that FEC 128 Pseudowire (Deprecated) is not defined for IPv6 in this document.

2. IPv4 Pseudowire Sub-TLVs

This document updates Section 3.2 and Sections 3.2.8 through 3.2.10 of [RFC4379] as follows and as indicated in Section 4 and Section 6. This is done to avoid any potential ambiguity, confusion, and backwards compatibility issues.

Sections 3.2.8 through 3.2.10 of [RFC4379] list the PW sub-TLVs and

state:

"FEC 128" Pseudowire (Deprecated)

"FEC 128" Pseudowire

"FEC 129" Pseudowire

These names and titles are now changed to:

IPv4 "FEC 128" Pseudowire (Deprecated)

IPv4 "FEC 128" Pseudowire

IPv4 "FEC 129" Pseudowire

Additionally, when referring to the PE addresses, these three sections state:

Sender's PE Address

Remote PE Address

These are now updated to say:

Sender's PE IPv4 Address

Remote PE IPv4 Address

3. IPv6 Pseudowire Sub-TLVs

3.1. IPv6 FEC 128 Pseudowire Sub-TLV

IPv6 FEC 128 Pseudowire sub-TLV has the consistent structure with FEC 128 Pseudowire sub-TLV as described in Section 3.2.9 of [RFC4379]. The encoding of IPv6 FEC 128 Pseudowire sub-TLV is as follows:

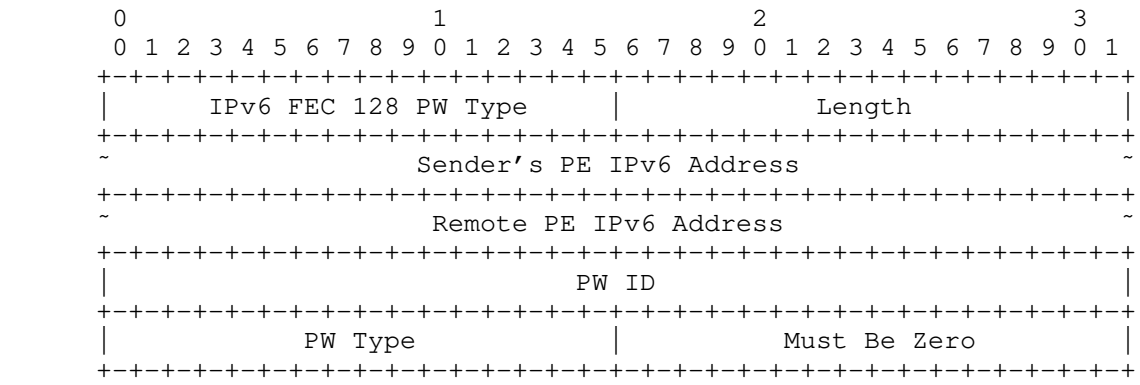


Figure 1: IPv6 FEC 128 Pseudowire

IPv6 FEC 128 PW: TBD.

Length: it defines the length in octets of the value field of the sub-TLV and its value is 38.

Sender's PE IPv6 Address: The source IP address of the target IPv6 LDP session.

Remote PE IPv6 Address: The destination IP address of the target IPv6 LDP session.

PW ID: Same as FEC 128 Pseudowire [RFC4379].

PW Type: Same as FEC 128 Pseudowire [RFC4379].

3.2. IPv6 FEC 129 Pseudowire Sub-TLV

IPv6 FEC 129 Pseudowire sub-TLV has the consistent structure with FEC 129 Pseudowire sub-TLV as described in Section 3.2.10 of [RFC4379]. The encoding of IPv6 FEC 129 Pseudowire is as follows:

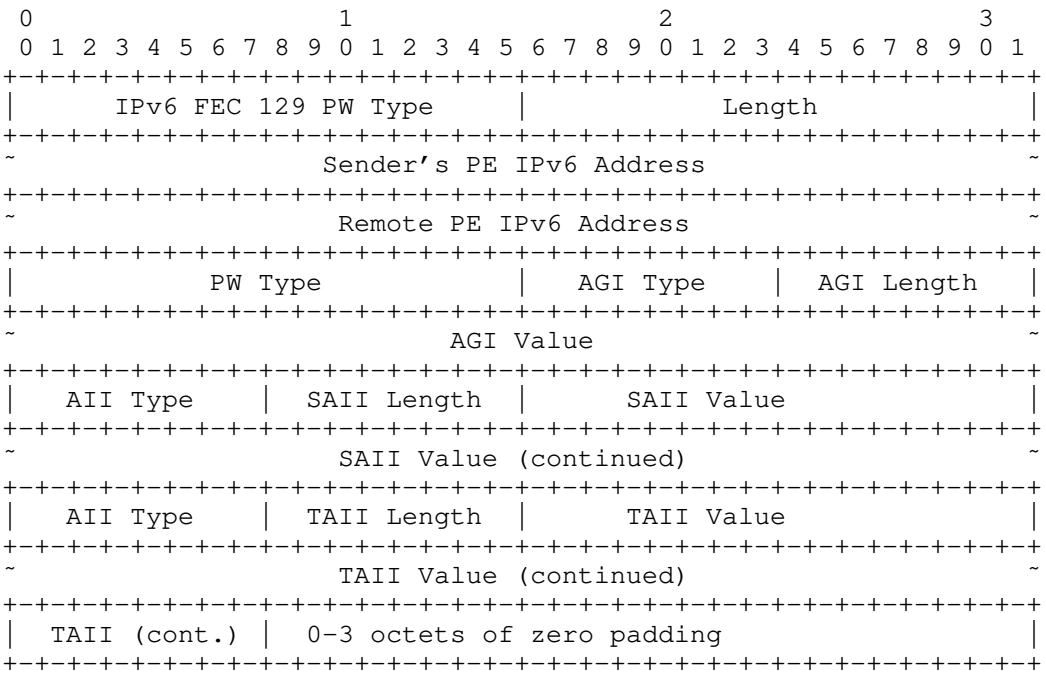


Figure 2: IPv6 FEC 129 Pseudowire

IPv6 FEC 129 PW: TBD.

The Length of this TLV is 40 + AGI length + SAII length + TAII length. Padding is used to make the total length a multiple of 4; the length of the padding is not included in the Length field.

Sender's PE IPv6 Address: The source IP address of the target IPv6 LDP session.

Remote PE IPv6 Address: The destination IP address of the target IPv6 LDP session.

The other fields are same as FEC 129 Pseudowire [RFC4379].

4. Summary of Changes

Section 3.2 of [RFC4379] tabulates all the sub-TLVs for the Target FEC Stack. Per the change described in Section 2 and Section 3, the table would show the following:

Sub-Type	Length	Value Field
...		
9	10	IPv4 "FEC 128" Pseudowire (deprecated)
10	14	IPv4 "FEC 128" Pseudowire
11	16+	IPv4 "FEC 129" Pseudowire
...		
TBD	38	IPv6 "FEC 128" Pseudowire
TBD	40+	IPv6 "FEC 129" Pseudowire

5. Operation

This document does not define any new procedures. The process described in [RFC4379] MUST be used.

6. IANA Considerations

IANA is requested to perform the following assignments in the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, "TLVs and sub-TLVs" sub-registry.

[RFC Editor: To be REMOVED prior to publication. This registration should take place at <http://www.iana.org/assignments/mpls-lsp-ping-parameters/mpls-lsp-ping-parameters.xml#mpls-lsp-ping-parameters-7>]

Update the Value fields of these three Sub-TLVs, adding the "IPv4" qualifier (see Section 2), and update the Reference to point to this document:

Type	Sub-Type	Value Field
1	9	IPv4 "FEC 128" Pseudowire (Deprecated)
1	10	IPv4 "FEC 128" Pseudowire
1	11	IPv4 "FEC 129" Pseudowire

Create two new entries for the Sub-Type field of Target FEC TLV (see Section 3):

Type	Sub-Type	Value Field
1	TBD1	IPv6 "FEC 128" Pseudowire
1	TBD2	IPv6 "FEC 129" Pseudowire

7. Security Considerations

This draft does not introduce any new security issues, the security mechanisms defined in [RFC4379] apply here.

8. Acknowledgements

The authors gratefully acknowledge review and comments of Vanson Lim, Tom Petch, and Spike Curtis.

9. References

9.1. Normative References

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

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

9.2. Informative References

[I-D.ietf-mpls-ldp-ipv6] Asati, R., Manral, V., Papneja, R., and C. Pignataro, "Updates to LDP for IPv6", draft-ietf-mpls-ldp-ipv6-05 (work in progress), August 2011.

Authors' Addresses

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd
No. 3 Xinxi Road, Shang-di, Hai-dian District
Beijing 100085
China

Email: mach@huawei.com

Ping Pan
Infinera
US

Email: ppan@infinera.com

Carlos Pignataro
Cisco Systems
7200-12 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: cpignata@cisco.com

Rajiv Asati
Cisco Systems
7025-6 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: rajiva@cisco.com

Network Working Group J. Dong
Internet-Draft H. Wang
Intended status: Standards Track Huawei Technologies
Expires: May 26, 2013 November 22, 2012

Pseudowire Redundancy on S-PE
draft-dong-pwe3-redundancy-spe-04

Abstract

This document describes Multi-Segment Pseudowire (MS-PW) protection scenarios in which the pseudowire redundancy is provided on the Switching-PE (S-PE). Operations of the S-PEs which provide PW redundancy are specified. Signaling of the preferential forwarding status as defined in [I-D.ietf-pwe3-redundancy-bit] is reused. This document does not require any change to the T-PEs of MS-PW.

Requirements Language

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 [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 26, 2013.

Copyright Notice

Table of Contents

1. Introduction

[RFC6718] describes the framework and requirements for pseudowire (PW) redundancy, and [I-D.ietf-pwe3-redundancy-bit] specifies Pseudowire (PW) redundancy mechanism for scenarios where a set of redundant PWs is configured between provider edge (PE) nodes in single-segment pseudowire (SS-PW) [RFC3985]applications, or between terminating provider edge (T-PE) nodes in multi-segment pseudowire (MS-PW) [RFC5659] applications.

In some MS-PW scenarios, there are some benefits to provide PW redundancy on S-PEs, such as reducing the burden on the access T-PE nodes, and faster protection switching. This document describes some scenarios in which PW redundancy is provided on S-PEs, and specifies the operations of the S-PEs. Signaling of the preferential forwarding status as defined in [I-D.ietf-pwe3-redundancy-bit] is reused. This document does not require any change to the T-PEs of MS-PW.

2. PW Redundancy on S-PE

In some MS-PW deployment scenarios, there are some benefits to provide PW redundancy on S-PEs. This section gives some examples of PW redundancy on S-PE.

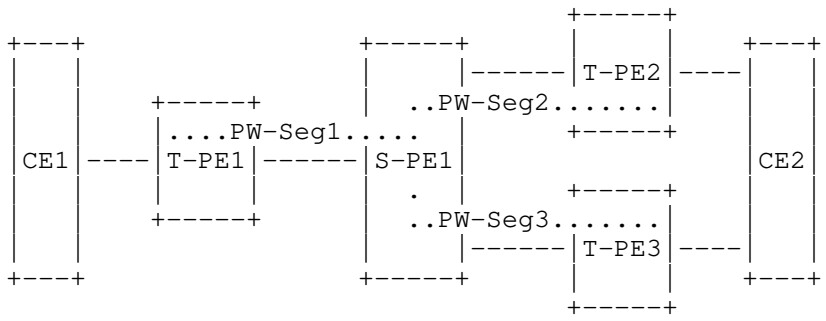


Figure 1.MS-PW Redundancy on S-PE

As illustrated in Figure 1, CE1 is connected to T-PE1 while CE2 is dual-homed to T-PE2 and T-PE3. T-PE1 is connected to S-PE1 only, and S-PE1 is connected to T-PE2 and T-PE3. The MS-PW is switched on S-PE1, and PW-Seg2 and PW-Seg3 provides resiliency on S-PE1 for failure of T-PE2 or T-PE3 or the connected ACs. PW-Seg2 is selected as primary PW segment, and PW-Seg3 is secondary PW segment.

MS-PW redundancy on S-PE is beneficial for the scenario in Figure 1 since T-PE1 as an access node may not be able to provide PW

redundancy, especially when the PW-Seg1 between T-PE1 and S-PE1 is statically configured. And with PW redundancy on S-PE, the number of PW segments needed between T-PE1 and S-PE1 is only half of the number of PW segments needed for end-to-end MS-PW redundancy. In addition, PW redundancy on S-PE could provide faster protection switching than end-to-end protection switching of MS-PW.

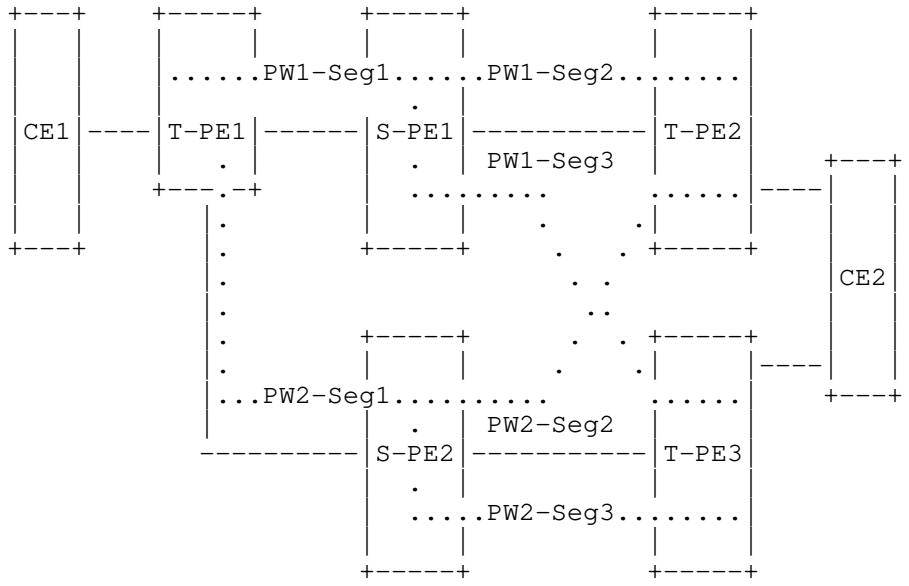


Figure 2. MS-PW Redundancy on S-PE with S-PE protection

As illustrated in Figure 2, CE1 is connected to T-PE1 while CE2 is dual-homed to T-PE2 and T-PE3. T-PE1 is connected to S-PE1 and S-PE2, and both S-PE1 and S-PE2 are connected to T-PE2 and T-PE3. There are two MS-PWs which are switched at S-PE1 and S-PE2 respectively to provide S-PE node protection. For MS-PW1, the S-PE1 provides resiliency using PW1-Seg2 and PW1-Seg3. For MS-PW2, the S-PE2 provides resiliency using PW2-Seg2 and PW2-Seg3. MS-PW1 is the primary PW and PW1-Seg2 is the primary PW segment.

MS-PW redundancy on S-PE is beneficial for the scenario in Figure 2 since it reduces the number of end-to-end MS-PWs required for both T-PE and S-PE protection. In addition, PW redundancy on S-PE could provide faster protection switching than end-to-end protection switching of MS-PW.

3. S-PE Operations

For an S-PE which provides PW redundancy, it is important to

advertise proper preferential forwarding status to the PW segments on both sides and perform protection switching according to the received status. This section specifies the operations of S-PEs on which PW redundancy is provisioned. This document does not make any change to the T-PEs of MS-PW.

The S-PE SHOULD work as a Slave node for the single-connected side, and SHOULD work in Independent mode for the multi-connected side. The S-PE SHOULD pass the preferential forwarding status received from the single-connected side unchanged to the PW segments on the multi-connected side. The S-PE SHOULD advertise Standby status to the single-connected side if it receives Standby status from all the PW segments on the multi-connected side, and it SHOULD advertise Active status to the single-connected side if it receives Active status from any of the PW segments on the multi-connected side. For the single-connected side, the active PW segment is determined by the T-PE on this side, which works as the Master node. On the multi-connected side, the PW segment which has both local and remote Preferential Forwarding status as Active SHOULD be selected for traffic forwarding.

The Signaling of Preferential Forwarding bit defined in [I-D.ietf-pwe3-redundancy-bit] is reused in these scenarios.

For the scenario in Figure 1, assume the AC from CE2 to T-PE2 is active. In normal operation, S-PE1 would receive Active Preferential Forwarding status bit on the single-connected side from T-PE1, then it would advertise Active Preferential Forwarding status bit on both PW-Seg2 and PW-Seg3. T-PE2 and T-PE3 would advertise Active and Standby preferential status bit respectively to S-PE1, reflecting the forwarding state of the two ACs to CE2. By matching the local and remote Up/Down status and Preferential Forwarding status, PW-Seg2 would be used for traffic forwarding.

On failure of the AC between CE2 and T-PE2, the forwarding state of AC on T-PE3 is changed to Active. T-PE3 then advertises Active Preferential Status to S-PE1, and T-PE2 would advertise the Preferential Status bit of Standby to S-PE1. S-PE1 would perform the switchover according to the updated local and remote Preferential Forwarding status, and select PW-Seg3 for traffic forwarding. Since S-PE1 still connects to an Active PW segment on the multi-connected side, it will not advertise any change of the PW Preferential Forwarding status to T-PE1. T-PE1 would not be aware of the switchover on S-PE1.

For scenario of Figure 2, assume the AC from CE2 to T-PE2 is active. T-PE1 works in Master mode and it would advertise Active and Standby Preferential Forwarding status bit respectively to S-PE1 and S-PE2.

According to the received Preferential Forwarding status bit, S-PE1 would advertise Active Preferential Forwarding status bit to both T-PE2 and T-PE3, and S-PE2 would advertise Standby Preferential Forwarding status bit to both T-PE2 and T-PE3. T-PE2 would advertise Active Preferential Forwarding status bit to both S-PE1 and S-PE2, and T-PE3 would advertise Standby Preferential Forwarding status bit to both S-PE1 and S-PE2, reflecting the forwarding state of the two ACs to CE2. By matching the local and remote Up/Down Status and Preferential Forwarding status, PW1-Seg2 from S-PE1 to T-PE2 would be used for traffic forwarding. Since S-PE1 connects to the Active PW segment on the multi-connected side, it would advertise Active Preferential Forwarding status bit to T-PE1, and S-PE2 would advertise Standby Preferential Forwarding status bit to T-PE1 since it does not have any Active PW segment on the multi-connected side.

On failure of the AC between CE2 and T-PE2, the forwarding state of AC on T-PE3 is changed to Active. T-PE3 would then advertise Active Preferential Forwarding status bit to both S-PE1 and S-PE2, and T-PE2 would advertise Standby Preferential Forwarding status bit to both S-PE1 and S-PE2. S-PE1 would perform the switchover according to the updated local and remote Preferential Forwarding status, and select PW1-Seg3 for traffic forwarding. Since S-PE1 still has an Active PW segment on the multi-connected side, it would not advertise any change of the PW status to T-PE1. Thus T-PE1 would not be aware of the switchover on S-PE1.

If S-PE1 fails, T-PE1 would notice this through some detection mechanism and then advertise the Active Preferential Forwarding status bit to S-PE2, and PW2-Seg1 would be selected by T-PE1 for traffic forwarding. On receipt of the newly changed Preferential Forwarding status, S-PE2 would advertise the Active Preferential Forwarding status to both T-PE2 and T-PE3. T-PE2 and T-PE3 would also notice the failure of S-PE1 by some detection mechanism. Then by matching the local and remote Up/Down and Preferential Forwarding status, PW2-Seg2 would be selected for traffic forwarding.

4. VCCV Considerations

PW VCCV [RFC5085] CC type 1 "PW ACH" can be used with S-PE redundancy mechanism. VCCV CC type 2 "Router Alert Label" is not supported for MS-PW as specified in [RFC6073]. If VCCV CC type 3 "TTL Expiry" is to be used, the hop count from one T-PE to the remote T-PE needs to be obtained in advance. This can be achieved either by control plane SP-PE TLVs or through data plane tracing of the MS-PW.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

This document has the same security properties as in the PWE3 control protocol [RFC4447] and [I-D.ietf-pwe3-redundancy-bit].

7. Acknowledgements

The authors would like to thank Mach Chen, Lizhong Jin, Mustapha Aissaoui, Luca Martini, Matthew Bocci and Stewart Bryant for their comments and discussions.

8. References

8.1. Normative References

[I-D.ietf-pwe3-redundancy-bit] Muley, P. and M. Aissaoui, "Pseudowire Preferential Forwarding Status Bit", draft-ietf-pwe3-redundancy-bit-08 (work in progress), September 2012.

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

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC5659] Bocci, M. and S. Bryant, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.

[RFC6718] Muley, P., Aissaoui, M., and M. Bocci, "Pseudowire Redundancy", RFC 6718, August 2012.

8.2. Informative References

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for

Pseudowires", RFC 5085, December 2007.

[RFC6073] Martini, L., Metz, C., Nadeau, T., Bocci, M., and M. Aissaoui, "Segmented Pseudowire", RFC 6073, January 2011.

Authors' Addresses

Jie Dong
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: rainsword.wang@huawei.com

Pseudowire Emulation Edge to Edge
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

H. Hao
Y. Ma
ZTE Corporation
W. Cheng
China Mobile
D. Cohn

M. Daikoku
KDDI Corporation

October 22, 2012

ICCP extension for the MSP application
draft-hao-pwe3-iccp-extension-for-msp-04

Abstract

This document specifies extensions to the Inter-Chassis Communication Protocol (ICCP) to support inter-chassis linear multiplex section protection (MSP) as described in G.841 and automatic protection switching as defined in ANSI T1.105.01. This document considers an application where a CE device or access network is attached to two PEs through Synchronous Digital Hierarchy (SDH) circuits, and MSP or APS is used to protect the attachment circuits. ICCP is used to support configuration and state synchronization between two chassis. CE device or access network attached to more than two PEs is out of the scope of this document.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 25, 2013.

Table of Contents

1. Introduction

[I-D:ietf-pwe3-iccp] has specified an inter-chassis communication protocol that enables Provider Edge (PE) device redundancy for Virtual Private Wire Service (VPWS) and Virtual Private LAN Service (VPLS) applications. The protocol runs within a set of two or more PEs, forming a redundancy group (RG), for the purpose of synchronizing data amongst the systems. In the ICCP draft, it specifies the ICCP TLVs for the Pseudowire Redundancy application and the multi-chassis LACP(mLACP) application. This document extends the ICCP TLVs for SDH attachment circuit redundancy using inter-chassis linear multiplex section protection (MSP) application. The application also supports SONET attachment circuits using automatic protection switching (APS). Unless otherwise stated, all requirements in this document are also applicable to SONET/APS, and all references to MSP equally apply to APS.

Inter-chassis linear multiplex section protection (MSP) application also adopts the topology described in Figure 1 of [I-D:ietf-pwe3-iccp]. In other words, the redundancy mechanism employed towards the access node/network is inter-chassis linear MSP which is commonly used in mobile backhaul networks. Packet transport technology is widely used in mobile backhaul networks, with either Ethernet or SDH as attachment circuit technology.

In packet transport mobile backhaul networks, 3G access nodes that typically connect to the network using Ethernet interfaces coexist with 2G access nodes that typically connect to the network using SDH interfaces. In Figure 1, the attachment circuit can be Ethernet or SDH. Ethernet access interfaces are typically protected using LAG, while SDH access interfaces are typically protected using MSP.

Linear MSP is a protection mechanism which protects the multiplex section layer. There are different implementations that extend this mechanism to support SDH sections that are terminated in different chassis. This document proposes using a new ICCP application to synchronize state and configuration data between two chassis to support multi-chassis MSP in the scenario shown in figure 1.

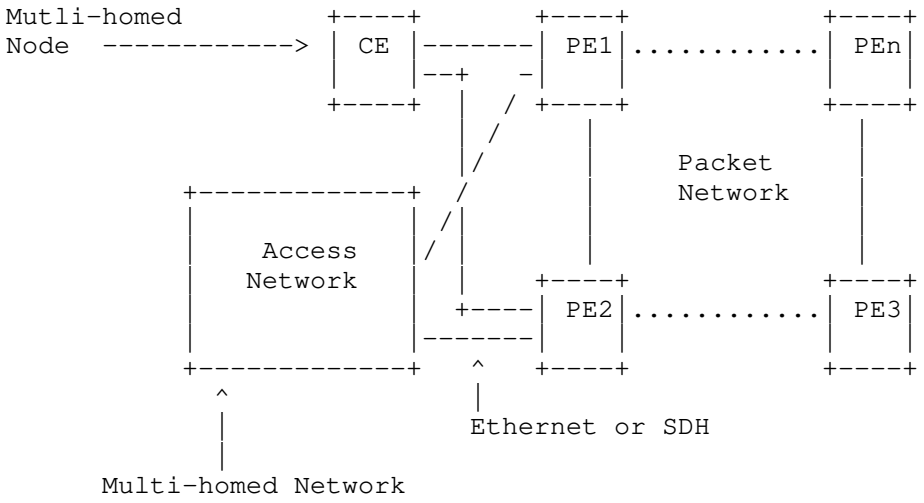


Figure 1: Attachment circuit multi-homed to Packet Network

1.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Terminology

- o AC: Attachment Circuit
- o AN: Access Network
- o CE: Customer Edge
- o ICCP:Inter-Chassis Communication Protocol
- o LACP: Link Aggregation Control Protocol
- o MSP: Multiplex Section Protection
- o PW: Pseudowire
- o RG: redundancy group
- o SDH: Synchronous Digital Hierarchy

3. ICCP extension requirements

3.1. Multi-chassis MSP Protection Model

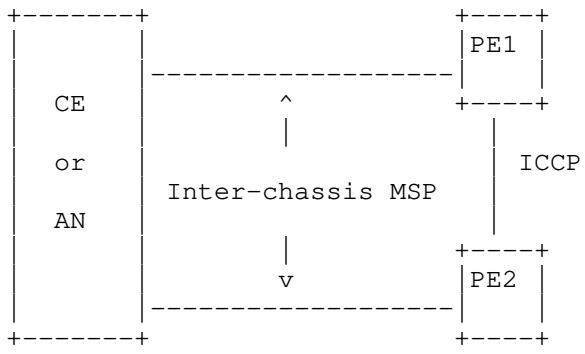


Figure 2: Generic Multi-chassis MSP Protection Model

Figure 2 describes the model where inter-chassis MSP is used as the AC redundancy mechanism. The SDH sections between CE/AN and PE1/PE2 form an inter-chassis protection group where one acts as the working section and the other as a protection section.

The PE that terminates the protection section SHALL process the MSP requests and calculate the bridge and selector states and the K1/K2 byte values to be transmitted, following MSP logic as specified in [G.841].

Whenever the output of the MSP logic changes, and when the MSP application initializes, the PE that terminates the protection section SHALL send the MSP group state to the other PE.

Each PE shall use the MSP group state to decide whether the PE is active or standby from an ICCP perspective.

For example, when the section between CE/AN and PE1 fails, the MSP group state at PE1 will change and PE1 will send a state update to PE2. After receiving and processing the information, the MSP group state at PE2 will change (assuming no other MSP requests exist) and PE2 will send an MSP group state update to PE1. As a result of this, PE2 will become the active PE and will act according to the procedures set out in [I-D.ietf-pwe3-iccp].

The same will occur as a result of external commands being applied to any of the PEs.

The ICCP application described in this document is responsible for

the state synchronization between different chassis forming a RG.

3.2. ICCP aspects

ICCP is specified in the [I-D:ietf-pwe3-iccp]. It allows synchronization of state and configuration data between a set of two or more PEs forming a RG. ICCP provides reliable message transport and in-order delivery between nodes in a RG with secure authentication mechanisms built into the protocol. Furthermore, it provides a common set of procedures by which applications on one PE can connect to their counterparts on another PE, for purpose of inter-chassis communication in the context of a given RG. The prerequisite for establishing an application connection is to have an operational ICCP RG connection between the two endpoints. When an application has information to transfer over ICCP, it triggers the transmission of an Application Data message. Currently, the ICCP draft has specified the ICCP's TLVs for the Pseudowire Redundancy application and the multi-chassis LACP (mLACP) application.

This draft extends ICCP TLVs to support MSP as an AC redundancy mechanism.

4. ICCP TLV extensions for MSP

The following sections specify the format of MSP application TLVs.

4.1. MSP connect TLV

This TLV is included in the RG Connect message to signal the establishment of MSP application connection.

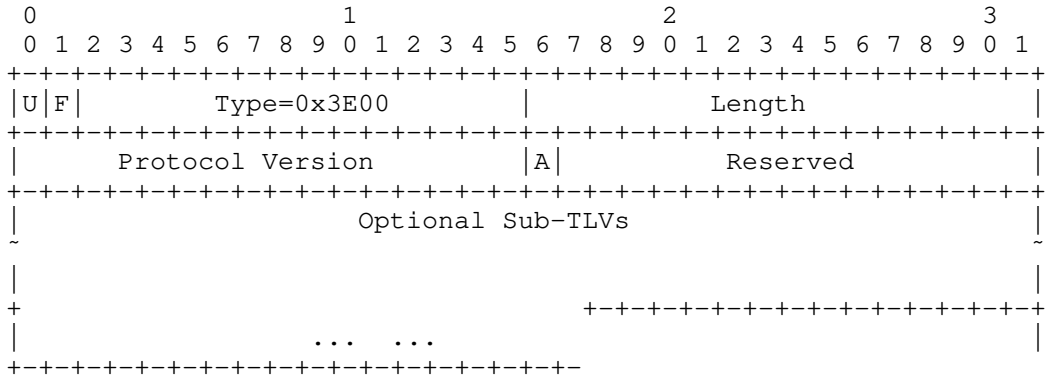


Figure 3: MSP connect TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E00 for "MSP connect TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Protocol Version
The version of this particular protocol for the purposes of ICCP. This is set to 0x0001.

- A Bit
Acknowledgement Bit. Set to 1 if the sender has received a MSP Connect TLV from the recipient. Otherwise, set to 0.

- Reserved
Reserved for future use.

- Optional Sub-TLVs
There are no optional Sub-TLVs defined for this version of the Protocol.

4.2. MSP disconnect TLV

This TLV is used in an RG Disconnect Message to indicate that the connection for the MSP application is to be terminated.

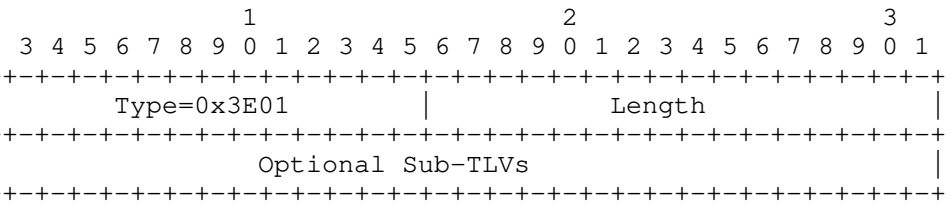


Figure 4: MSP disconnect TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E01 for "MSP disconnect TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Optional Sub-TLVs
There are no optional Sub-TLVs defined for this version of the Protocol.

4.2.1. MSP disconnect cause TLV

This TLV is used in an RG Disconnect Message to indicate the cause of disconnect message.

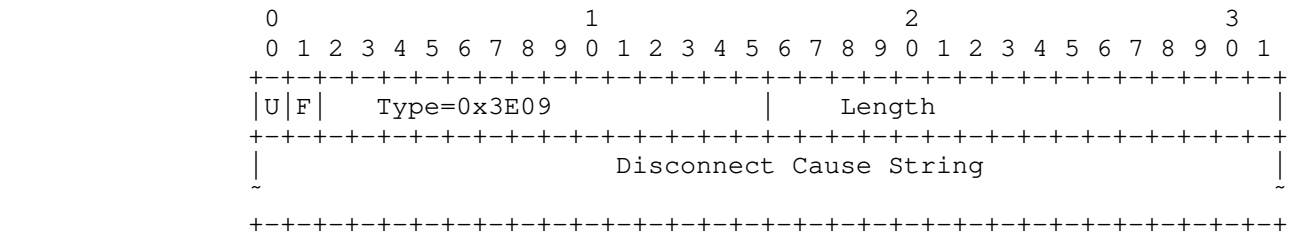


Figure 5: MSP disconnect TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E09 for "MSP disconnect cause TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Disconnect Cause String
Variable length string specifying the reason for the disconnect message. Used for network management.

4.3. MSP group config TLV

The MSP configuration TLV is sent in the RG application data message. This TLV is used to notify RG peers about the local configuration of protect group.

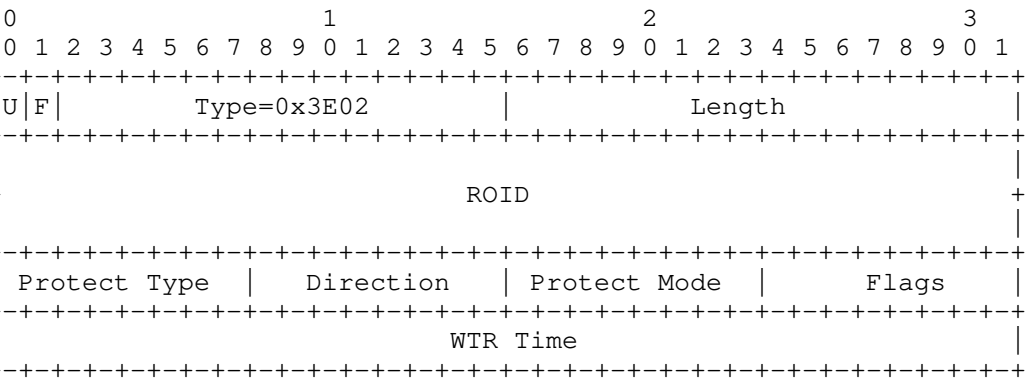


Figure 6: MSP group config TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E02 for "MSP group config TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- ROID
Defined in the [I-D:ietf-pwe3-iccp].Eight octets, uniquely identifies the Redundant Object.

- Protect Type
One octet encoding the protect type of the MSP protect group as follows:
0x00 1+1
0x01 1:1
0x02-0xFF reserved

- Direction
One octet encoding the architecture of the network as follows:
0x00 unidirectional
0x01 bidirectional

- Reversion Mode
One octet encoding the mode of operation as follows:
0x00 non-revertive operation
0x01 revertive operation

- Flags
One octet. Valid values are:
-i Synchronized (0x01)
Indicates that the sender has concluded transmitting all group configuration information.
-ii Purge Configuration (0x02)
Indicates that the group is no longer configured for MSP operation.

- WTR Time
Four octets. The time of waiting to restore, is used in the revertive mode of operation.

4.4. MSP port config TLV

The MSP port configuration TLV is sent in the RG application data message. This TLV is used to notify RG peers about the local port configuration.

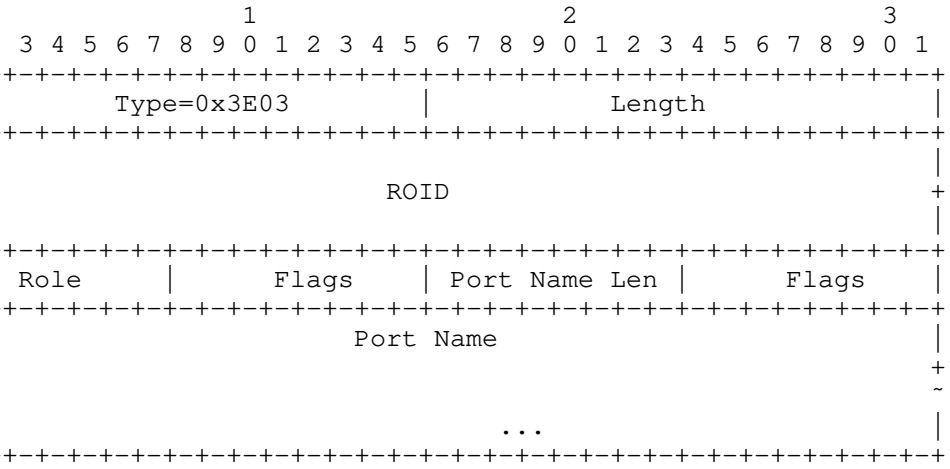


Figure 7: MSP port config TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E03 for "MSP group config TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- ROID
Defined in the [I-D:ietf-pwe3-iccp].Eight octets, uniquely identifies the Redundant Object.

- Role
One octet encoding the role of the section as follows:
0x00 working
0x01 protection

- Flags
One octet. Valid values are:
-i Synchronized (0x01)
Indicates that the sender has concluded transmitting all group configuration information.
-ii Purge Configuration (0x02)
Indicates that the group is no longer configured for MSP operation.

- Port Name Len
One octet, length of the "Port Name" field in octets.

- Port Name
Port name encoded in UTF-8 format, up to a maximum of 32 characters.

4.5. MSP section state TLV

The MSP section state TLV is sent in the RG application data message. This TLV announces the local section state to the RG peers.

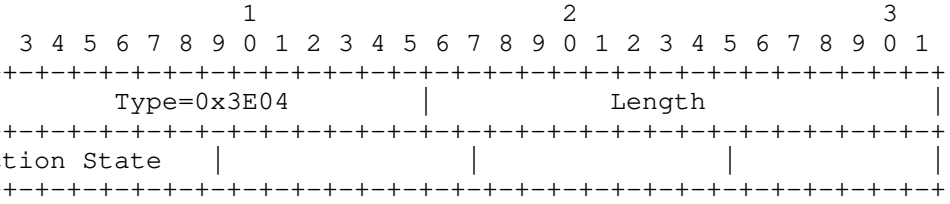


Figure 8: MSP section state TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E04 for "MSP section state TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length

fields.

- Section State
One octet encoding the section state as follows:
0x00 the signal is ok
0x01 Signal fail high priority
0x02 Signal fail low priority
0x03 Signal degrade high priority
0x04 Signal degrade low priority

4.6. MSP switch command TLV

The MSP configuration TLV is sent in the RG application data message. This TLV is used to notify RG peers about the local configuration of protect group.

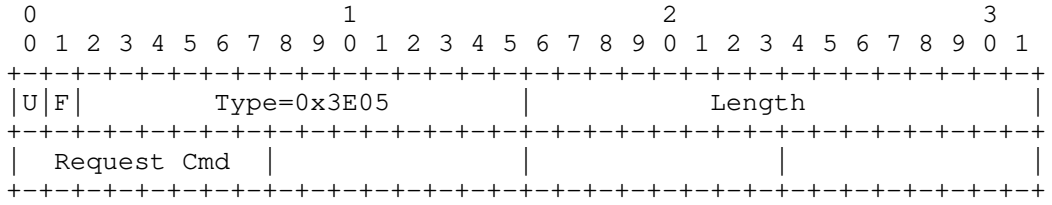


Figure 9: MSP switch command TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E05 for "MSP switch command TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Request Cmd
One octet.The switch command issued at the MSP APS controller interface. The following are the possible values, in order of priority from highest to lowest:
(1111) Clear
(1101) Lockout of protection(LP)
(1011) Forced Switch working-to-protection
(1001) Forced Switch protection-to-working
(0111) Manual switch working-to-protection
(0101) Manual switch protection-to-working
(0100) Exercise

4.7. MSP group state TLV

The MSP group state TLV is sent in the RG application data message. This TLV is used by the PE terminating the protection section to report the state of the MSP group to the other PE in the same RG.

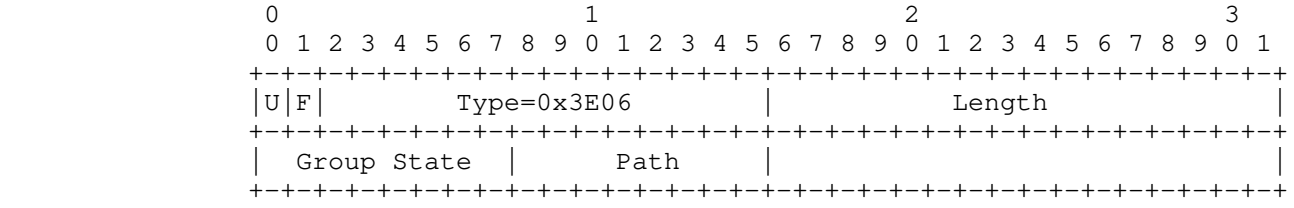


Figure 10: MSP group state TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E06 for "MSP group state TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Group State
One octet encoding the current state of the MSP protect group as follows:
0x00 No request
0x01 Do not revert
0x02 Reverse request
0x03 Unused
0x04 Exercise
0x05 Unused
0x06 Wait-to restore
0x07 Unused
0x08 Manual switch
0x09 Unused
0x0A Signal degrade low priority
0x0B Signal degrade high priority
0x0C Signal fail low priority
0x0D Signal fail high priority
0x0E Forced switch
0x0F Lockout of protection

- Path
One octet encoding the active path of the MSP protect group as follows:
0x00 the active path is the working section
0x01 the active path is the protection section

4.8. MSP Synchronization Request TLV

The MSP synchronization request TLV is used in the RG application data message. This TLV is used by a device to request from its peer to re-transmit configuration or operational state.

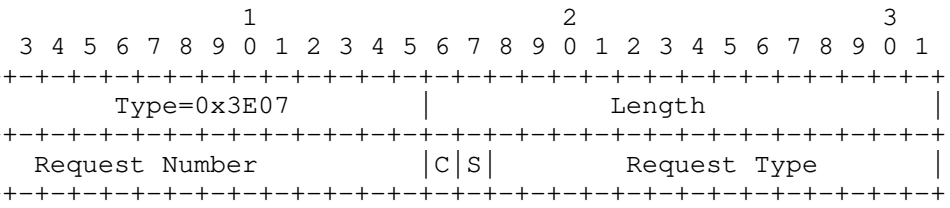


Figure 11: MSP Synchronization Request TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E07 for "MSP Synchronization Request TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Request Number
Two octets. Unsigned integer uniquely identifying the request. Be used to match the request with a response. The value of 0 is reserved for unsolicited synchronization, and MUST NOT be used in the MSP synchronization request TLV.

- C Bit
Set to 1 if request is for configuration data. Otherwise, set to 0.

- S Bit
Set to 1 if request is for running state data. Otherwise, set to 0.

- Request Type
14-bits specifying the request type, encoded as follows:

0x00 Request Data for specified protect group
0x01 Request Data for all groups in specified service(s)

4.9. MSP Synchronization Data TLV

The purpose of MSP Synchronization Data TLV is similar to the PW-RED Synchronization Data TLV defined in the [I-D:ietf-pwe3-iccp].It is used in the RG Application Data message. A pair of these TLVs is used by a device to delimit a set of TLVs that are sent in response to a MSP Synchronization Request TLV. The delimiting TLVs signal the start and end of the synchronization data, and associate the response with its corresponding request via the Request Number field.

The MSP Synchronization Data TLVs are also used for unsolicited advertisements of complete MSP configuration and operational state data. In this case, the Request Number field MUST be set to 0.

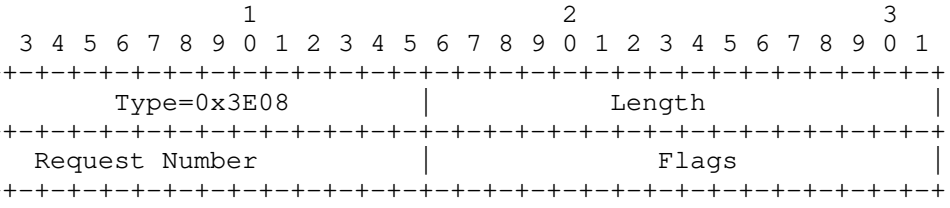


Figure 12: MSP group notify TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E08 for "MSP Synchronization Data TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.

- Request Number
Two octets. Unsigned integer is identifying the Request Number from the "MSP Synchronization Request TLV" which solicited this synchronization data response.

- Flags
Two octets, response flags encoded as follows:
0x00 Synchronization Data Start
0x01 Synchronization Data End

5. PE Node Failure

Section 9.2.3 of [I-D.ietf-pwe3-iccp] specifies the behavior in the event of PE node failure. Additionally. if the PE node detecting the remote PE failure is the one that terminates the protection section, it SHOULD transmit a signal fail request for the working section (SF-W) over the K1 byte and follow normal MSP procedure for this condition.

6. Security Considerations

The extensions of this document are based on ICCP and only some TLVs are added which will not change the security of existing network.

7. IANA Consideration

TBD.

8. References

8.1. Normative References

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

8.2. Informative References

[G.841] ITU-T Recommendation G.841, "Types and characteristics of SDH network protection architectures", 1998.

[I-D.ietf-pwe3-iccp] Luca Martini,Samer Salam,Ali Sajassi, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-07 .

Authors' Addresses

Hongjie Hao
ZTE Corporation

Email: hao.hongjie@zte.com.cn

Yuxia Ma
ZTE Corporation

Email: ma.yuxia@zte.com.cn

Weiqiang Cheng
China Mobile

Email: chengweiqiang@chinamobile.com

Daniel Cohn

Email: daniel.cohn.ietf@gmail.com

Masahiro Daikoku
KDDI Corporation

Email: ms-daikoku@kddi.com

Wanming Cao
ZTE Corporation

Email: cao.wanming@zte.com.cn

Jinghai Yu
ZTE Corporation

Email: yu.jinghai@zte.com.cn

Network Working Group
Internet-Draft
Updates: 4447, 6073 (if approved)
Intended status: Standards Track
Expires: January 4, 2013

L. Jin(ed.)
ZTE
R. Key(ed.)
Huawei
S. Delord
Alcatel-Lucent
T. Nadeau
Juniper
S. Boutros
Cisco Systems, Inc.
July 3, 2012

Pseudowire Control Word Negotiation Mechanism Update
draft-ietf-pwe3-cbit-negotiation-05.txt

Abstract

The control word negotiation mechanism specified in RFC4447 has a problem when a PE changes the preference for the use of the control word from NOT PREFERRED to PREFERRED. This document updates RFC4447 and RFC6073 by adding the Label Request message to resolve this control word negotiation issue for single-segment and multi-segment pseudowires.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 4, 2013.

Table of Contents

1. Introduction

The control word negotiation mechanism specified in [RFC4447] section 6.2 encounters a problem when a PE (Provider Edge) changes the preference for the use of the control word from NOT PREFERRED to PREFERRED. [RFC4447] specifies that if both endpoints prefer the use of the control word, then the pseudowire control word should be used. However, in the case whereby a PE changes its preference from NOT PREFERRED to PREFERRED and both ends of the PW (pseudowire) PE have the use of control word set as PREFERRED, an incorrect negotiated result of the control word as "not used" occurs. This document updates the control word negotiation mechanism in [RFC4447] by adding Label Request message to resolve this negotiation issue for single-segment PW. Multi-segment PW in [RFC6073] inherits the control word negotiation mechanism in [RFC4447], and this document updates [RFC6073] by adding the processing of Label Request message on S-PE. When PE changes the preference for the use of control word from PREFERRED to NOT PREFERRED, it should follow [RFC4447], and there is no problem.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Problem Statement

[RFC4447] section 6 describes the control word negotiation mechanism. Each PW endpoint has a configurable parameter that specifies whether the use of the control word is PREFERRED or NOT PREFERRED. During control word negotiation whereby one PE advertises a C bit set 0 in the label mapping message with its locally configured use of control word as PREFERRED and a corresponding peer PE changes its use of control word from NOT PREFERRED to PREFERRED causes an incorrect negotiated control word result of "not used".

The following case will describe the negotiation problem in detail:

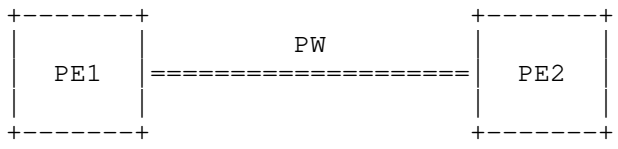


Figure 1

1. Initially, the use of control word on PE1 is configured as PREFERRED, and on PE2 as NOT PREFERRED.

2. The negotiation result for the control word of this PW is not used, and ultimately PE1 sends the Label Mapping message with C bit set to 0 according to [RFC4447] section 6.2.

3. PE2 then changes its use of control word configuration from NOT PREFERRED to PREFERRED, by deleting PW configuration with NOT PREFERRED use of control word, and configuring the PW again with PREFERRED use of control word.

4. PE2 will then send the Label Withdraw message to PE1, and correspondingly will receive the Label Release message from PE1.

5. According to the control word negotiation mechanism, the previously received Label Mapping message on PE2 from PE1 carries the C bit set to 0, therefore PE2 will still send the Label Mapping message with C bit set to 0.

The negotiation result for the control word is still not used, even though the use of control word configuration on both PE1 and PE2 are PREFERRED.

4. Control word renegotiation by Label Request message

The control word negotiation mechanism in [RFC4447] section 6 is updated to add the Label Request message described in this section.

The renegotiation process begins when the local PE has received the remote Label Mapping message with the C bit set to 0 and at the point a change occurs of its use of control word from NOT PREFERRED to PREFERRED. The following additional procedure will be carried out:

i. The local PE MUST send a Label Release message to remote PE. If local PE has previously sent a Label Mapping message, it MUST send a Label Withdraw message to remote PE, and wait until it has received a Label Release message from the remote PE. Note: the above Label Release message and Label Withdraw message sending

does not require specific sequence.

ii. The local PE MUST send a Label Request message to peer PE, and then MUST wait until it receives a Label Mapping message containing the peer's current configured preference for use of control word.

iii. After receiving the remote peer PE Label Mapping message with C bit, local PE MUST follow the procedures defined in [RFC4447] section 6 when sending its Label Mapping message.

The remote PE will follow [RFC4447], and once the remote PE has successfully processed the Label Withdraw message and Label Release message, it will reset its use of control word with the locally configured preference. Then the remote PE will send Label Mapping message with locally configured preference for use of control word as a response of Label Request message as specified in [RFC5036].

Note: for the local PE, before processing new configuration changing request, the above message exchanging process should be finished. The FEC (Forwarding Equivalence Class) element in the Label Request message should be the PE's local PW FEC element. As a response of the Label Request message, the peer PE should send Label Mapping message with its own local PW FEC element. The Label Request message format and procedure is described in [RFC5036].

4.1. Control word renegotiation for multi-segment PW

The multi-segment PW case for a T-PE (Terminating Provider Edge) operates similarly as the PE in single-segment PW described in the above section. An initial passive role is defined in [RFC6073] for S-PE (Switching Provider Edge) for the processing of Label Mapping message. [RFC6073] is updated by applying this passive role to the processing of Label Request message. When an S-PE receives a Label Request message from one of its adjacent PEs (may be S-PE or another T-PE), it MUST send a matching Label Request message to other adjacent PE (again, it may be an S-PE or a T-PE). This is necessary since an S-PE does not have complete information of the interface parameter field in the FEC advertisement. When the S-PE receives a Label Release message from remote PE, it MUST send a corresponding Label Release message to the other remote PE when it holds a label for the PW from the remote PE.

Note: because the local T-PE will send Label Withdraw message before sending Label Request message to the remote peer, the S-PE MUST process the Label Withdraw message before the Label Request message. When the S-PE receives the Label Withdraw message, it should process this message to send a Label Release message as a response and a

Label Withdraw message to upstream S-PE/T-PE. The S-PE will then process the next LDP message, e.g. the Label Request message.

When the local PE changes the use of control word from PREFERRED to NOT PREFERRED, the local PE would then renegotiate the PW control word to be not used by deleting the PW configuration with PREFERRED use of control word, and configuring the PW again with NOT PREFERRED use of control word. All of these procedures have been defined in [RFC4447] section 5.4.1.

The diagram in Appendix A in this document updates the control word negotiation diagram in [RFC4447] Appendix A.

4.2. Control word re-negotiation use case

The procedure of PE1 and PE2 for the use case in figure 1 will become as follows:

1. PE2 changes locally configured preference for use of control word to PREFERRED.

2. PE2 will then send the Release messages to PE1. PE2 will also send the Label Withdraw message, and wait until it has received the Label Release message from PE1.

3. PE1 will send the Label Release message in response to the Label Withdraw message from PE2. After processing the Label Release from PE2, PE1 will then reset the use of control word to the locally configured preference as PREFERRED.

4. Upon receipt of the Label Release message from PE1, PE2 will send the Label Request message to PE1, and proceed to wait until a Label Mapping message is received.

5. PE1 will send a Label Mapping message with C bit set to 1 again to PE2 as response of the Label Request message.

6. PE2 receives the Label Mapping message from PE1 and gets the remote label binding information. PE2 will wait for the PE1 Label Mapping message before sending its Label Mapping message with C bit set.

7. PE2 will send the Label Mapping to PE1 with C bit set to 1, and follow procedures defined in [RFC4447] section 6.

While it is assumed that PE1 is configured to prefer the use of the control word, in step 5 if PE1 doesn’t prefer or support the control word, PE1 would then send the Label Mapping message with C bit set to

0. As a result, PE2 in step 7 would send a Label Mapping message with C bit set 0 as per [RFC4447] section 6.

By sending a Label Request message, PE2 will get the locally configured preference for use of control word of peer PE1 in the received Label Mapping message. By using the new C bit from the Label Mapping message received from peer PE1 and the locally configured preference for use of control word, PE2 should determine the use of PW control word according to [RFC4447] section 6.

5. Backward Compatibility

Since control word negotiation mechanism is updated by adding Label Request message, and still follows the basic procedure described in [RFC4447] section 6, this document is fully compatible with existing implementations. For single-segment pseudowire, the remote PE (PE1 in figure 1) which already implements [RFC4447] and Label Request message as defined in [RFC5036] could be compatible with the PE (PE2 in figure 1) following the mechanism of this document. For the multi-segment pseudowire, the T-PE is same as PE in single-segment pseudowire; the S-PE should be upgraded with the mechanism defined in this document.

6. Security Considerations

The security considerations specified in [RFC4447] and [RFC6073] also apply to this document, and this document does not introduce any additional security constraints.

7. IANA Considerations

This document does not require IANA assignment.

8. Acknowledgements

The authors would like to thank Stewart Bryant, Andrew Malis, Nick Del Regno, Luca Martini, Venkatesan Mahalingam, Alexander Vainshtein, Adrian Farrel and Spike Curtis for their discussion and comments.

9. Contributing Authors

Vishwas Manral
Hewlett-Packard Co.

19111 Pruneridge Ave, Bldg 44,
Cupertino, CA 95014-0691
Email: vishwas.manral@hp.com

Reshad Rahman
Cisco Systems, Inc.
2000 Innovation Drive Ottawa,
Ontario K2K 3E8
CANADA
Email: rrahman@cisco.com

10. Normative references

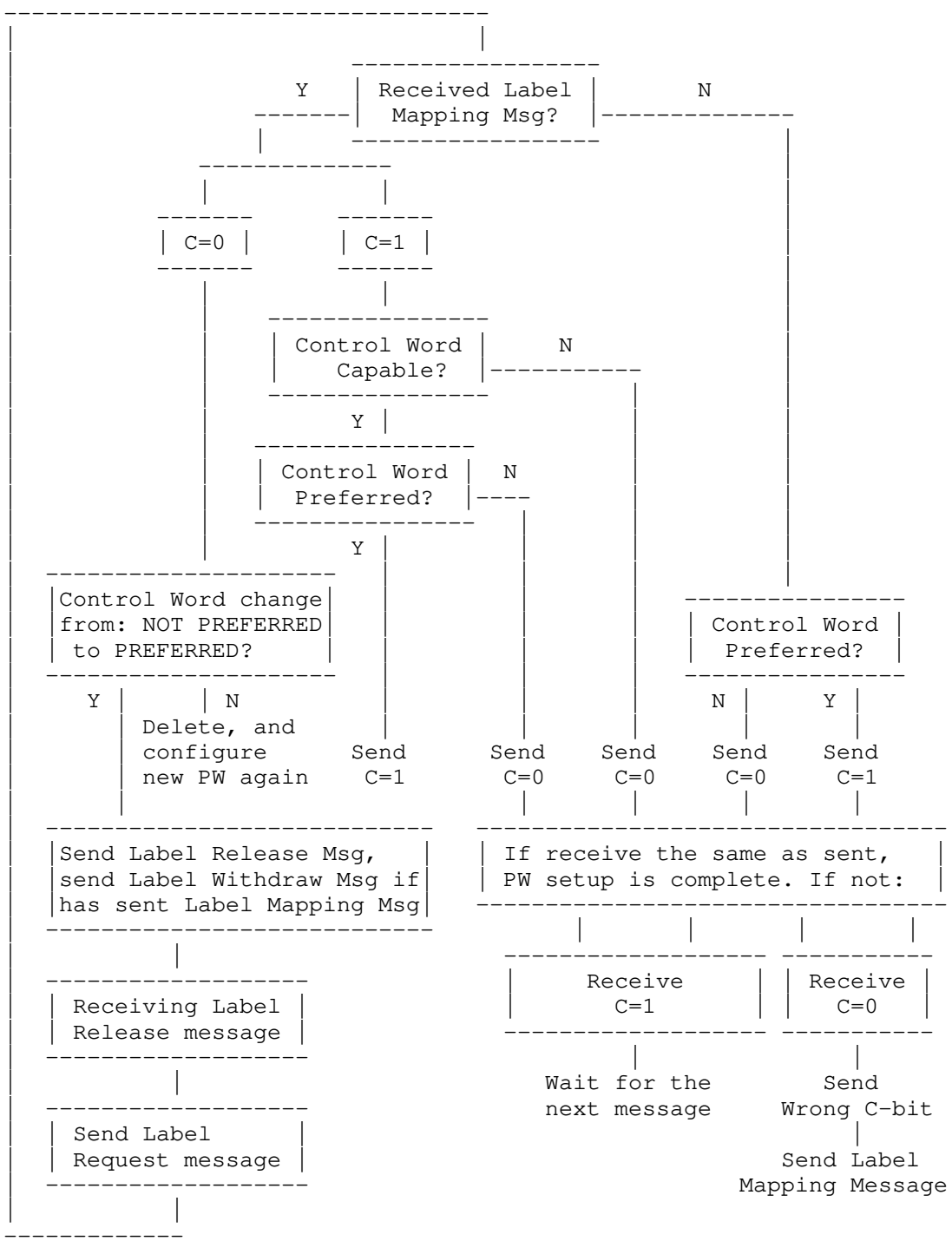
[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC6073] Martini, L., Metz, C., Nadeau, T., Bocci, M., and M. Aissaoui, "Segmented Pseudowire", RFC 6073, January 2011.

Appendix A. Updated C-bit Handling Procedures Diagram



Authors' Addresses

Lizhong Jin (editor)
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China

Email: lizhong.jin@zte.com.cn

Raymond Key (editor)
Huawei

Email: raymond.key@ieee.org

Simon Delord
Alcatel-Lucent

Email: simon.delord@gmail.com

Thomas Nadeau
Juniper

Email: tnadeau@juniper.net

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way
San Jose, California 95134, USA

Email: sboutros@cisco.com

INTERNET-DRAFT
PWE3 WG
Intended Status: Standard Track
Expires: November 2011

David L. Black (ed.)
EMC Corporation
Linda Dunbar(ed.)
Huawei Technologies
Moran Roth
Infinera
Ronen Solomon
Orckit-Corrigent
May 3, 2011

Encapsulation Methods for Transport of Fibre Channel Traffic over MPLS Networks

draft-ietf-pwe3-fc-encap-16.txt

Status of this Memo

Copyright Notice

Abstract

A Fibre Channel pseudowire (PW) is used to carry Fibre Channel traffic over an MPLS network. This enables service providers to take advantage of MPLS to offer "emulated" Fibre Channel services. This document specifies the encapsulation of Fibre Channel traffic within a pseudowire. It also specifies the common procedures for using a PW to provide a Fibre Channel service.

Conventions used in this document

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

Table of Contents

1. Introduction

Fibre Channel (FC) is a high-speed communications technology, used primarily for Storage Area Networks (SANs). Within a single site (e.g., data center), an FC-based SAN connects servers to storage systems, and FC can be extended across sites. When FC is extended across multiple sites, the most common usage is storage replication in support of recovery from disasters (e.g., flood or fire that takes a site out of operation). This is particularly the case over longer distances where network latency results in unacceptable performance for a server whose storage is not at the same site. Fibre Channel is standardized by INCITS Technical Committee T11 [T11] and multiple methods for encapsulating and transporting FC traffic over other networks have been developed [FC-BB-6].

FCIP, as described in [RFC3821] and [FC-BB-6], interconnects otherwise isolated FC SANs over IP Networks. FCIP uses FC Frame Encapsulation [RFC3643] to encapsulate FC frames for tunneling over an IP-based network. Since IP networks may drop or reorder packets, FCIP relies on TCP to retransmit dropped frames and restore the delivery order of reordered frames. Due to possible delay variation and TCP timeouts, special timing mechanisms are required to ensure correct Fibre Channel operation over FCIP [FC-BB-6].

MPLS networks can be provisioned and operated with very low loss rates and very low probability of reordering, making it possible to directly interconnect Fibre Channel ports over MPLS. A Fibre Channel pseudowire (FC PW) is a method to transparently transport FC traffic over an MPLS network resulting in behavior similar to a pair of FC ports that are directly connected by a physical FC link. The result is simpler control processing by comparison to FCIP.

This document specifies the encapsulation of FC traffic into an MPLS pseudowire and related PW procedures to transport FC traffic over MPLS PWs. The complete FC pseudowire specification consists of this document and the FC PW portion of the T11 [FC-BB-6] standard. The following subsections describe some of the requirements for transporting FC traffic over an MPLS network.

1.1. Transparency

Transparent extension of an FC link is a key requirement for transporting FC traffic over a PW. This requires the FC PW to emulate an FC Link between two FC ports, similar to the approach defined for FC over GFPT in [FC-BB-6]. GFPT is an Asynchronous Transparent Generic Framing Procedure specified by ITU-T, see [FC-BB-6] for details and reference to the ITU-T specifications. This results in

transparent forwarding of FC traffic over the MPLS network from both the FC Fabric and the network operator points of view.

Transparency distinguishes the FC PW approach from FCIP. An FC PW logically connects the FC port on the FC link attached to one end of the PW directly with the FC port on the far end of the FC link attached to the other end of the PW, whereas FCIP introduces FC B_Ports at both ends of the extended FC link; each FC B_Port is connected to an FC E_Port in an FC switch on the same side of the link extension.

1.2. Bandwidth Efficiency

The bandwidth allocated to a PW may be less than the rate of the attached FC port. When there is no data exchange on a native FC link, Idle Primitive Signals are continuously exchanged between the two FC ports. In order to improve the bandwidth efficiency across the MPLS network, it is necessary for the FC PW PE to suppress (or drop) the Idle Primitive signals generated by its adjacent FC ports. The far end FC PW PE regenerates Idle Primitive signals to send to its adjacent FC port as required, see [FC-BB-6].

FC link control protocols require an FC port to continuously send the same FC Primitive Sequence [FC-FS-2] until a reply is received or some other event occurs. To improve bandwidth efficiency, the FC PW PE encapsulates a subset of repeated FC Primitive Sequences to send across the WAN [FC-BB-6]. For example, in a sequence of identical received primitives, only every fourth primitive may be sent across the MPLS network. Alternatively, a time-based approach may be used to send a copy of the repeated FC Primitive Sequence once every few milliseconds. The far end FC PW PE regenerates the FC link behavior by continuously sending the Primitive Sequence most recently received from the WAN until a new primitive signal, primitive sequence or data frame is received from the WAN.

The sending FC PW PE may unilaterally choose any convenient subset for sending the same FC Primitive Sequence. This is acceptable because the receiving FC PW PE generates a continuous stream of the most recently received FC Primitive Sequence on the outgoing native FC link, independent of the arrival rate of that FC Primitive Sequence from the WAN. In practice, a 10:1 reduction in FC Primitive Sequence transmission rate achieves 90% of the bandwidth benefits without loss of FC functionality and sending a copy every few milliseconds does not pose a serious risk of exceeding the timeouts specified in Section 5 below.

These bandwidth efficiency techniques may cause changes in the FC traffic that traverses an FC PW (e.g., number of IDLE signals or number of identical Primitive Sequences), but the far end FC PW PE's regeneration of FC link behavior on the attached FC port is transparent to the FC ports connected to each PW PE.

1.3. Reliability

Fibre Channel does not employ a native frame retransmission protocol, and treats most frame delivery failures as errors. FC SAN traffic requires a very low frame loss rate because the typical result of a failure to deliver a frame is an I/O operation failure. Recovery from such I/O failures involves I/O operation retries after what may be a significant delay (30 second and 60 second timeouts are common). In addition, such retries are likely to be logged as errors indicating possible problems with FC equipment or cables. Hence, drops, errors and discards of FC frames must be very rare for an FC PW.

FC SAN implementations have limited tolerance for frame reordering. Any reordering affecting more than a few frames within a single higher level operation (e.g., a read or write I/O) is usually treated as an error by the destination FC port, resulting in discards of the frames involved; some deployed FC implementations treat all such within-operation frame reordering as errors that result in frame discards. As a result, FC frame reordering must be minimized for an FC PW.

The FC PW does not compensate for frame drops, discards or reordering. The MPLS network that hosts the FC PW is expected to be designed and operated in a fashion that makes such events very rare.

In contrast to the TTL field in an IP packet, FC uses a constant delivery timeout value (R_A_TOV) for which 10 seconds is the default. Each FC frame must be delivered or discarded within that timeout period after it is sent, see Section 5.

2. Reference Model

An FC PW extends a native FC link over an MPLS network. This document specifies the PW encapsulation for FC. Figure 1 describes the reference models (derived from [RFC3985]) that support the FC PW. FC traffic is received by PE1's FC attachment channel, encapsulated at PE1, transported across MPLS network, decapsulated at PE2, and transmitted onward via the PE2's FC attachment channel. This document assumes that a pseudowire can be provisioned statically or via a signaling protocol as defined in [RFC4447].

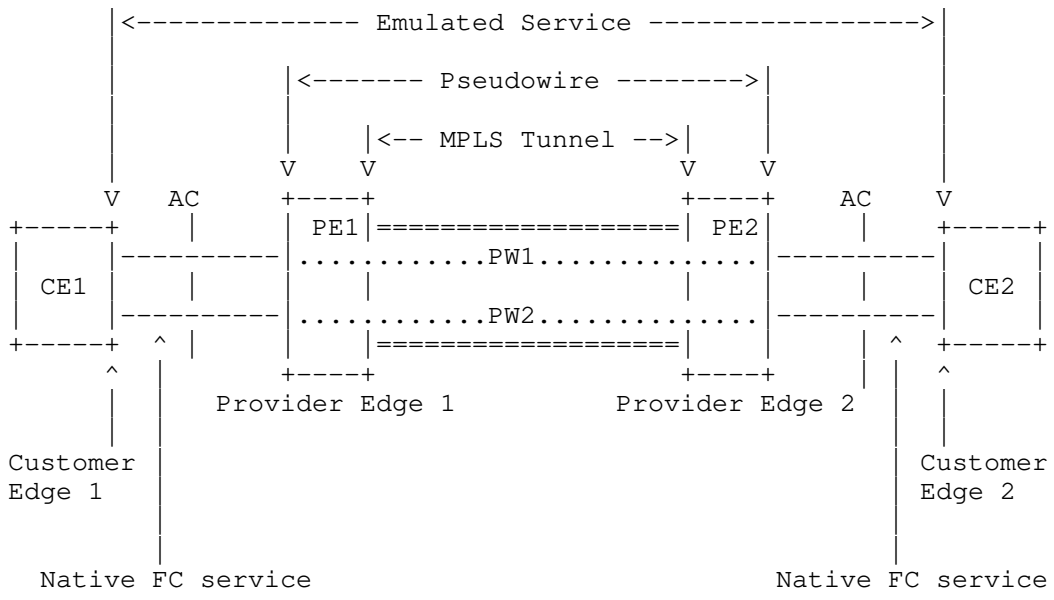


Figure 1: PWE3 FC Interface Reference Configuration

The following reference model describes the termination point of each end of the PW within the PE:

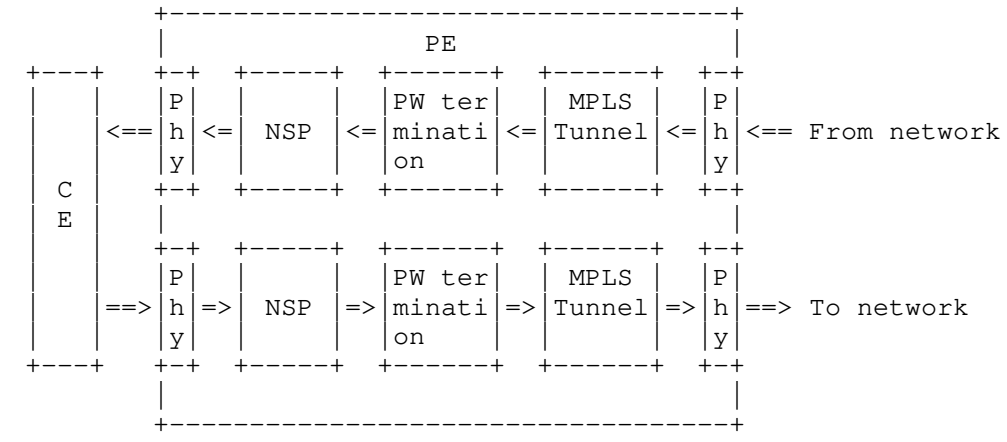


Figure 2: PW reference diagram

The Native Service Processing (NSP) function includes the following functionality:

- Idle Suppression: any FC Idle signals received from the source PE's attached FC port are suppressed and re-generated at the destination PE to send on its attached FC port when there is no other FC traffic to send;

- FC Primitive Sequence Reduction: a subset of repetitive FC Primitive Sequences received from the attached FC port at the source PE is selected for WAN transmission, with the destination PE sending the FC Primitive Sequence most recently received from the WAN on the destination PE's attached FC port continuously until a new packet is received from the WAN; and

- Flow Control: the Alternate Simple Flow Control (ASFC) protocol is used for buffer management in concert with the peer PW PE's NSP function so that FC traffic is not dropped. ASFC is a simple pause/resume protocol that allows operation repetition; the receiver responds to the first pause or resume operation in an identical sequence of operations, and ignores the rest of the sequence.

The NSP flow control functionality is required to extend FC's credit-based flow control to address situations where the number of buffer credits available to an FC link is insufficient to utilize the available bandwidth over the additional distance and latency represented by the FC pseudowire. The NSPs avoid this problem by inserting ASFC into FC's link flow control used on the attached FC ports, see [FC-BB-6].

In contrast, Idle Suppression and FC Primitive Sequence Reduction are bandwidth optimizations that are included in the NSP for clarity in this document. Analogous optimizations are not treated as part of the NSP by other pseudowires (e.g., ATM idle frame suppression is not considered to be an NSP function by [RFC4717]).

The NSP function is specified in detail by [FC-BB-6].

3. Encapsulation

This specification provides port to port transport of FC encapsulated traffic. There are a number of port types defined by Fibre Channel, including:

- An N_port is a port on the node (e.g. host or storage device) used with both FC-P2P (Point to Point) or FC-SW (Switched fabric) topologies. Also known as a Node port.
- An NL_port is a port on the node used with an FC-AL (Arbitrated Loop) topology. Also known as a Node Loop port.
- An F_port is a port on the switch that connects to a node point-to-point (i.e. connects to an N_port). Also known as a Fabric port. An F_port is not loop capable.
- An FL_port is a port on the switch that connects to a FC-AL loop (i.e. to NL_ports). Also known as Fabric Loop port.
- An E_port is a port used to connect two Fibre Channel switches. Also known as an Expansion port. When E_ports between two switches are connected to form a link, that link is referred to as an inter-switch link (ISL).

Among the port types listed above, only the following FC connections (as specified in [FC-BB-6]) are supported by an FC PW over MPLS:

- N_Port to N_Port, established by an FC PLOGI (Port Login) operation
- N_Port to F_Port, established by an FC FLOGI (Fabric Login) operation
- E_Port to E_Port, established by an FC ELP (Exchange Link Parameters) operation

FC traffic flowing over an FC PW is subdivided into four payload types (PT) that are encoded in the PW Control Word (see Section 3.1):

1. FC login traffic (PT = 1): FC login operations and responses that establish connections between FC ports. The three FC login operations are PLOGI, FLOGI, and ELP. These operations and their responses may require the NSP to allocate buffer resources, see the specification of Login Exchange Monitors in [FC-BB-6].

2. FC data traffic (PT = 0): All FC frames other than those involved in an FC login operation.

3. FC Primitive Sequences and Signals (PT = 2): Native FC link control operations - 4-character primitive sequences and signals that are not encapsulated in FC frames. See [FC-BB-6] and [FC-FS-2].

4. FC PW Control (PT = 6): FC PW control operations exchanged only between the endpoints of the PW. FC PW control operations are used for ASFC flow control, ping (e.g., for round trip latency measurement) and reporting native FC link errors, see [FC-BB-6].

This FC PW specification is limited to use with FC service classes 2, 3 and F (see [FC-FS-2]). Other FC service classes (e.g., 1, 4 and 6) MUST NOT be used with an FC PW. Numbered FC service classes are used for end-to-end FC traffic, whereas service class F is used for inter-switch traffic in an FC switched fabric.

This FC PW specification is limited to native FC attachment links that employ an 8b/10b transmission code (see [FC-FS-2]). The protocol specified in this document converts a received 10b code to its 8b counterpart for PW encapsulation, and hence does not support attached FC links that use a 64b/66b transmission code (e.g., 10GFC, 16GFC); such links MUST NOT be attached to an FC PW PE unless their link speed can be negotiated to one that uses 8b/10b encoding. If an invalid 10b code that cannot be converted to an 8b code is received from an FC link, the PE sends an FC PW control frame to report the error, see [FC-BB-6].

3.1. The Control Word

The Generic PW Control Word, as defined in "PWE3 Control Word" [RFC4385] MUST be used for FC PW to facilitate the transport of short packets (by setting the Length field as detailed below), and convey the flag bits defined below. The structure of the Control Word for the FC PW is as follows:

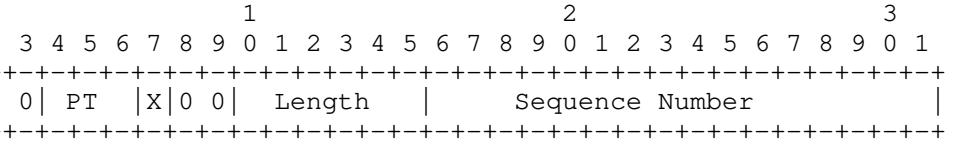


Figure 3 - Control Word Structure

The first four bits of the PW Control Word MUST be set to 0 by the ingress PE to indicate PW data.

Three of the four Flags bits are used to convey the PT - Payload Type indication. The 3-bit binary value in this field identifies the payload type carried by a PW packet. The following types are defined:

PT = 0: FC data frame.

PT = 1: FC login frame.

PT = 2: FC Primitive Sequence(s) and/or Primitive Signal(s).

PT = 6: FC PW Control Frame (refer to [FC-BB-6] for usage).

Packets with other values in the PT field are not valid for the FC PW and MUST be discarded by the receiving FC PW PE.

X - This flag bit is not used by this version of the protocol. It SHOULD be set to zero by the sender and MUST be ignored by the receiver.

The fragmentation bits (bits 8-9) are not used by the FC PW protocol. These bits may be used in the future for FC specific indications as defined in [RFC4385]. The fragmentation bits SHOULD be set to zero by the ingress PE and MUST be ignored by the egress PE.

The Length field enables recovery of the original pseudowire packet when a short packet is padded to the minimum 64 octet packet size required for Ethernet, see [RFC4385]. The Length field MUST be used

for packets shorter than 64 octets, MUST be set to zero for longer packets, and MUST be processed according to the rules specified in [RFC4385].

The sequence number is not used for the FC PW and MUST be set to 0 by the ingress PE, and MUST be ignored by the egress PE.

3.2. MTU Requirements

The MPLS network MUST be able to transport the largest Fibre Channel frame after encapsulation, including the overhead associated with the encapsulation. The maximum FC frame size is 2164 octets without PW and MPLS labels (refer to Figure 4); this maximum size is a constant value that is required for all FC implementations [FC-FS-2]. The MPLS network SHOULD accommodate frames of up to 2500 octets in order to support possible future increases in the maximum FC frame size.

Fragmentation, as described in [RFC4623], SHALL NOT be used for an FC PW, therefore the network MUST be configured with a minimum MTU that is sufficient to transport the largest encapsulated FC frame.

3.3. Mapping of FC traffic to PW packets

FC frames, Primitive Sequences, and Primitive Signals are transported over the PW. All packet types are carried over a single PW. In addition to the PW Control Word, an FC Encapsulation Header is included in the PW packet. This FC Encapsulation Header is not used in this version of the protocol; it SHOULD be set to zero by the sender and MUST be ignored by the receiver.

3.3.1. FC Data Frames (PT=0) and FC Login Frames (PT=1)

FC data frames and FC login frames share a common encapsulation format, except that the PT field in the FC PW control word is set to 0 for data frames and is set to 1 for login frames. An FC login frame contains an FC PLOGI, FLOGI or ELP operation or response that requires special processing by the NSP in support of flow control, see [FC-BB-6].

Each FC data frame or login frame is mapped to a PW packet, including the Start Of Frame (SOF) delimiter, frame header, CRC field and the End Of Frame (EOF) delimiter, as shown in figure 4.

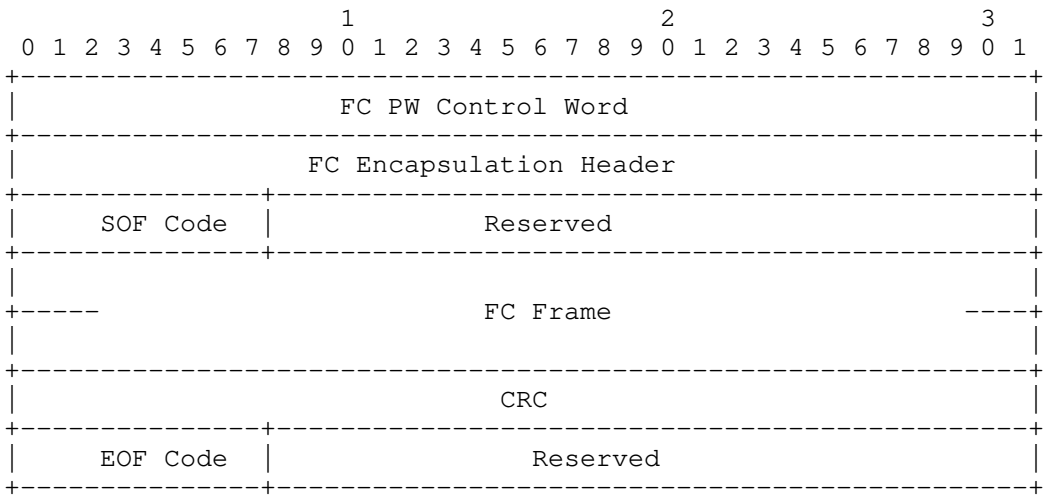


Figure 4 - FC frame (SOF/Data/CRC/EOF) encapsulation in PW packet

The SOF and EOF frame delimiters are each encoded into a single octet as specified in [RFC3643], except that the codes for delimiters that apply only to FC service class 4 (SOFi4, SOFc4, SOFn4, EOFdt, EOFdti, EOFrt, EOFrti - see [FC-FS-2]) MUST NOT be used.

The CRC in the frame is obtained directly from the FC attachment channel, so that the PW PE is not required to re-calculate the CRC or to check the CRC in the received frame. The CRC will be checked by the FC port that receives the frame, ensuring that coverage is provided for data errors that occur between the PW endpoints. This CRC behavior differs from the FCS retention technique for PWs defined in [RFC4720] which states that "as usual, the FCS MUST be examined at the ingress PE, and errored frames MUST be discarded."

3.3.2. FC Primitive Sequences and Primitive Signals (PT=2)

FC Primitive Sequences and Primitive Signals are FC ordered sets. On an 8b/10b-coded FC link, an ordered set consists of four 10b characters, starting with the K28.5 character, followed by three Dxx.y data characters. All FC ordered sets start with a K28.5 control character, but the three following Dxx.y data characters differ depending on the ordered set. A Kxx.y control character has a different 10b code from the corresponding Dxx.y data character, but uses the same 8b code (e.g., K28.5 and D28.5 both use the 8b code 0xBC). Here are two examples of ordered sets:

- Idle(IDLE) is K28.5 - D21.4 - D21.5 - D21.5. This FC primitive signal is sent when the FC link is idle; it is suppressed by the FC PW NSP and not sent over the WAN.

- Link Reset Response(LRR) is K28.5 - D21.1 - D31.5 - D9.2 (this FC primitive sequence is used as part of FC link initialization and recovery).

Each ordered set is encapsulated in a PW packet containing the encoded K28.5 control character [FC-BB-6], followed by three encoded data characters, as shown in Figure 5.

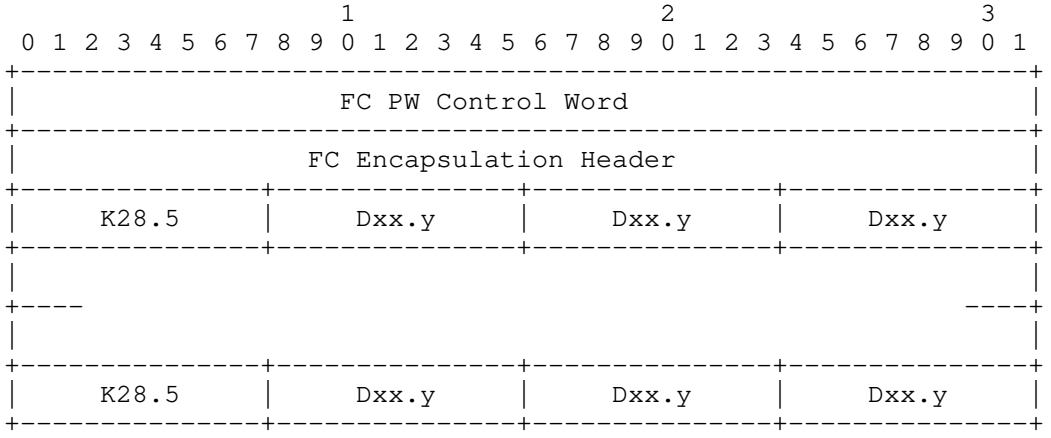


Figure 5 - FC Ordered Sets encapsulation in PW packet

The K28.5 10b control character received from the PE's attached FC link is encoded for the FC PW as its 8b counterpart (0xBC). Because the same 8b value (0xBC) is used to encode a D28.5 data word, the receiving FC PW PE:

- MUST check for presence of an 8b K28.5 value (0xBC) at the start of each ordered set (see Figure 5), and MUST send that value as a 10b K28.5 character on the attached FC link.

- MUST send the following three Dxx.y 8b values as Dxx.y 10b characters on the attached FC link and MUST NOT send any of these Dxx.y 8b values as 10b Kxx.y characters on the attached FC link.

A PW packet may contain one or more encoded FC Ordered sets [FC-BB-6]. The Length field in the FC PW Control Word is used to indicate the packet length when the PW packet contains multiple Ordered Sets.

For this reason, FC PW packets that contain FC Ordered Sets MUST NOT be larger than 60 octets (8 octets of header words plus at most 13 ordered sets), in order to ensure that the Length field contains a non-zero value, see [RFC4385].

Idle Primitive Signals could be carried over the PW in the same manner as Primitive Sequences. However, [FC-BB-6] requires that Idle Primitive Signals be dropped by the Ingress PE and re-generated by the egress PE in order to reduce bandwidth consumption (see [FC-BB-6] for further details).

The egress PE extracts the Primitive Sequence or Primitive Signal from the received PW packet. For a Primitive Sequence, the PE continues transmitting the same FC Ordered Set to its attached FC port until an FC frame or another ordered set is received over the PW; see Section 1.2 above for discussion of ingress PE transmission behavior for Primitive Sequences. A Primitive Signal is sent once, except that Idle Primitive Signals are sent continuously when there is nothing else to send.

3.3.3. FC PW Control Frames (PT=6)

FC PW Control Frames are transported over the PW, by encapsulating each frame in a PW packet with PT=6 in the Control Word. FC PW Control Frame payloads are generated and terminated by the corresponding FC entity. FC PW Control frames are used for FC PW flow control (ASFC), ping and transmission of error indications. [FC-BB-6] specifies the generation and processing of FC PW Control Frames. FC PW Control Frames are always shorter than 64 octets, and hence the Length field in the FC Control Word indicates their length.

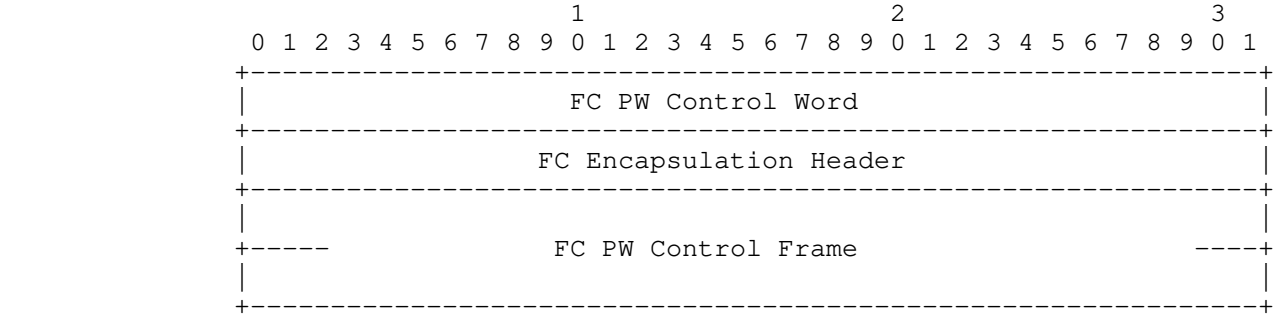


Figure 6 - FC PW Control frame encapsulation in PW packet

3.4. PW failure mapping

PW failures are detected through PW signaling failure, PW status notifications as defined in [RFC4447], or through PW OAM mechanisms and MUST be mapped to emulated signal failure indications. Sending the FC link failure indication to its attached FC link is performed by the NSP, as defined by [FC-BB-6].

4. Signaling of FC Pseudowires

RFC4447 specifies the use of the MPLS Label Distribution Protocol, LDP, as a protocol for setting up and maintaining pseudowires. This section describes the use of specific fields and error codes used to control FC PW.

The PW Type field in the PWid FEC element and PW generalized ID FEC elements MUST be set to the "FC Port Mode" value in section 8 below.

The Control Word is REQUIRED for FC pseudowires. Therefore the C-Bit in the PWid FEC element and PW generalized ID FEC elements MUST be set. If the C-Bit is not set, the pseudowire MUST NOT be established and a Label Release MUST be sent with an "Illegal C-Bit" status code [RFC4447].

The Fragmentation Indicator (Parameter ID = 0x09) is specified in [RFC4446] and its usage is defined in [RFC4623]. Since fragmentation is not used in FC PW, the fragmentation indicator parameter MUST be omitted from the Interface Parameter Sub-TLV.

The Interface MTU Parameter (Parameter ID = 0x01) is specified in [RFC4447]. Since all FC interfaces have the same MTU, this parameter MUST be omitted from the Interface Parameter Sub-TLV.

The FCS Retention Indicator (Parameter ID = 0x0A) is specified in [RFC4720]. Since the CRC treatment defined in this document differs from one that is specified in [RFC4720], this parameter MUST be omitted from the Interface Parameter Sub-TLV.

5. Timing Considerations

Correct Fibre Channel link operation requires that the FC link latency between CE1 and CE2 (refer to Figure 1) be:

- no more than one-half of the R_T_TOV (Receiver Transmitter Timeout Value, default value: 100 milliseconds) of the attached devices for Primitive Sequences;

- o no more than one-half of the E_D_TOV (Error Detect Timeout Value, default value: 2 seconds) of the attached devices for frames; and

- o within the R_A_TOV (Resource Allocation Timeout Value, default value: 10 seconds) of the attached fabric(s), if any. The FC standards require that the E_D_TOV value for each FC link be set so that the R_A_TOV value for the fabric is respected when the worst case latency occurs for each link, see [FC-FS-2].

An FC PW MUST adhere to these three timing requirements and MUST NOT be used in environments where high or variable latency may cause these requirements to be violated.

These three timeout values are ordered (R_T_TOV < E_D_TOV < R_A_TOV), so adherence to one-half of R_T_TOV for all FC PW traffic is sufficient. See [FC-FS-2] for definitions of the FC timeout values.

The R_T_TOV is used by the FC link initialization protocol. If an FC PW's latency exceeds one-half R_T_TOV, initialization of the FC link that is encapsulated by the FC PW may fail, leaving that FC link in a non-operational state.

The E_D_TOV is used to detect failures of operational FC links. If an FC PW's latency exceeds the one-half E_D_TOV requirement, the FC link that is encapsulated by the FC PW may fail. The usual FC response to such a link failure is to attempt to recover the FC link by initializing it. That initialization will also fail if the FC PW latency exceeds one-half R_T_TOV (a tighter requirement).

The R_A_TOV is used to determine when FC communication resources (e.g., values that identify FC frames) may be reused. If an FC PW's violation of the one-half E_D_TOV requirement is sufficient to also cause the FC fabric to violate the R_A_TOV requirement, then FC reuse of frame identification values after an R_A_TOV timeout may result in multiple FC frames with the same identification values, causing incorrect Fibre Channel operation. For example, if two such frames are swapped between I/O operations, the result may corrupt data in the I/O operations.

The PING and PING_ACK FC PW control frames defined in Section 6.4.7 of [FC-BB-6] SHOULD be used to measure the current FC pseudowire latency between the CE devices. If the measured latency violates any of the timing requirements, then the FC PW PE MUST generate a WAN Down event as specified in [FC-BB-6].

The WAN Down event causes the PE to continuously send NOS (an FC primitive sequence) on the native FC link to the FC Port at the other

end of that link (typically an E_Port on a switch in this case). This immediately causes the FC link that is carried by the PW to become non-operational, halting transmission of FC traffic. However, it is not necessary to tear down the pseudowire itself in this situation (e.g., destroy the MPLS path set up by LDP).

The Transparent FC-BB initialization state machine in [FC-BB-6] specifies the protocol used to attempt to recover from a WAN Down event (i.e., bring the WAN back up). If that protocol brings the WAN back up, FC traffic will resume and the standard FC link recovery protocol will bring the encapsulated FC link back up. If the previous pseudowire was destroyed, attempts will be made to re-establish the path via LDP as part of recovering from the WAN Down event. If the PW round-trip latency remains above R_T_TOV, the initialization protocol for the FC PW will repeatedly time out in attempting to recover from the WAN Down event, preventing recovery of the FC link carried by the PW, see [FC-BB-6].

6. Security Considerations

The FC PW is an MPLS pseudowire; for MPLS pseudowire security considerations, see the security considerations sections of [RFC3985] and [RFC4385].

The protocols used to implement security in a Fibre Channel fabric are defined in [FC-SP]. These protocols operate at higher layers of the FC hierarchy and are transparent to the FC PW.

The FC timing requirements (see Section 5) create an exposure of the FC PW to inserted latency. Injection of latency sufficient to cause the round trip time for an FC PW to exceed R_T_TOV (default: 100ms) may cause the FC PW to fail in an active fashion because the FC link initialization protocol repeatedly times out. OAM functionality for deployed FC PWs SHOULD monitor for persistence of this situation and respond accordingly (e.g., shut down the FC PW in order to avoid wasting WAN bandwidth on an FC PW whose FC link cannot be successfully initialized due to excessive latency).

7. Applicability Statement

FC PW allows the transparent transport of FC traffic between Fibre Channel ports while saving network bandwidth by removing FC Idle Signals and reducing the number of FC Primitive Sequences.

- The pair of CE devices operates as if they were directly connected by an FC link. In particular they react to Primitive Sequences on their local FC links as specified by the FC standards.

- The FC PW carries only FC data frames, FC Primitive Signals and a subset of the copies of an FC Primitive Sequence. Idle Primitive Signals are suppressed, and long streams of the same Primitive Sequence are reduced over the PW thus saving bandwidth.

- The PW PE MUST generate Idle Primitive Signals to the attached FC link when there is no other traffic to transmit on the attached FC link [FC-FS-2].

- The PW PE MUST send Primitive Sequences continuously to the attached FC port, as required by the FC standards [FC-FS-2].

FC PW traffic should only traverse MPLS networks that are provisioned based on traffic engineering to provide dedicated bandwidth for FC PW traffic. The MPLS network should enforce ingress traffic policing so that delivery of FC PW traffic can be assured. To extend FC across a network that does not satisfy these requirements, FCIP SHOULD be used instead of an FC PW, see [RFC3821] and [FC-BB-6].

This document does not provide any mechanisms for protecting an FC PW against network outages. As a consequence, resilience of the emulated FC service to such outages is dependent upon the underlying MPLS network, which should be protected against failures. When a network outage is detected, the PE SHOULD use a WAN Down event (as specified in [FC-BB-6]) to convey the PW status to the CE, to enable faster outage handling.

8. IANA Considerations

IANA is requested to assign a new MPLS Pseudowire (PW) type as follows:

PW type	Description	Reference
0x001F	FC Port Mode	RFC XXXX

The above value is suggested as the next available value and has been reserved for this purpose by IANA.

RFC Editor: Please replace RFC XXXX above with the RFC number of this document and remove this note.

IANA should reserve the following Pseudowire Interface Parameters Sub-TLV Types that were tentatively allocated for FC PW and restrict them to prevent future allocation, citing this RFC as the reference for that reservation and restriction. These Sub-TLV types were used for the FC PW Selective Retransmission protocol, which the working

group has decided to eliminate. This action prevents future use of these values for other purposes, as there is at least one implementation of the Selective Retransmission protocol that has been deployed.

Parameter	ID Length	Reference
0x12	4	RFC XXXX
0x13	4	RFC XXXX
0x14	4	RFC XXXX
0x15	4	RFC XXXX

RFC Editor: Please replace RFC XXXX above with the RFC number of this document and remove this note.

9. Acknowledgments

Previous versions of this document were authored by Moran Roth, Ronen Solomon and Munefumi Tsurusawa; their efforts and contributions are gratefully acknowledged. The authors would like to thank Stewart Bryant, Elwyn Davies, Steve Hanna, Dave Peterson, Yaakov Stein, Alexander Vainshtein, and the members of the IESG for helpful comments on this document.

The protocol specified in this document is intended to be used in conjunction with the Fibre Channel pseudowire portion of the FC-BB-6 specification developed by INCITS Technical Committee T11. The authors would like to thank the members of both the IETF and T11 organizations who have supported and contributed to this work.

This document was prepared using 2-Word-v2.0.template.dot.

10. Normative References

[RFC3643] Weber, R., et al, "Fibre Channel (FC) Frame Encapsulation", RFC 3643, December 2003.

[RFC3985] Bryant, S., et al, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", RFC 4446, April 2006.

[RFC4447] Martini, L., et al, "Pseudowire Setup and Maintenance using the Label Distribution Protocol (LDP)", RFC4447, April 2006.

[RFC4385] Bryant, S., et al, "Pseudowire Emulation Edge-to-Edge(PWE3) Control Word for use over an MPLS PSN", RFC4385, February 2006.

[RFC4623] Malis, A., Townsley, M., "PWE3 Fragmentation and Reassembly", RFC 4623, August 2006.

[RFC4720] Malis, A., et al, "Pseudowire Emulation Edge-to-Edge (PWE3) Frame Check Sequence Retention", RFC 4720, November 2006.

[FC-BB-6] "Fibre Channel Backbone-6" (FC-BB-6), T11 Project 2159-D, Rev 1.02, October 2010.

RFC Editor: FC-BB-6 is a work in progress. Please treat [FC-BB-6] as a normative reference to a work in progress, and proceed as follows:

1. Assign an RFC number to this draft and communicate that number to the authors of this draft, one of whom (David Black) is the T11 designated liaison to IETF.
2. Place a reference hold on this draft until FC-BB-6 is published as an ANSI standard.
3. When FC-BB-6 is published as an ANSI standard, the draft authors will provide an update to the FC-BB-6 reference that includes an ANSI standard number. Update the FC-BB-6 reference using that information, remove the reference hold due to FC-BB-6, and remove this note.

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

[FC-FS-2] "Fibre Channel - Framing and Signaling-2 (FC-FS-2)", ANSI INCITS 424:2007, August 2007.

11. Informative references

[RFC3821] M. Rajogopal, E. Rodriguez, "Fibre Channel over TCP/IP (FCIP)", RFC 3821, July 2004.

[RFC4717] Martini, L., et al, "Encapsulation Methods for Transport of Asynchronous Transfer Mode (ATM) over MPLS Networks", RFC 4717, December 2006.

[T11] INCITS Technical Committee T11, http://www.t11.org, visited January, 2011.

[FC-SP] "Fibre Channel - Security Protocols" (FC-SP), ANSI INCITS 426:2007, February 2007.

Authors' Addresses

David L. Black (ed.)
EMC Corporation
176 South Street
Hopkinton, MA 01748
Phone: +1 (508) 293-7953
Email: david.black@emc.com

Linda Dunbar (ed.)
Huawei Technologies
1700 Alma Drive, Suite 500
Plano, TX 75075, USA
Phone: +1 (972) 543-5849
Email: ldunbar@huawei.com

Moran Roth
Infinera Corporation
169 Java Drive
Sunnyvale, CA 94089
Phone: (408) 572-5200
Email: MRoth@infinera.com

Ronen Solomon
Orckit-Corrigent Systems
126, Yigal Alon st.
Tel Aviv, ISRAEL
Phone: +972-3-6945316
Email: ronens@corrigent.com

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

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

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: September 11, 2012

Siva Sivabalan (Ed.)
Sami Boutros (Ed.)
Luca Martini
Cisco Systems

Frederic Jounay
Philippe Niger
France Telecom

Maciek Konstantynowicz
Juniper

Gianni Del Vecchio
Swisscom

Thomas D. Nadeau
CA Technologies

Yuji Kamite
NTT Communications

Simon Delord
Telstra

Lizhong Jin
ZTE

Laurent Ciavaglia
Martin Vigoureux
Alcatel-Lucent

March 11, 2012

Signaling Root-Initiated Point-to-Multipoint Pseudowire using LDP
draft-ietf-pwe3-p2mp-pw-04.txt

Status of this Memo

Abstract

This document specifies a mechanism to signal Point-to-Multipoint (P2MP) Pseudowires (PW) tree using LDP. Such a mechanism is suitable for any Layer 2 VPN service requiring P2MP connectivity over an IP or MPLS enabled PSN. A P2MP PW established via the proposed mechanism is root initiated.

Table of Contents

1. Introduction

A Point-to-Multipoint (P2MP) Pseudowire (PW) emulates the essential attributes of a unidirectional P2MP Telecommunications service such as P2MP ATM over PSN. A major difference between a Point-to-Point (P2P) PW outlined in [RFC3985] and a P2MP PW is that the former is intended for bidirectional service whereas the latter

is intended for both unidirectional, and optionally bidirectional service. Requirements for P2MP PW are described in [P2MP-PW-REQ].

P2MP PW can be constructed as either Single Segment (P2MP SS-PW) or Multi Segment (P2MP MS-PW) Pseudowires as mentioned in [P2MP-PW-REQ]. P2MP MS-PW is outside the scope of this document. A reference model for P2MP PW is depicted in Figure 1 below. A transport LSP associated with a P2MP SS-PW SHOULD be a P2MP MPLS LSP (i.e., P2MP TE tunnel established via RSVP-TE [RFC4875] or P2MP LSP established via mLDP [RFC6388]) spanning from the Root-PE to the Leaf-PE(s) of the P2MP SS-PW tree. For example, in Figure 1, PW1 can be associated with a P2MP TE tunnel or P2MP LSP setup using mLDP originating from PE1 and terminating at PE2 and PE3.

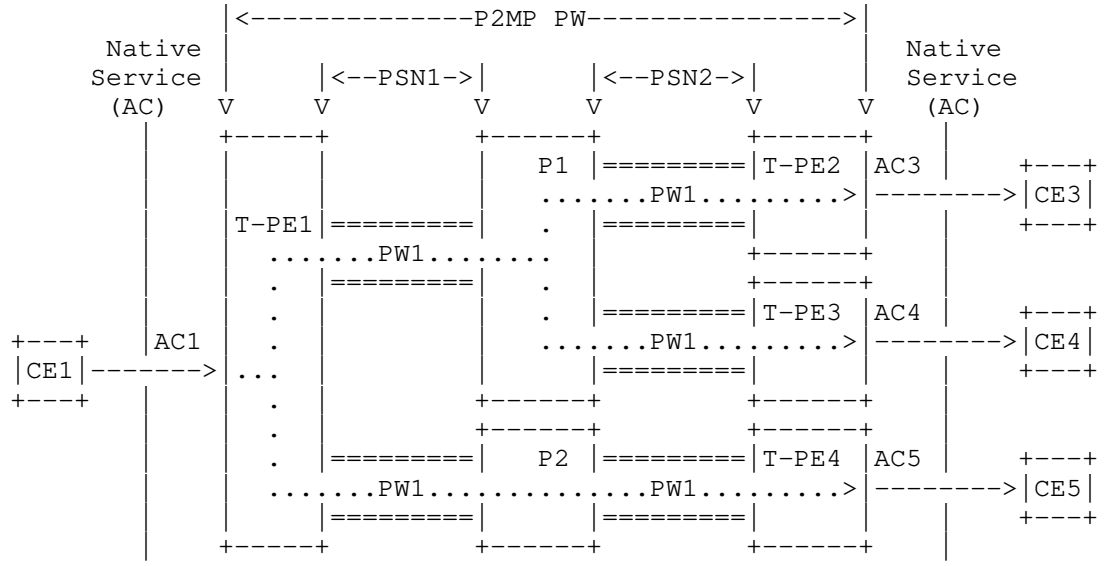


Figure 1: P2MP PW

Mechanisms for establishing P2P SS-PW using LDP are described in

[RFC4447]. In this document, we specify a method to signal P2MP PW using LDP. In particular, we define new FEC, TLVs, parameters, and status codes to facilitate LDP to signal and maintain P2MP PWs.

As outlined in [P2MP-PW-REQ], even though the traffic flow from a Root-PE (R-PE) to Leaf-PE(s) (L-PEs) is P2MP in nature, it may be desirable for any L-PE to send unidirectional P2P traffic destined only to the R-PE. The proposed mechanism takes such option into consideration.

The P2MP PW requires an MPLS LSP to carry the PW traffic, and the MPLS packets carried over the PW will be encapsulated according to the methods described in [RFC5332].

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 Error! Reference source not found..

2. Terminology

FEC: Forwarding Equivalence Class

LDP: Label Distribution Protocol

mLDP: Label Distribution Protocol for P2MP/MP2MP LSP

LSP: Label Switching Path

MS-PW: Multi-Segment Pseudowire

P2P: Point to Point

P2MP: Point to Multipoint

PE: Provider Edge

PSN: Packet Switched Network

PW: Pseudowire

SS-PW: Single-Segment Pseudowire

S-PE: Switching Provider Edge Node of MS-PW

TE: Traffic Engineering

R-PE: Root-PE - ingress PE, PE initiating P2MP PW setup.

L-PE: Leaf-PE - egress PE.

3. Signaling P2MP PW

In order to advertise labels as well as exchange PW related LDP messages, PEs must establish LDP sessions among themselves using the Extended Discovery Mechanisms. A PE discovers other PEs that are to be connected via P2MP PWs either via manual configuration or autodiscovery [RFC6074].

R-PE and each L-PE MUST be configured with the same FEC as defined in the following section.

P2MP PW requires that there is an active P2MP PSN LSP set up between R-PE and L-PE(s). Note that the procedure to set up the P2MP PSN LSP is different depending on the signaling protocol used (RSVP-TE or mLDP).

In case of mLDP, a Leaf-PE can decide to join the P2MP LSP at any time; whereas in the case of RSVP-TE, the P2MP LSP is set up by the R-PE, generally at the initial service provisioning time. It should be noted that local policy can override any decision to join, add or prune existing or new L-PE(s) from the tree. In any case, the PW setup can ignore these differences, and simply assume that the P2MP PSN LSP is available when needed.

A P2MP PW signaling is initiated by the R-PE simply by sending a P2MP-PW LDP label mapping message to the L-PE(s) belonging to that P2MP PW. This label mapping message will contain the following:

1. A FEC TLV containing P2MP PW Upstream FEC element that includes Transport LSP sub TLV.
2. An Interface Parameters TLV, as described in [RFC4447].
3. A PW Grouping TLV, as described in [RFC4447].
4. A label TLV for the upstream-assigned label used by R-PE for the traffic going from R-PE to L-PE(s).

The R-PE imposes the upstream-assigned label on the outbound packets sent over the P2MP-PW, and using this label an L-PE identifies the inbound packets arriving over the P2MP PW.

Additionally, the R-PE MAY send label mapping message(s) to one or more L-PE(s) to signal unidirectional P2P PW(s). The L-PE(s) can use such PW(s) to send traffic to the R-PE. This optional label mapping message will contain the following:

1. P2P PW Downstream FEC element.
2. A label TLV for the down-stream assigned label used by the corresponding L-PE to send traffic to the R-PE.

The LDP liberal label retention mode is used, and per requirements specified in [RFC5036], the Label Request message MUST also be supported.

The upstream-assigned label is allocated according to the rules in [RFC5331].

When an L-PE receives a PW Label Mapping Message, it MUST verify the associated P2MP PSN LSP is in place. If the associated P2MP PSN LSP is not in place, and its type is LDP P2MP LSP, the L-PE SHOULD attempt to join the P2MP LSP associated with the P2MP PW. If the associated P2MP PSN LSP is not in place, and its type is RSVP-TE P2MP LSP, the L-PE SHOULD wait till the P2MP transport LSP is signaled.

3.1. PW ingress to egress incompatibility issues

If an R-PE signals a PW with a pw type, CW mode, or interface parameters that a particular L-PE cannot accept, then the L-PE must

not enable the PW, and notify the user. In this case, a PW status message with status code of 0x00000001 (Pseudowire Not Forwarding) MUST also be sent to the R-PE.

Note that this procedure does not apply if the L-PE had not been provisioned with this particular P2MP PW. In this case according to the LDP liberal label retention rules, no action is taken.

3.2. P2MP PW FEC

[RFC4447] specifies two types of LDP FEC elements called "PWid FEC Element" and "Generalized PWid FEC Element" used to signal P2P PWs. We define two new types of FEC elements called "P2MP PW Upstream FEC Element" and "P2P PW Downstream FEC Element". These FEC elements are associated with a mandatory upstream assigned label and an optional downstream assigned label respectively.

FEC type of the P2MP PW Upstream FEC Element is 0x82 (pending IANA allocation) and is encoded as follows:

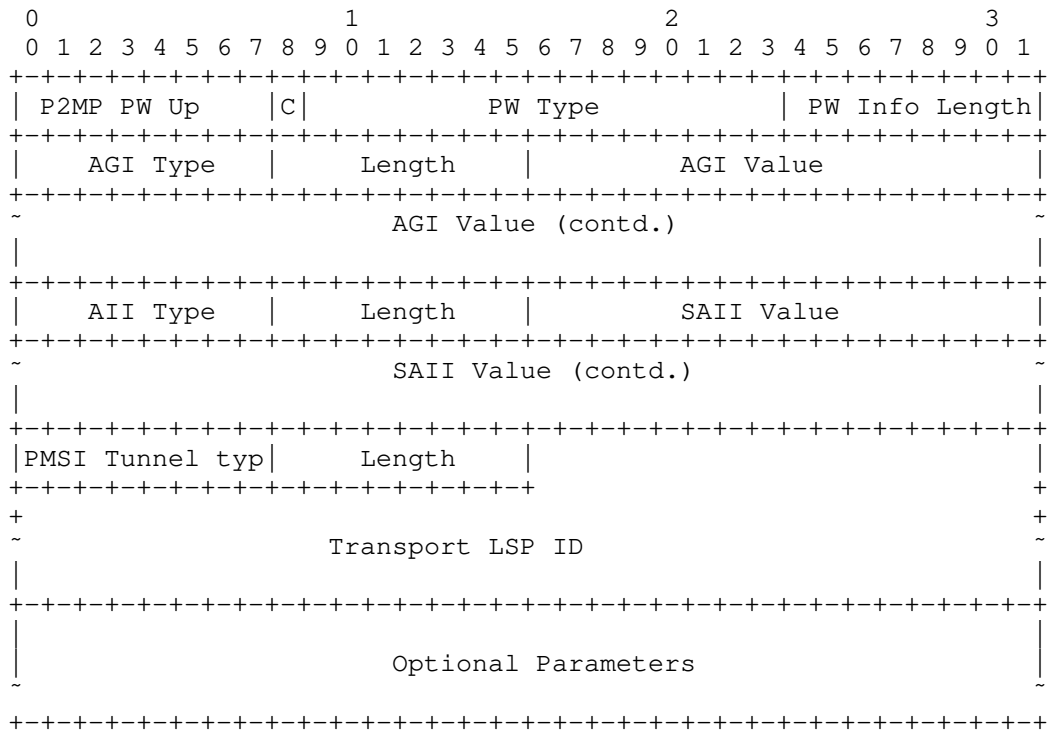


Figure 2: P2MP PW Upstream FEC Element

* PW Type:

  15-bit representation of PW type, and the assigned values are assigned by IANA.

* C bit:

  A value of 1 or 0 indicates whether control word is present or absent for the P2MP PW.

* PW Info Length:

Sum of the lengths of AGI, SAII, PMSI Tunnel info, and Optional Parameters field in octets. If this value is 0, then it references all PWs using the specified grouping ID. In this case, there are neither other FEC element fields (AGI, SAII, etc.) present, nor any interface parameters TLVs. Alternatively, we can use typed WC FEC described in section 3.3 to achieve the same or to have better filtering.

* AGI:

Attachment Group Identifier can be used to uniquely identify VPN or VPLS instance associated with the P2MP PW. This has the same format as the Generalized PWid FEC element [RFC4447].

* SAII:

Source Attachment Individual Identifier is used to identify the root of the P2MP PW. The root is represented using AII type 2 format specified in [RFC5003]. Note that the SAII can be omitted by simply setting the length and type to zero.

P2MP PW is identified by the Source Attachment Identifier (SAI). If the AGI is non-null, the SAI is the combination of the SAII and the AGI, if the AGI is null, the SAI is the SAII.

* PMSI Tunnel Type and Transport LSP ID:

A P2MP PW MUST be associated with a transport LSP which can be established using RSVP-TE or mLDP.

* PMSI Tunnel Type:

The PMSI tunnel type is defined in [L3VPN-MCAST].

When the type is set to mLDP P2MP LSP, the Tunnel Identifier is a P2MP FEC Element as defined in [RFC6388]. A new mLDP Opaque Value Element type for L2VPN-MCAST application needs to be allocated.

* Transport LSP ID:
  This is the Tunnel Identifier which is defined in [L3VPN-MCAST].

An R-PE sends Label Mapping Message as soon as the transport LSP ID associated with the P2MP PW is known (e.g., via configuration) regardless of the operational state of that transport LSP. Similarly, an R-PE does not withdraw the labels when the corresponding transport LSP goes down. Furthermore, an L-PE retains the P2MP PW labels regardless of the operational status of the transport LSP.

Note that a given transport LSP can be associated with more than one P2MP PWs and all P2MP PWs will be sharing the same R-PE and L-PE(s).

In the case of LDP P2MP LSP, when an L-PE receives the Label Mapping Message, it can initiate the process of joining the P2MP LSP tree associated with the P2MP PW.

In the case of RSVP-TE P2MP LSP, only the R-PE initiates the signaling of P2MP LSP.

* Optional Parameters:

The Optional Parameter field can contain some TLVs that are not part of the FEC, but are necessary for the operation of the PW. This proposed mechanism uses two such TLVs: Interface Parameters TLV, and Group ID TLV.

The Interface Parameters TLV and Group ID TLV specified in [RFC4447] can also be used in conjunction with P2MP PW FEC in a label message. For Group ID TLV, the sender and receiver of these TLVs should follow the same rules and procedures specified in [RFC4447]. For Interface Parameters TLV, the procedure differs from the one specified in [RFC4447] due to specifics of P2MP connectivity. When the interface parameters are signaled by a R-PE, each L-PE must check if its configured value(s) is less than or equal to the threshold value provided by the R-PE (e.g. MTU size (Ethernet), max number of concatenated ATM cells, etc)). For

other interface parameters like CEP/TDM Payload bytes (TDM), the value MUST exactly match the values signaled by the R-PE.

Multicast traffic stream associated with a P2MP PW can be selective or inclusive. To support the former, this document defines a new optional Selective Tree Interface Parameter sub-TLV (type is pending IANA allocation) according to the format described in [RFC4447]. The value of the sub-TLV contains the source and the group for a given multicast tree as shown in Figure 3. Also, if a P2MP PW is associated with multiple selective trees, the corresponding label mapping message will carry more than one instance of this Sub-TLV. Furthermore, in the absence of this sub-TLV, the P2MP PW is associated with all multicast traffic stream originating from the root.

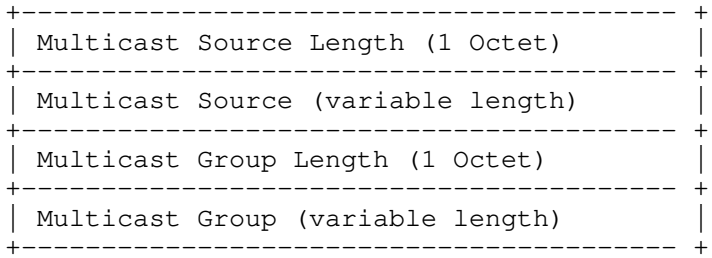


Figure 3: Selective Tree Interface Parameter Sub-TLV Value

Note that since the LDP label mapping message is only sent by the R-PE to all the L-PEs, it is not possible to negotiate any interface parameters.

The type of optional P2P PW Downstream FEC Element is 0x83 (pending IANA allocation), and is encoded as follows:

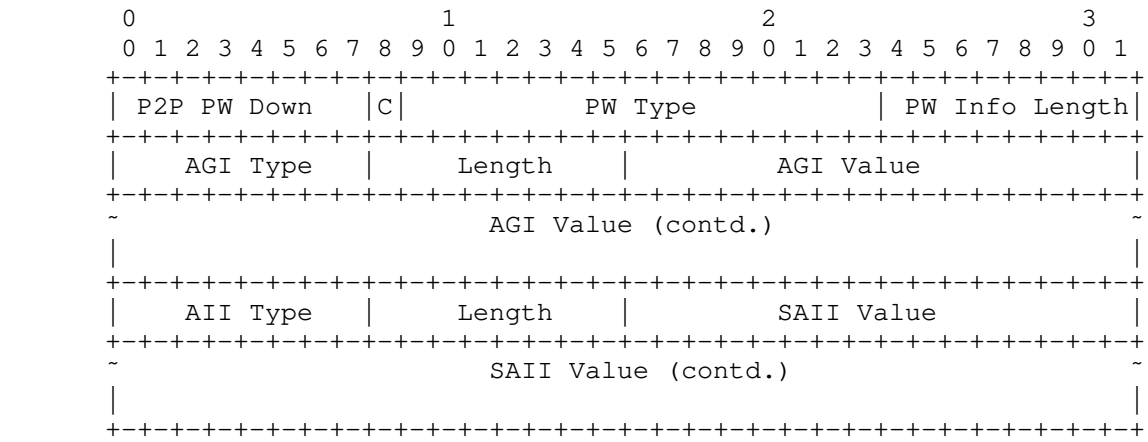


Figure 4: P2P PW Downstream FEC Element

The definition of the fields in the P2P PW Downstream FEC Element is the same as those of P2MP PW Upstream FEC Element.

3.3. Typed Wildcard FEC Format for new FEC

[RFC5918] defines the general notion of a "Typed Wildcard" FEC Element, and requires FEC designer to specify a typed wildcard FEC element for newly defined FEC element types. This document defines two new FEC elements, "P2MP PW Upstream" and "P2P PW Downstream" FEC element, and hence requires us to define their Typed Wildcard format.

[PW-TWC-FEC] defines Typed Wildcard FEC element format for other PW FEC Element types (PWid and Gen. PWid FEC Element) in section 2 as follows:

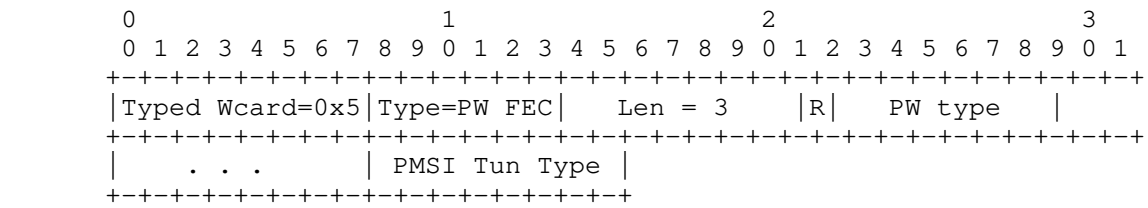


Figure 5: Typed Wildcard Format for P2MP PW FEC Elements

[PW-TWC-FEC] specifies that "Type" field can be either "PWid" (0x80) or "Generalized PWid" (0x81) FEC element type. This document reuses the existing typed wildcard format as specified in [PW-TWC-FEC] and illustrated in Figure 5. We extend the definition of "Type" field to also include "P2MP PW Upstream" and "P2P PW Downstream" FEC element types, as well as add an additional field "PMSI Tun Type". We reserve PMSI tunnel Type 0xFF to mean "wildcard" transport tunnel type. This "wildcard" transport tunnel type can be used in a typed wildcard p2mp FEC for further filtering. This field only applies to Typed wildcard P2MP PW Upstream FEC and MUST be set to "wildcard" for "P2P PW Downstream FEC" typed wildcard element.

3.4. Group ID usage

The Grouping TLV as defined in [RFC4447] contains a group ID capable of indicating an arbitrary group membership of a P2MP-PW. This group ID can be used in LDP "wild card" status, and withdraw label messages, as described in [RFC4447].

3.5. Generic Label TLV

As in the case of P2P PW signaling, P2MP PW labels are carried within Generic Label TLV contained in LDP Label Mapping Message. A Generic Label TLV is formatted and processed as per the rules and procedures specified in [RFC4447]. For a given P2MP PW, a single upstream-assigned label is allocated by the R-PE, and is advertised to all L-PEs using the Generic Label TLV in label mapping message containing the P2MP PW Upstream FEC element.

The R-PE can also allocate a unique label for each L-PE from which it intends to receive P2P traffic. Such a label is advertised to the L-PE using Generic Label TLV and P2P PW Downstream FEC in label mapping message.

4. LDP Capability Negotiation

The capability of supporting P2MP PW must be advertised to all LDP peers. This is achieved by using the methods in [RFC5561] and

advertising the LDP "P2MP PW Capability" TLV. If an LDP peer supports the dynamic capability advertisement, this can be done by sending a new Capability message with the S bit set for the P2MP PW capability TLV. If the peer does not supports dynamic capability advertisement, then the P2MP PW Capability TLV MUST be included in the LDP Initialization message during the session establishment. An LSR having P2MP PW capability MUST recognize both P2MP PW Upstream FEC Element and P2P PW Downstream FEC Element in LDP label messages.

In line with requirements listed in [RFC5561], the following TLV is defined to indicate the P2MP PW capability:

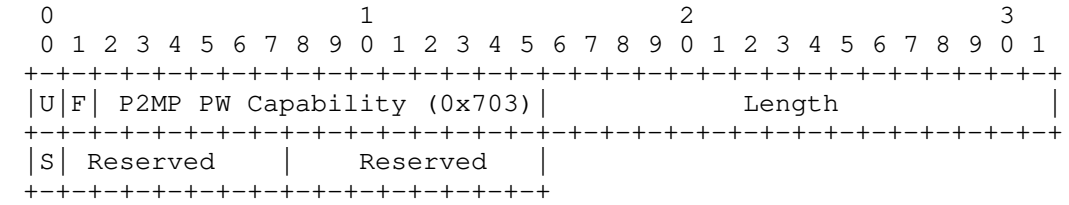


Figure 7: LDP P2MP PW Capability TLV

Note: TLV number pending IANA allocation.

* U-bit:

  SHOULD be 1 (ignore if not understood).

* F-bit:

  SHOULD be 0 (don't forward if not understood).

* P2MP PW Capability TLV Code Point:

  The TLV type, which identifies a specific capability. The P2MP PW capability code point is requested in the IANA allocation section below.

* S-bit:

The State Bit indicates whether the sender is advertising or withdrawing the P2MP PW capability. The State bit is used as follows:

1 - The TLV is advertising the capability specified by the TLV Code Point.

0 - The TLV is withdrawing the capability specified by the TLV Code Point.

* Length:
MUST be set to 2 (octet).

5. P2MP PW Status

In order to support the proposed mechanism, a node MUST be capable of handling PW status. As such, PW status negotiation procedure described in [RFC4447] is not applicable to P2MP PW.

Once an L-PE successfully processes a Label Mapping Message for a P2MP PW, it MUST send appropriate PW status according to the procedure specified [RFC4447] to notify the PW status. If there is no PW status notification required, then no PW status notification is sent (for example if the P2MP PW is established and operational with a status code of Success (0x00000000), pw status message is not necessary). PW status message sent from any L-PE to R-PE contains P2P PW Downstream FEC to identify the PW.

An R-PE also sends PW status to L-PE(s) to reflect its view of a P2MP PW state. Such PW status message contains P2MP PW Upstream FEC to identify the PW.

Connectivity status of the underlying P2MP LSP that P2MP PW is associated with, can be verified using LSP Ping and Traceroute procedures described in [P2MP-LSP-PING].

6. Security Considerations

The security measures described in [RFC4447] is adequate for the proposed mechanism.

7. IANA Considerations

7.1. FEC Type Name Space

This document uses two new FEC element types, number 0x82 and 0x83 will be requested as an allocation from the registry "FEC Type Name Space" for the Label Distribution Protocol (LDP RFC5036):

Value	Hex	Name	Reference
130	0x82	P2MP PW Upstream FEC Element	RFCxxxx
131	0x83	P2P PW Downstream FEC Element	RFCxxxx

7.2. LDP TLV Type

This document uses a new LDP TLV types, IANA already maintains a registry of name "TLV TYPE NAME SPACE" defined by RFC5036. The following values are suggested for assignment:

TLV type	Description:
0x0703	P2MP PW Capability TLV

7.3. mLDP Opaque Value Element TLV Type

This document requires allocation of a new mLDP Opaque Value Element Type from the LDP MP Opaque Value Element type name space defined in [mLDP].

The following value is suggested for assignment:

TLV type	Description
0x3	L2VPN-MCAST application TLV

7.4. Selective Tree Interface Parameter sub-TLV Type

This document requires allocation of a sub-TLV from the registry "Pseudowire Interface Parameters Sub-TLV Type".

The following value is suggested for assignment:

TLV type	Description
0x0D	Selective Tree Interface Parameter.

7.5. WildCard PMSI tunnel type.

This document requires the allocation of PMSI tunnel Type 0xFF to mean wildcard transport tunnel type

8. Acknowledgment

Authors would like thank Andre Pelletier and Parag Jain for their valuable suggestions.

9. References

9.1. Normative References

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

[RFC4447] "Transport of Layer 2 Frames Over MPLS", Martini, L., et al., RFC 4447, April 2006.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5003] C. Metz, L. Martini, F. Balus, J. Sugimoto, "Attachment Individual Identifier (AII) Types for Aggregation", RFC5003, September 2007.

[RFC5331] R. Aggarwal, Y. Rekhter, E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.

[RFC5332] T. Eckert, E. Rosen, Ed.,R. Aggarwal, Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.

[RFC6388] I. Minei, K. Kompella, I. Wijnands, B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and

Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

[RFC4875] R. Aggarwal, Ed., D. Papadimitriou, Ed., S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs).", RFC 4875, May 2007.

[L3VPN-MCAST] R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-bgp-08.txt, Work in Progress, October 2009.

[RFC5561] B.Thomas, K.Raza, S.Aggarwal, R.Agarwal, JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.

[RFC5918] R. Asati, I. Minei, and B. Thomas, "LDP Typed Wildcard Forwarding Equivalence Class", RFC 5918, August 2010.

[PW-TWC-FEC] K. Raza, S. Boutros, and C. Pignataro, "LDP Typed Wildcard FEC for PWid and Generalized PWid FEC Elements", draft-ietf-pwe3-pw-types-wc-fec-03.txt, work in progress, February 2012.

9.2. Informative References

[RFC3985] Stewart Bryant, et al., "PWE3 Architecture", RFC3985

[RFC6074] E. Rosen,W. Luo,B. Davie,V. Radoaca "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC6074, January 2011.

[P2MP-PW-REQ] F. Jounay, et. al, "Requirements for Point to Multipoint Pseudowire", draft-ietf-pwe3-p2mp-pw-requirements-03.txt, Work in Progress, August 2010.

[P2MP-LSP-PING] A. Farrel, S. Yasukawa, "Detecting Data Plane Failures in Point-to-Multipoint Multiprotocol Label Switching (MPLS) - Extensions to LSP Ping", draft-ietf-mpls-p2mp-lsp-ping-15.txt, Work In Progress, January 2011.

Author's Addresses

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario, K2K 3E8
Canada
Email: msiva@cisco.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way
San Jose, California 95134
USA
Email: sboutros@cisco.com

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO, 80112
United States
Email: lmartini@cisco.com

Maciek Konstantynowicz
Juniper Networks
UNITED KINGDOM
e-mail: maciek@juniper.net

Gianni Del Vecchio
Swisscom (Schweiz) AG
Zentweg 9
CH-3006 Bern
Switzerland
e-mail: Gianni.DelVecchio@swisscom.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive
Portsmouth, NH 03801
USA
e-mail: thomas.nadeau@ca.com

Frederic Jounay
France Telecom

2, avenue Pierre-Marzin
22307 Lannion Cedex
FRANCE
Email: frederic.jounay@orange-ftgroup.com

Philippe Niger
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex
FRANCE
Email: philippe.niger@orange-ftgroup.com

Yuji Kamite
NTT Communications Corporation
Tokyo Opera City Tower
3-20-2 Nishi Shinjuku, Shinjuku-ku
Tokyo 163-1421
Japan
Email: y.kamite@ntt.com

Lizhong Jin
ZTE
889 Bibo Road,
Shanghai, 201203
P.R.China
Email: lizhong.jin@zte.com.cn

Martin Vigoureux
Alcatel-Lucent
Route de Villejust
Nozay, 91620
France
Email: martin.vigoureux@alcatel-lucent.com

Laurent Ciavaglia
Alcatel-Lucent
Route de Villejust
Nozay, 91620
France
Email: Laurent.Ciavaglia@alcatel-lucent.com

Simon Delord
Alcatel-Lucent
E-mail: simon.a.delord@team.telstra.com

Kamran Raza
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario, K2K 3E8
Canada
Email: skraza@cisco.com

Acknowledgment

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

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: October 30, 2012

Parag Jain, Ed.
Sami Boutros
Cisco Systems, Inc.

Sam Aldrin
Huawei Technologies

April 26, 2012

Definition of P2MP PW TLV for LSP-Ping Mechanisms
draft-jain-mpls-p2mp-pw-lsp-ping-02.txt

Abstract

LSP-Ping is a widely deployed Operation, Administration, and Maintenance (OAM) mechanism in MPLS networks. This document describes a mechanism to verify connectivity of Point-to-Multipoint (P2MP) Pseudowires (PW) using LSP Ping.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on December 28, 2011.

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction

A Point-to-Multipoint (P2MP) Pseudowire (PW) emulates the essential attributes of a unidirectional P2MP Telecommunications service such as P2MP ATM over PSN. Requirements for P2MP PW are described in [PPWREQ]. P2MP PWs are carried over P2MP MPLS LSP. The Procedure for P2MP PW signaling using LDP for single segment P2MP PWs are described in [PPWPWE3]. Many P2MP PWs can share the same P2MP MPLS LSP and this arrangement is called Aggregate P-tree. The aggregate P2MP trees require an upstream assigned label so that on the tail of the P2MP LSP, the traffic can be associated with a VPN or a VPLS instance. When a P2MP MPLS LSP carries only one VPN or VPLS service instance, the arrangement is called Inclusive P-Tree. For Inclusive P-Trees, P2MP MPLS LSP label itself can uniquely identify the VPN or VPLS service being carried over P2MP MPLS LSP. The P2MP MPLS LSP can also be used in Selective P-Tree arrangement for carrying multicast traffic. In a Selective P-Tree arrangement, traffic to each multicast group in a VPN or VPLS instance is carried by a separate

unique P-tree. In Aggregate Selective P-tree arrangement, traffic to a set of multicast groups from different VPN or VPLS instances is carried over a same shared P-tree.

The P2MP MPLS LSP are setup either using MLDP [MLDP] or P2MP RSVP-TE [RFC4875]. Mechanisms for fault detection and isolation for data plane failures for P2MP MPLS LSPs are specified in [PLSPPING]. This document describes a mechanism to detect data plane failures for P2MP PW carried over P2MP MPLS LSPs.

This document defines a new FEC 130 Pseudowire sub-TLV for Target FEC Stack for P2MP PW. The FEC 130 Pseudowire sub-TLV is added in Target FEC Stack TLV by the originator of the echo request to inform the receiver at P2MP MPLS LSP tail, of the P2MP PW being tested.

Multi-segment Pseudowires support is out of scope of this document at present and may be included in future.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

The term "FEC-Type" is used to refer to a tuple consisting of <FEC Element Type, Address Family>.

3. Terminology

ATM: Asynchronous Transfer Mode

LSR: Label Switching Router

MPLS-OAM: MPLS Operations, Administration and Maintenance

P2MP-PW: Point-to-Multipoint PseudoWire

PW: PseudoWire

TLV: Type Length Value

4. Identifying a P2MP PW

This document introduces a new LSP Ping Target FEC Stack sub-TLV, FEC 130 Pseudowire sub-TLV, to identify the P2MP PW under test at the P2MP LSP Tail/Bud node.

4.1. FEC 130 Pseudowire Sub-TLV

The FEC 130 Pseudowire sub-TLV fields are taken from P2MP PW FEC Element (FEC Type 0x82) defined in [PPWPWE3]. The PW Type is a 15-bit number indicating the encapsulation type. It is carried right justified in the field below PW Type with the high-order bit set to zero. All the other fields are treated as opaque values and copied directly from P2MP PW FEC Element (FEC Type 0x82) format.

The FEC 130 Pseudowire sub-TLV has the format shown in Figure 1. This TLV will be included in the echo request sent over P2MP PW by the originator of request.

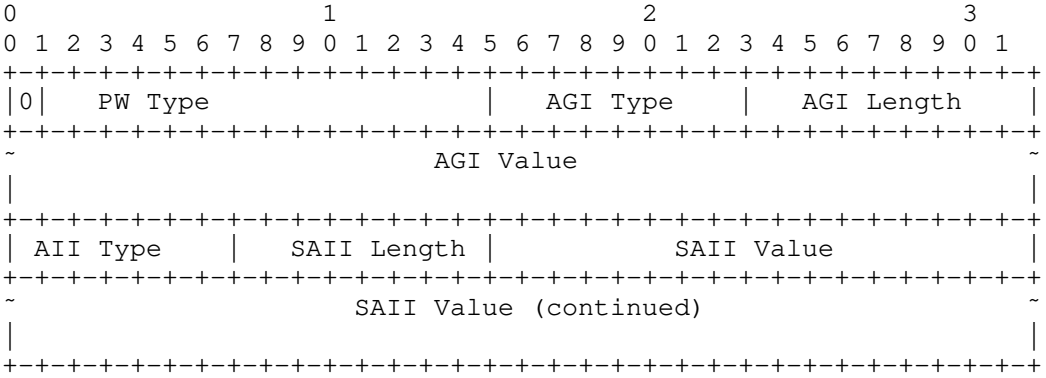


Figure 1: FEC 130 Pseudowire sub-TLV format

For Inclusive and Selective P2MP MPLS P-trees, the echo request will be sent using the P2MP MPLS LSP label.

For Aggregate Inclusive and Aggregate Selective P-trees, the echo request will be sent using a label stack of <P2MP MPLS P-tree label, upstream assigned P2MP PW label>. The P2MP MPLS P-tree label is the outer label and upstream assigned P2MP PW label is inner label.

5. Operations

In this section, we explain the operation of the LSP Ping over P2MP PW. Figure 2 shows a P2MP PW PW1 setup from T-PE1 to remote PEs (T-

PE2, T-PE3 and T-PE4). The transport LSP associated with the P2MP PW1 can be MLDP P2MP MPLS LSP or P2MP TE tunnel.

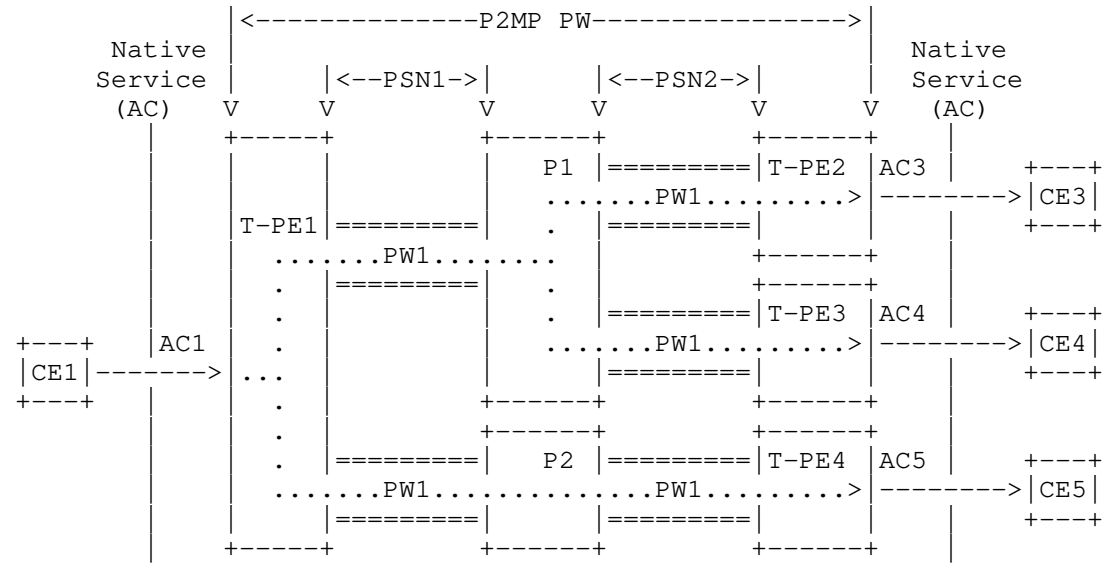


Figure 2: P2MP PW

When an operator wants to perform a connectivity check for the P2MP PW1, the operator initiate a LSP-Ping request with the Target FEC Stack TLV containing FEC 130 Pseudowire sub-TLV in the echo request packet. The echo request packet is sent over the P2MP MPLS LSP using the P2MP MPLS LSP label for Inclusive P-tree or with a label stack with Upstream assigned P2MP PW label as bottom label and P2MP MPLS LSP label as the top label. The intermediate P router will do swap and replication based on the MPLS LSP label. Once the packet reaches remote terminating PEs, the T-PEs will process the packet and perform checks for the FEC 130 Pseudowire sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC4379] and respond according to [RFC4379] processing rules.

6. Echo Reply using Downstream Assigned Label

Root of a P2MP PW may send an optional downstream assigned p2p MPLS label in the LDP Label Mapping message for the P2MP PW signaling. If the root of a P2MP PW expects leaf to send echo reply using the downstream assigned label signaled in the Label Mapping message of the P2MP PW message, the Reply Mode value of 4 "Reply via application level control channel" should be used in Reply Mode field described in Section 3 in [RFC4379] in echo request message for the P2MP PW.

7. Controlling Echo Responses

The procedures described in [PLSPPING] for preventing congestion of Echo Responses (Echo Jitter TLV) and limiting the echo reply to a single egress node (Node Address P2MP Responder Identifier TLV) can be applied to P2MP PW LSP Ping.

8. Security Considerations

The proposal introduced in this document does not introduce any new security considerations beyond that already apply to [PLSPPING].

9. IANA Considerations

This document defines a new sub-TLV type to be included in Target FEC Stack TLV (TLV Type 1) [RFC4379] in LSP Ping.

IANA is requested to assign a sub-TLV type value to the following sub-TLV from the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

FEC 130 Pseudowire sub-TLV (See Section 3). Suggested value 24.

10. References

10.1. Normative References

[RFC4379] K. Kompella, G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

[PPWPWE3] Martini, L. et. al, "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-pwe3-p2mp-pw-03.txt, Work in Progress, March 2011.

[PLSPPING] Saxena, S et. Al, "Detecting Data Plane Failures in Point-to-Multipoint Multiprotocol Label Switching (MPLS) - Extensions to LSP. draft-ietf-mpls-p2mp-lsp-ping-17, Work in Progress, June 2011

10.2. Informative References

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

[RFC5085] T. Nadeau, et. al, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires ", RFC 5085, December 2007.

[MLDP] Minei, I., Kompella, K., Wijnands, I., and Thomas, B., "LDP Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-10.txt, Work in Progress, July 2010.

[RFC4875] Aggarwal, R., Papadimitriou, D., and Yasukawa, S., "Extensions to Resource Reservation Protocol" Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[PPWREQ] F. Jounay, et. al, "Requirements for Point to Multipoint Pseudowire", draft-ietf-pwe3-p2mp-pw-requirements-03.txt, Work in Progress, August 2010.

11. Acknowledgments

The authors would like to thank Shaleen Saxena, Michael Wildt, Tomofumi Hayashi, Danny Prairie for their valuable input and comments.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Parag Jain
Cisco Systems, Inc.,
2000 Innovation Drive,
Kanata, ON K2K3E8, Canada.
E-mail: paragj@cisco.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way,
San Jose, CA 95134, USA.
E-mail: sboutros@cisco.com

Sam Aldrin
Huawei Technologies, co.
2330 Central Express Way,
Santa Clara, CA 95051, USA.
E-mail: aldrin.ietf@gmail.com

Network Working Group
Internet Draft
Intended status: Informational
Expires: April 2012

B. Mack-Crane
L. Yong
Huawei

October 17, 2011

Shortest Path Bridging (SPB) over an MPLS Packet Switched Network
draft-mack-crane-l2vpn-spb-o-mpls-00.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the DNSEXT working group mailing list: <rbridge@postel.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Abstract

This informational document describes ways to interconnect a Shortest Path Tree (SPT) Region over WAN connections using MPLS Pseudo Wires (PWs) with existing SPB and MPLS standards. It also describes how a combination of SPB and MPLS can provide a hierarchical scalable L2VPN.

Table of Contents

1. Introduction

The IEEE Shortest Path Bridging (SPB) standard [802.1aq] provides optimal pair-wise data frame forwarding with little or no configuration in multi-hop networks of arbitrary topology. This network behavior is implemented by Shortest Path Tree (SPT) Bridges that automatically confederate (i.e., recognize compatibly configured neighbors) to form SPT Regions within which shortest path bridging is provided. The data plane controlled by SPT Bridges is unchanged from earlier bridging standards except for the addition of a reverse path forwarding check option. The ECMP project [802.1Qbp] will add support for multipath load spreading for both unicast and multicast traffic. SPB enables a new method to construct enterprise and cloud data center networks.

This document describes use cases for SPB over an MPLS Packet Switched Network (PSN) and introduces a new hierarchical L2VPN architecture that uses SPB and IP/MPLS and documents the related

configurations and references for proper interworking. In the use cases described the SPBM mode (MAC address based) is used, implying the existence of a Provider Backbone Edge Bridge function (MAC-in-MAC encapsulation) [802.1Q] at the boundary of the SPT Region.

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 [RFC2119].

Acronyms used in this document include the following:

AC - Attachment Circuit

CE - Customer Edge

IS-IS - Intermediate System to Intermediate System

MPLS - Multi-Protocol Label Switching

PE - Provider Edge

PPP - Point to Point Protocol

PW - Pseudo Wire

SPB - Shortest Path Bridging

SPT - Shortest Path Tree

VSI - Virtual Switching Instance

2. Use Cases

SPT Regions at different locations may be interconnected by networks that are implemented with different technologies to form one larger SPT Region. This section describes use cases assuming that IP/MPLS technology is available. From the MPLS network view, SPT Bridges act as Customer Edge (CE) devices and connect to PEs via an attachment circuit (AC). SPT Bridges [802.1aq] support deterministic forwarding behavior over point-to-point links. Section 2.1 describes SPT Region interconnection over a single point-to-point link provided by an MPLS network. Section 2.2 describes interconnecting multiple SPT Regions using multiple PWs. Section 2.3 introduces a hierarchical L2VPN solution that uses SPT Bridges and MPLS in a tiered architecture.

2.1. Point-To-Point Interconnection

Two SPT Bridges are interconnected by either an Ethernet or PPP PW over a MPLS network. The PW is configured between a pair of PEs to provide part of the point-to-point link between two SPT Bridges. Figure 1 illustrates this architecture. Each SPT Bridge connects to a PE via an AC and acts as a CE device. The MPLS PSN is bounded by the PEs. The link across the IP/MPLS PSN enables the site A and site B SPT Bridges to form one SPT Region.

MPLS supports many pseudo wire transport encapsulations [RFC4446]. Two types of links between Bridges have been standardized: Ethernet [RFC4448] and PPP [RFC3518, RFC4618]. A Bridge port connected to an AC may be mapped to a PW with Ethernet encapsulation [RFC4448]. The PW between two PEs can be auto-configured [RFC4447] or manually configured; the two Bridges then appear directly interconnected with an Ethernet link.

When the Bridge ports connected to the ACs are configured with PPP, the PEs may be configured as a PW with PPP encapsulation [RFC4618]. After the PW is established between two PEs, the two RBridges then appear directly interconnected with a PPP link. Because the frames between the bridges are encapsulated within PPP, if the PEs have the capability to add or remove PPP encapsulation, it is an independent decision for each AC and for the PW whether each is PPP or Ethernet.

An SPB adjacency is automatically established over an Ethernet link or PPP link. The PW provides transparent transport between ACs.

Note: For Ethernet PW configuration, PE SHOULD use the raw mode and non-service-delimiting options.

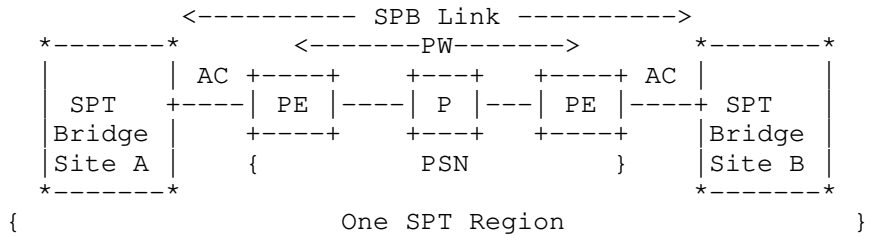


Figure 1 P2P SPB Link over IP/MPLS PSN Use Case I

As networks converge, it is possible that one operator controls both the SPT Region as well as the core MPLS network. Figure 2 illustrates this use case, in which SPT Bridges are also MPLS PE enabled. The interworking between the SPT network and the MPLS PSN is within one device. In this case, a virtual Ethernet interface is configured between the SPT Bridge component and PE component on the SPT/PE device and a Packet-PW is configured between two PE components on two devices to emulate the virtual Ethernet link. An SPB adjacency is established between two RB/PE devices after the PW is established. In this case, SPB runs in the client layer and MPLS runs in the Server Layer; SPB/PE devices support both client and server layer control plane and data plane functions.

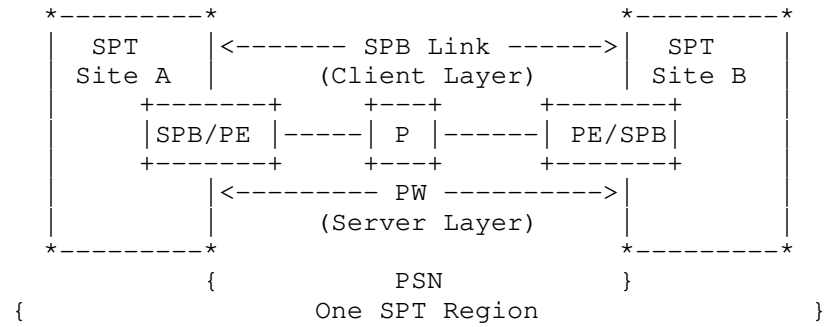


Figure 2 P2P SPB-Link over IP/MPLS PSN Use Case II

In both case I and II, the PE treats an SPT Bridge as a generic CE and has no awareness of SPB capability on the CE. Use case I enables the business models when the SPT Region and Core MPLS may be operated by different operators or the same operator. In the case of different operators, the core MPLS operator can sell a VPWS service to the SPB operator. Use case II provides the model where the SPT Region and the core network are operated by the same operator but use different technologies in edge and core domains of the network.

A PW may cross multiple MPLS domains [RFC5659]. In this case, SPT Bridges connect to T-PEs and it works in the same way as single domain.

2.2. Multiple Interconnections

More than two SPT sites may be interconnected by a full or partial mesh of PWs. The PWs provide a set of links interconnecting the SPT sites and enable the formation of one SPT Region. Interconnecting

multiple sites using PWs is preferable to using a VPLS (VLAN) service because it allows deterministic control of traffic placement and traffic engineering (assuming the PWs provide a bandwidth SLA).

PWs can provide multiple connections to a single physical interface if VLAN tags are used for service selection (Ethernet VLAN ACs). Virtual ports can be provisioned on the SPT Bridge by using a port-mapping S-VLAN component [802.1Qbc]. The S-VID is then used for service selection to map traffic to each PW connection. Figure 3 shows the use of PWs to interconnect three SPT Bridges. One SPT Region is formed across three different sites. Three PWs are configured, providing a full mesh between the three sites. Each SPT site connects to the others via PWs selected by the service-delimiting S-VID on the AC. So in this use case the PEs should use raw mode with service-delimiting.

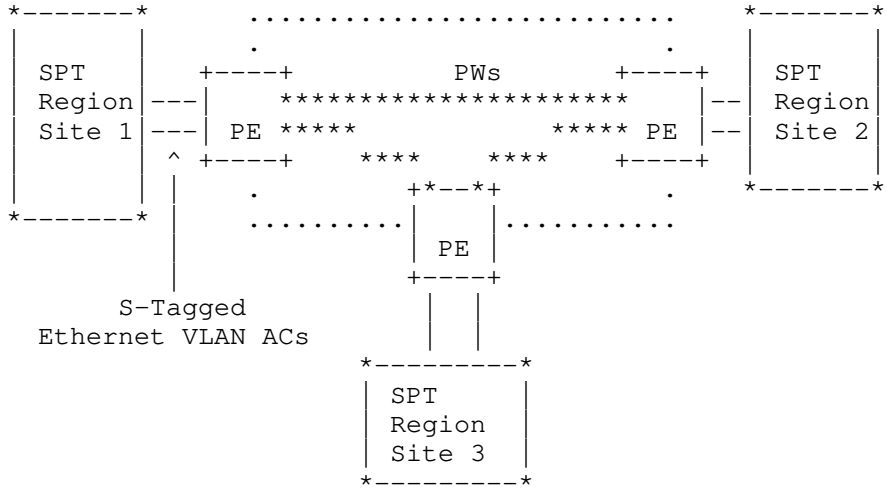


Figure 3 Multiple SPT sites interconnected by PWs

The scenario in Figure 3 can also be applied to interconnect multiple SPT Bridges when a device serves both SPT Bridge and PE functions. This use case is addressed in the following section.

Note: If CEs at a site happen to be regular C-VLAN bridges, the site may be connected to a SPT Bridge via a virtual port bound to an I-Component. This enables MAC-in-MAC encapsulation to be performed

before the traffic enters the SPT Region without requiring upgrade at the C-VLAN bridging site. In this case the PW at the PE connected to the C-VLAN bridging site could be configured as raw mode, non service-delimiting.

2.3. Hierarchical L2VPN with SPB and MPLS

H-VPLS in [RFC4762] describes a two-tier hierarchical solution for the purpose of pseudo wire (PW) scalability improvement. This improvement is achieved by reducing the number of PE devices connected in a full-mesh topology through connecting CE devices via the lower-tier access network, which in turn is connected to the top-tier core network. However, H-VPLS solutions in [RFC4762] require learning and forwarding based on customer MAC addresses, which poses scalability issues as the number of VPLS instances and customer MAC addresses increase. [PBB-VPLS] describes how to use PBB (Provider Backbone Bridges) at the lower-tier access network to solve the scalability issue, in which the transit network nodes only learn and forward on PBB port MAC addresses instead of customer MAC addresses.

Figure 4 depicts the hierarchical L2VPN architecture with SPT Bridge/MPLS technologies. An IP/MPLS network serves the top-tier core network function while an SPT Region serves as the low-tier access network function. A SPB/PE enabled device is placed at the border of the two-tier networks. Ethernet PWs, as described in Section 2.1, are configured between pairs of PE components in the top-tier IP/MPLS network, which construct a full mesh of links among the SPB/PE devices. The SPT Bridge component on a SPB/PE device and other SPT Bridges at the same site serve as the low-tier access network. Customer CEs connect to SPT Bridges at each site directly.

This architecture provides E-LAN or E-VLAN connectivity among customer CEs connecting to the SPT Region sites. The transit SPT Bridge node only forwards and learns other SPT Bridge addresses and the number of PWs in the top-tier core network is not related to the number of devices connecting to Customer CEs. This makes the solution scale very well. In addition, SPB technology supports multiple links from one SPT Bridge to multiple other SPT Bridges and prevents loops, which provides the flexibility to construct the networks based on traffic demands and dynamically reroute traffic when necessary. Figure 4 shows that one SPT Bridge in campus site 1 connects to two SPB/PE devices and one SPB/PE device connects two SPT Bridges at Site 3.

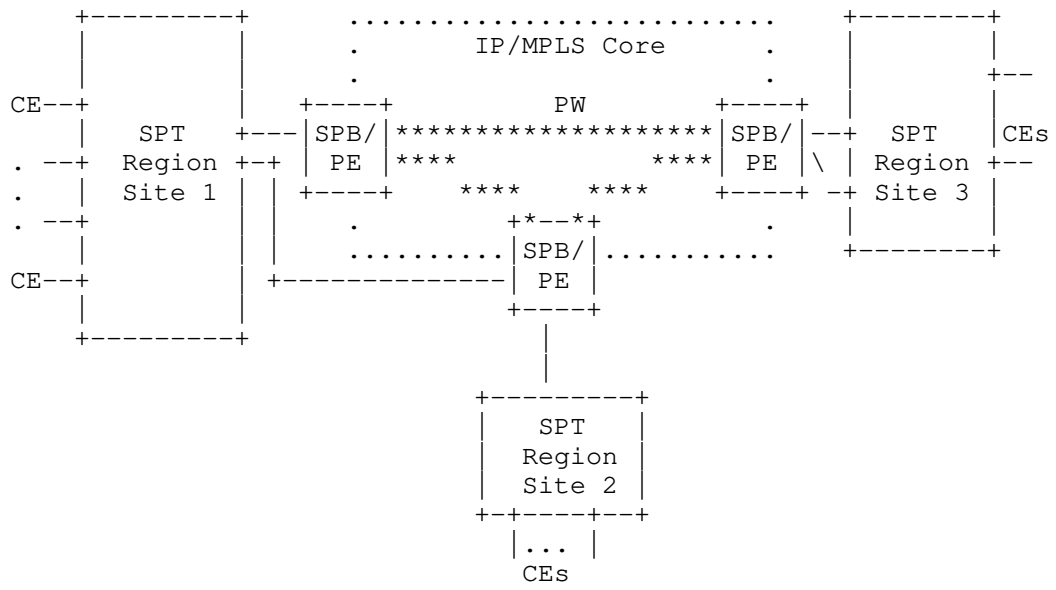


Figure 4 Hierarchical L2VPN with SPB and MPLS

There are several advantages to using SPT Bridge/MPLS based L2VPNs: 1) Scalability improvement; 2) Auto-configuration; 3) Good efficiency and loop prevention; 4) Multipath support (based on 802.1Qbp).

The solution also has advantages over some alternative solutions:

1. SPT Bridges provide deterministic forwarding behavior, allowing network tuning and traffic engineering;

2. SPB supports shortest path for both unicast and multicast traffic;

3. SPT Bridge core interfaces do not have to be upgraded to support a new encapsulation;

4. I-SID supports over 16M tenants; 5) Mature OAM functionality, Ethernet OAM (802.1ag and Y.1731) can be applied to SPB VLANs.

Note: It is possible to construct a Tiered L2VPN in the combination of Figure 4 and 3, i.e. some locations use SPB/PE enabled device and some location use separated SPT Bridge and PE devices in a Hierarchical L2VPN.

3. Security Considerations

The IS-IS authentication mechanism [RFC5304] [RFC5310] can be used to prevent fabrication of link-state control messages including those discussed in this document.

The use cases do not introduce any new security considerations for MPLS networks.

4. IANA Considerations

This document requires no IANA actions.

5. Acknowledgements

The authors would like to acknowledge the contributions of Donald E. rd Eastlake, 3 , Sue Hares, and Sam Aldrin.

6. References

6.1. Normative References

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

[RFC3518] Higashiyama, M., etc, "Point-to-Point Protocol (PPP) Bridging Control Protocol (BCP)", RFC 3518, April 2003.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006.

[RFC4447] Martini, L., etc, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC4447, April 2006.

[RFC4448] Martini, L., "Encapsulation Methods for Transport of Ethernet over MPLS Networks", BCP 116, RFC 4446, April 2006.

[RFC4618] Martini, L., "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) over MPLS Networks", BCP 116, RFC 4618, September 2006.

[RFC4762] Lasserre, M., and Kompella, V., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC4762, January 2007

[RFC5304] Li, T. and Atkinson, R, "IS-IS Cryptographic Authentication," RFC 5304, October 2008

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009

[RFC5659] Bocci, M and Bryant, S, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.

[802.1Q] IEEE Std 802.1Q 2011, Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks, August 2011.

[802.1Qbc] IEEE Std 802.1Qbc 2011, Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks-Amendment 16: Provider Bridging-Remote Customer Service Interfaces, September 2011.

6.2. Informative References

[PBB-VPLS] Sajassi, A, etc, "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-pbb-vpls-interop, work in progress, 2011

Authors' Addresses

Ben Mack-Crane
Huawei Technologies
5340 Legacy Drive
Plano, TX 75025

Phone: 630-810-1132
Email: ben.mackcrane@huawei.com

Lucy Yong
Huawei Technologies
5340 Legacy Drive
Plano, TX 75025

Phone: 469-227-5837
Email: lucy.yong@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 21, 2012

P. Kwok
P. Dutta
Alcatel-Lucent
F. Jounay
France Telecom

May 20, 2012

Pseudowire Communities
draft-pkwok-pwe3-pw-communities-03

Abstract

[RFC4447]describes a set of procedures for Pseudowire set-up and maintenance using LDP as signaling protocol. [I-D.ietf-pwe3-dynamic-ms-pw] extends the mechanisms described in [RFC4447] for dynamic placement of multi- segment pseudowires.

This document describes an extension to [RFC4447]procedures which may be used to pass additional information to S-PE/T-PEs when SS-PWs or MS-PWs are set-up.

The intention of the proposed technique is to aid in policy administration, specifically during MS-PW set-up across various S-PEs. The proposed method is very generic so that it can support the management of various parameters or rules while setting up pseudowires with minimal overhead.

Requirements Language

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 [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on November 21, 2012.

Table of Contents

1. Introduction

A Multi-Segment PW (MS-PW) is defined as a set of two or more contiguous segments that behave and function as a single point-to-point PW. An MS-PW enables service providers to extend the reach of PWs across multiple PSN domains.

To facilitate and simplify the control of dynamic MS-PW set-up across S-PEs, this document proposes a grouping or "community" of PWs so that PW set-up decision can also be based on the identity of the group. Such a scheme is expected to significantly simplify dynamic MS-PW signaling [I-D.ietf-pwe3-dynamic-ms-pw] that controls the MS-PW set-up across the Switching Provider Edge (S-PE) devices.

MS-PW spans across multiple autonomous systems or administrative domains. For security reasons, strict access control is required at S-PEs through which a PW enters another administrative domain. One way is for operators to define a policy at the S-PE that would match the PW set-up requests based on Target Attachment Individual Identifier (TAII) or Source Attachment Individual Identifier (SAII) or Attachment Group Identifier (AGI) etc. Such policies can be complex or very large, leading to administrative overheads or configuration mistakes. Rather, operators could define several tags/colors which can be associated with individual PWs when they are signaled. S-PEs can then apply PW policies based on the received tags, accordingly. This example application eliminates the primary motivation for a complex policy database that may result in the generation of very large PW prefix-based filter rules. A smaller policy database such as this also requires less maintenance, so shortening or eliminating out-of-band maintenance delays.

Another application of PW policies is in underlying transport applications. Each S-PE independently chooses a unidirectional PSN tunnel to map a set of PW segments to their next S-PE or T-PE. Such PSN Tunnels could be Label Distribution Protocol (LDP) [RFC5036] or Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) [RFC3209] or Labeled BGP [RFC3107] based LSPs. There is currently no signaling support in [I-D.ietf-pwe3-dynamic-ms-pw] to signal a preference for the type of PSN tunnel to bind a PW to at the S-PEs when multiple tunnel types are available. For example, LDP can be preferred over BGP tunnels when both forms of tunnels are available at an S-PE. Secondly, it is also possible that only a specific RSVP tunnel or class of RSVP tunnels based in Admin Groups is preferable to provide a traffic class or QoS treatment, or protection capability, and some form of control is required that LSPs are correctly used by S-PEs. One possible way is to manually configure filter rules by PW ID or AGI/SAII/TAII, but such rules can create significant maintenance overhead and be prone to configuration errors. Further, signaling

each of the various types of PSN tunnel selection criteria/ preferences in PW set-up messages adds significant burden to LDP label mapping procedures.

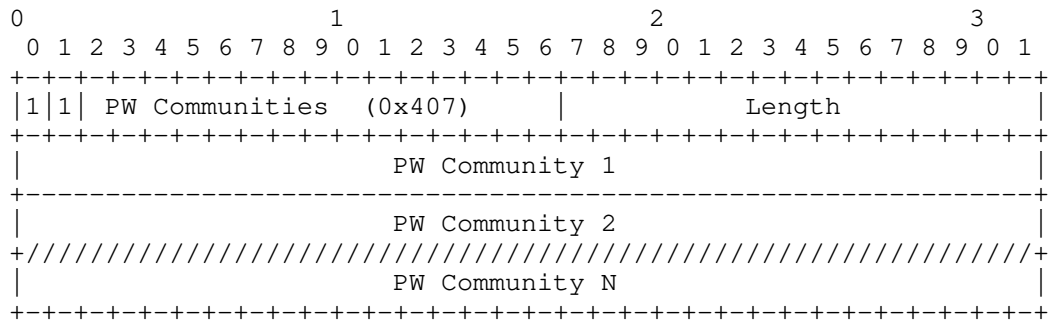
In Dynamic MS-PW, a T-PE or S-PE may need to choose one next-hop from several Equal Cost Multi-Path (ECMP) next-hops provided by best matching PW Route. One way to do ECMP selection is to apply some form of hash function on AGI/SAII/TAII of the PW but that strictly limits the MS-PW addressing schemes in order to get proper load distribution of MS-PWs across all next-hops. Operators need a predictable way for load balancing MS-PW across ECMP next-hops which is independent of MS-PW addressing schemes.

To address such policy management issues, this draft proposes a very simple solution that allows minimal manual intervention and configuration with no overhead in PW signaling. It introduces a concept of "PW Communities" that can be thought of as templates provisioned at a S-PE/T-PE, based on which of a certain set of rules are applied to all PWs that are tagged as belonging to same community.

Note that PW Community is different from PW Grouping (as defined using PW Group ID) defined in [RFC4447]]. PW Grouping is associated with binding of a set of PWs to a common event group for reduced signaling of various intensive events such as Label withdraw or PW Status Notification etc. However, PW Communities can be thought of a grouping of PWs from policy management perspective. It is not necessary that PW Grouping and PW Communities associated with a PW be correlated.

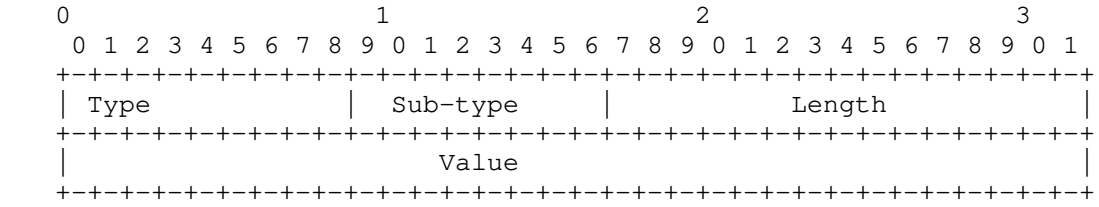
2. PW Communities

The PW Communities is an OPTIONAL TLV defined as follows which is in the format of LDP TLV [RFC5036].



U/F bits MUST be set to 1. Length is variable. Value field of the TLV contains a set of "PW Communities".

A PW Community is defined as follows:



Type field indicates the specific PW Community Type. The types are introduced to provide a broad classification of various PW communities based on the scope of applicability. Each community type further provides the flexibility to define sub-types within it. Length of a PW community is variable and to be defined by Type and Sub-Type associated with a PW community.

3. Defined PW Community Types

This section introduces a few PW community types and defines the format of the PW Community for those types.

3.1. PW Template Community

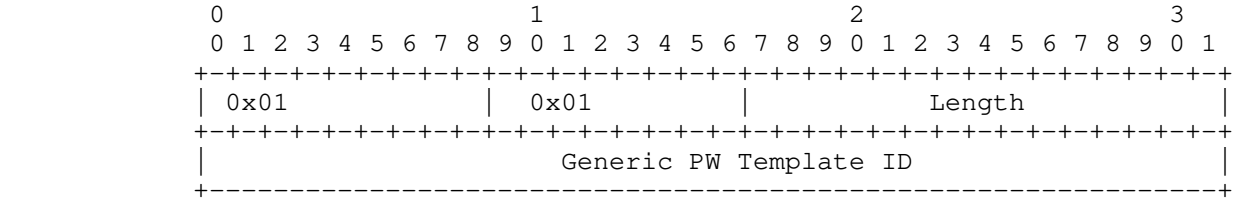
A PW Template community (PW Community Type 0x01) can be considered as a template that has a set of rules defined locally by a T-PE or S-PE. Each T-PE or S-PE can define its own set of rules and its upto the administrative domain to maintain congruities among PW community rules through which PW set-up process would follow. A LDP peer may use this community to control information it accepts, prefers or distributes to other peers.

A LDP peer receiving a PW set-up request (label mapping message) that does not carry the PW Template Community MAY append a PW Template Community TLV when propagating the label mapping message to next S-PE/T-PE.

A LDP peer receiving a PW set-up request with PW Template Community MAY modify the PW community according to local policy while propagating the request to the next-hop. Following sub-types of PW Template Community are defined in this document.

3.1.1. PW Generic Template Community

PW Generic Template Community is defined as sub-type 0x1 of the PW Template Community. The length field is 4 octets and contains a 32 bit generic identifier.

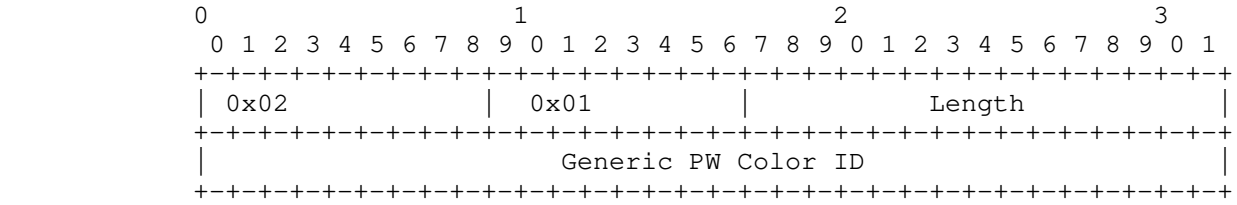


3.2. PW Color Community

A PW Color community (PW Community type 0x2) can be considered as a "coloring" of the PW that may be used by T-PE and S-PE in performing various hash functions required during PW set-up. One such application is in selection of PW signaling next-hop from multiple ECMP next-hops provided by the matching PW Route.

3.2.1. PW Generic Color Community

PW Generic Color Community is defined as sub-type 0x1 of the PW Color Community. The length field is 4 octets and contains a 32 bit generic identifier.



4. IANA Considerations

This document proposes an OPTIONAL LDP PW Communities TLV, with a proposed type of 0x407, to be allocated from the LDP TLV type registry.

5. Security Considerations

This document does not impose additional security considerations to what is defined in [RFC5036],[RFC4447] and [I-D.ietf-pwe3-dynamic-ms-pw]

6. Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions from Mathew Bocci, Mustapha Aissaoui and Wim Henderickx.

7. References

7.1. Normative References

[I-D.ietf-pwe3-dynamic-ms-pw]
Martini, L., Bocci, M., and F. Balus, "Dynamic Placement of Multi Segment Pseudowires", draft-ietf-pwe3-dynamic-ms-pw-14 (work in progress), July 2011.

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

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

7.2. References

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

Authors' Addresses

Paul Kwok
Alcatel-Lucent
701 E Middlefield Road
Mountain View, CA 94043
USA

Email: paul.kwok@alcatel-lucent.com

Pranjal Kumar Dutta
Alcatel-Lucent
701 E Middlefield Road
Mountain View, CA 94043
USA

Email: pranjal.dutta@alcatel-lucent.com

Frederic Jounay
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cidex,
France

Email: frederic.jounay@orange-ftgroup.com

TRILL Working Group
INTERNET-DRAFT
Intended status: Informational

Lucy Yong
Donald Eastlake
Sam Aldrin
Huawei R&D USA
Jon Hudson
Brocade

Expires: September 9, 2012

March 10, 2012

Transparent Interconnection of Lots of Links (TRILL) over MPLS Pseudo Wires
<draft-yong-trill-trill-o-mpls-01.txt>

Abstract

This informational document describes ways to interconnect TRILL RBridges by using MPLS Pseudo Wire (PW) services with existing TRILL and MPLS standards so as to form a unified TRILL campus.

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list <rbridge@postel.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) standard [RFC6325] [RFC6326] provides optimal pair-wise data frame forwarding without configuration in multi-hop networks with arbitrary topology and link technology, and supports multipathing of both unicast and multicast traffic. TRILL enables a new method to construct a campus or data center network. Devices that implement TRILL are called RBridges (Routing Bridges) or TRILL Switches.

This document describes the use of MPLS Pseudo Wire or VPLS links by TRILL.

1.1 Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Acronyms used in this document include the following:

AC - Attachment Circuit

IS-IS - Intermediate System to Intermediate System

MPLS - Multi-Protocol Label Switching

PE - Provider Edge

PPP - Point-to-Point Protocol

PW - Pseudo Wire

QoS - Quality of Service

RB - RBridge

RBridge - Routing Bridge

TRILL - Transparent Interconnection of Lots of Links

TRILL Switch - An alternative term for an RBridge

VPLS - Virtual Private LAN Service

VSI - Virtual Service Instance

2. Use Cases

TRILL campuses at different locations may interconnect by networks that are implemented with different technologies to form a unified RBridge campus. This section describes use cases assuming that IP/MPLS technology is available. From the MPLS network view, a pair of RBridges can be directly connected with a pseudo wire or an RBridge can act as a Customer Edge device that connects to a Provider Edge device via an attachment circuit. RBridge ports [RFC6325], by default, support both point-to-point links and multi-access links.

Section 2.1 describes point-to-point links, i.e. TRILL over an Ethernet or PPP point-to-point link that is over an MPLS network. Section 2.2 describes TRILL over a bridged LAN or equivalent that is implemented by MPLS/VPLS.

2.1 Point-To-Point Interconnection

Either an Ethernet or PPP link over an MPLS network interconnets two RBridge ports. This can be either a direct pseudo wire between the RBridges or by attachment circuits from each RBridge port to Provider Edge devices that provide a transparent tunnel between the provider edge attachment points.

MPLS already supports many pseudo wire transport encapsulations [RFC4446]. Two types of TRILL links between RBridges have been standardized are Ethernet [RFC6325] and PPP [RFC6361]. Pseudo wire encapsulations for these two interfaces are specified in [RFC4448] and [RFC4618], respectively.

The method described in 2.1.1 below is typically suitable when the TRILL and MPLS facilities have common management while the method described in 2.1.2 is typically suittable when the TRILL and MPLS facilities are separately managed. In the case of different management, the core MPLS operator can sell a VPWS service to an RBridge operator.

In both cases, the MPLS label switched routers involved need no awareness of TRILL.

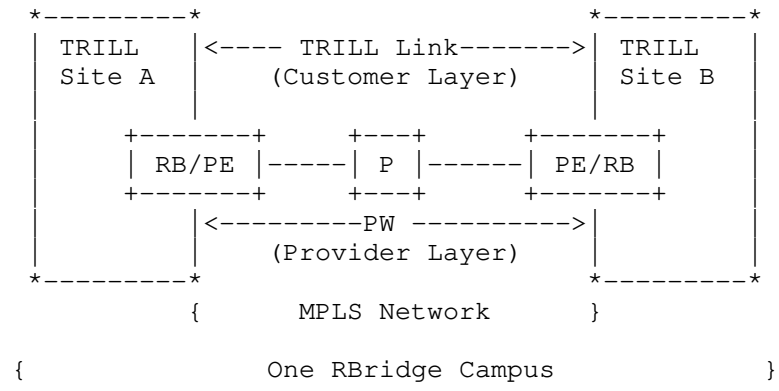
A pseudo wire may cross multiple MPLS domains [RFC5659]. In these cases, RBridges may be considered to connect to T-PEs and it works in the same way as a single domain. The MPLS network can provide transport resiliency for a pseudo wire. The dual homing (two attachment circuits) can be used for attachment circuit protection. In this case, two TRILL links are established; RBridges can perform load balancing over two links.

2.1.1 Direct Point-to-Point

Two RBridge ports can be connected directly by an MPLS pseudo wire. This implies that the RBridges, which are TRILL routers, are also acting as label switched routers. The pseudo wire can be either Ethernet over MPLS or PPP over MPLS but PPP over MPLS is recommended because it saves 16 bytes per frame. The pseudo wire between two RBridge ports can be auto-configured [RFC4447] or manually configured; the two RBridges then appear directly interconnected with a transparent link.

(Technically speaking, it is possible to create a specially designated TRILL encapsulated pseudo wire for point-to-point TRILL over MPLS. However, the authors think that this is not worth the effort in this case because of available technologies, particularly the highly-efficient PPP link technology.)

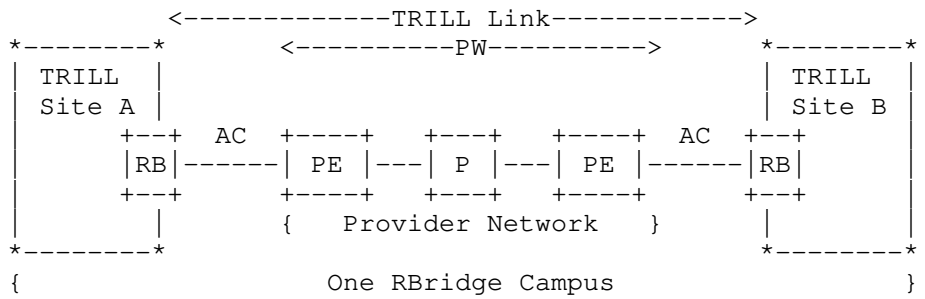
From a customer/provider point of view, this can also be though of as shown in the following diagram:



The interworking between the RBridge network and the MPLS network is within the combined TRILL/MPLS device. This has a similar architecture to MPLS/VPLS [RFC4762].

2.1.2 Provider Point-to-Point Service

Two RBridge ports may also be connected by attachment circuits (ACs) to Provider Edge (PE) devices that are part of a typically separately managed provider network. The provider network then provides a transparent path between these attachment circuits, connecting the RBridge ports. The following diagram illustrates this arrangement:



If the attachment circuits are Ethernet, the two PEs would be configured to form a pseudo wire with Ethernet encapsulation [RFC4448]. This pseudo wire can be auto-configured [RFC4447] or manually configured; the two RBridges then appear directly interconnected with an Ethernet link. The provider edge devices SHOULD use the Raw mode and non-service-delimiting, which provides a transparent Ethernet transport.

If the attachment circuits are PPP link type, the two PEs must be configured to form a PW with PPP encapsulation [RFC4618]. After the pseudo wire is established between two PEs, the two RBridges then appear directly connected with a PPP link. The PPP link configuration will be more efficient than the Ethernet point-to-point configuration; it saves about 16 bytes per frame by replacing the TRILL Outer.MacDA, Outer.MacSA, Outer.VLAN, and outer Ethertype with a PPP code point [RFC6361].

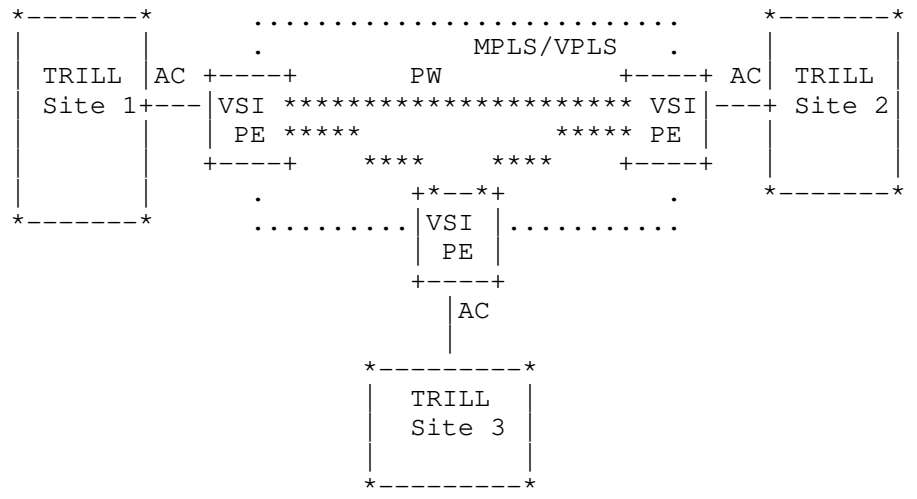
In both case, the pseudo wire provides transparent transport between attachment circuits. The TRILL link and the connectivity over that link is automatically discovered and established through TRILL IS-IS control messages.

2.2 Multi-Access Link Interconnection

Multiple RBridges may interconnect via an [802.1Q] Bridged LAN that acts as a hub. The bridged LAN simply forwards on the outer Ethernet header of the TRILL over Ethernet frames. This configuration creates what appears to each connected RBridge as a multi-access link. In other words, each RBridge connecting to a bridged LAN normally has connectivity to every other RBridge connecting to the same bridged LAN.

MPLS/VPLS can provide the same capability when multiple parts of an RBridge campus are interconnected over an IP/MPLS network and make each RBridge attaching to the VPLS appear as having a multi-access TRILL link. The diagram below shows one RBridge campus is split between three different sites and the use of MPLS/VPLS for RBridge

interconnection. Ethernet attachment circuits and pseudo wires are assumed.



One VPLS instance is configured on three Provider Edge (PE) devices and the pseudo wires are configured for the VPLS instance. Each RBridge Site connects to the VSI on a PE via an attachment circuit. The VSI on a PE forwards TRILL frames based on the outer Ethernet header of the frames [RFC6325]. Either BGP [RFC4761] or LDP [RFC4762] protocol can be used to automatically construct the VPLS instance on the PEs.

The choice of three VSIs and three singly connected sites was for illustrative purposes. There could be two or more than three. Furthermore, each TRILL site could have multiple connections to the VPLS network and/or direct connection via other technologies or VPLS networks to one or more of the other sites. TRILL sorts all this out and router properly.

A PE may connect to several different RBridge campuses that belong to different customers. Separated VPLS instances are configured for individual customers and customer traffic is isolated by VPLS instance. The PE treats an RBridge as a generic Ethernet customer devices and has no awareness of TRILL. The outer Ethernet MAC of TRILL frames may be either a next-hop RBridge MAC address (for unicast frames) or one of TRILL defined multicast addresses (ALL-IS-IS-RBridges and All-RBridges) [RFC6325]. The VSI at each PE learns the source MAC addresses on each VSI interface and forward the frame based on the destination MAC. For the multicast frames, the VSI replicates the frames to all pseudo wires it associates. If a VPLS is configured with some optimization capability [VPLS-BCAST], the multicast frames can be delivered over a point-to-multipoint pseudo wire while unicast frames are carried over a point-to-point pseudo

wire.

The scenario above can also be extended to multiple RBridge interconnections when a device serves both the RBridge and PE functions, similarly to the case in Section 2.1.2 and 2.1.1 above.

Note: If the customer devices associated with one VPLS instances happen to include some RBridges and some end stations or IEEE 802.1Q bridges providing paths to end stations, TRILL will, by default, be able to handle this by providing both through service and end station service. However, the end station addresses will be visible to the VPLS instance. If, in such a case, all the RBridge ports connected to the VPLS are configured as trunk ports (see Section 4.9.2 of [RFC6325]), then they will not provide any end station service.

3. RBridge Behavior for MPLS Pseudo Wire

This section describes RBridge behaviors for TRILL Ethernet or TRILL PPP links over MPLS pseudo wire (PW) as described in Sections 2.1.

1. For two RBridge ports connecting via a PPP pseudo wire, the ports MUST be configured as IS-IS point-to-point because there are no subnetwork point of attachment (SNPA/MAC) addresses of the end points at the PPP protocol layer. Thus TRILL will use IS-IS P2P Hellos that, as described in "Point-to-Point IS to IS Hello PDU" (section 9.7 of [IS-IS]), do not use Neighbor TLVs or require SNPAs. However, as described section 4.2.4.1 of [RFC6325], three-way IS-IS handshake using extended circuit IDs is required.

2. Any MPLS forwarder within an MPLS label switched path does not change the TRILL Header Hop Count. RBridges are not aware of the packet forwarders in with the MPLS network.

3. If it is desired for MPLS label switched routers to perform QoS in the same way as RBridges do, an Ethernet path MUST be used and RBridges MUST be configured to send an Outer.VLAN tag on the RBridge port leading to the pseudo wire. The PE can then copy the priority value from the Outer.VLAN tag to the COS field of the pseudo wire label prior to the forwarding [RFC5462].

4. TRILL MTU-probe and TRILL MTU-ack messages (section 4.3.2 of [RFC6325]) are not needed on a pseudo wire link. Implementations MUST NOT send MTU-probe and SHOULD NOT reply to these messages. The MTU pseudo wire interface parameter SHOULD be used instead. PE MUST configure the MTU size as the originating RBridges Size specified in Section 4.3.1 of [RFC6325].

4. IANA Considerations

No IANA action is required by this document. RFC Editor: Please remove this section before publication.

5. Security Considerations

The IS-IS authentication mechanism [RFC5304] [RFC5310], at the TRILL IS-IS layer, can be used to prevent fabrication of link-state control messages over TRILL links including those discussed in this document.

For general TRILL protocol security considerations, see [RFC6325].

Use cases in which the path between RBridges transits a provider network under separate administration may represent a substantial increase in the threat of observation, deletion, modification, or insertion of data or control information. Under such circumstances consideration should be give to the use of security at the TRILL link level, such as [802.1AE] if the path between the RBridge ports is Ethernet or security as suggested in [RFC6361] if that path is PPP.

6. Acknowledgements

The authors sincerely acknowledge the contributions of Ben Mack-Crane and Sue Hares.

7. Normative References

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

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to EdgeEmulation (PWE3)", BCP 116, RFC 4446, April 2006.

[RFC4447] Martini, L., Ed., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC4618] Martini, L., "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) over MPLS Networks", BCP 116, RFC 4618, September 2006.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC4761, January 2007.

[RFC4762] Lasserre, M. and Kompella, V, "Virtual Private LAN Service(VPLS) Using Label Distribution Protocol (LDP) Signaling",RFC4762, January 2007

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

[RFC5304] Li, T. and Atkinson, R, "IS-IS Cryptographic Authentication," RFC 5304, October 2008

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009

[RFC5462] Andersson, L. and Asati, R., "Multiprotocol Label Switching (MPLS) Label Stack entry: "Exp" Field Rename to "Traffic Class" Field", RFC5462, February 2009

[RFC5659] Bocci, M and Bryant, S, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October2009.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC6325, July 2011.

[RFC6326] Eastlake 3rd, D., Banerjee, A., Dutt, D., Perlman, R.,

andGhanwani, A. "Transparent Interconnection of Lots of Links(TRILL) Use of IS-IS", RFC6326, July 2011.

[RFC6361] Carlson, J., and D. Eastlake, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC6361, August 2011.

8. Informative References

[802.1AE] "IEEE Standard for Local and metropolitan area networks / Media Access Control (MAC) Security", 802.1AE-2006, 18 August 2006.

[802.1Q] IEEE 802.1, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, May 2011.

[IS-IS] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO/IEC10589:2002, Second Edition, Nov 2002

[VPLS-BCAST] Delord, S, and Key, R., "Extension to LDP-VPLS for Ethernet Broadcast and Multicast", draft-ietf-l2vpn-ldp-vpls-broadcast-exten-02, work in progress, 2011.

Authors' Addresses

Lucy Yong
Huawei R&D USA
5340 Legacy Drive
Plano, TX 75025

Phone: +1-469-227-5837
Email: lucy.yong@huawei.com

Donald E. Eastlake, 3rd
Huawei R&D USA
155 Beaver Street
Milford, MA 01757 USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Sam Aldrin
Huawei R&D USA
2330 Central Expressway
Santa Clara, CA 95050

Phone: +1-408-330-4517
Email: sam.aldrin@huawei.com

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134

Phone: +1-408-333-4062
jon.hudson@brocade.com

Copyright and IPR Provisions