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.

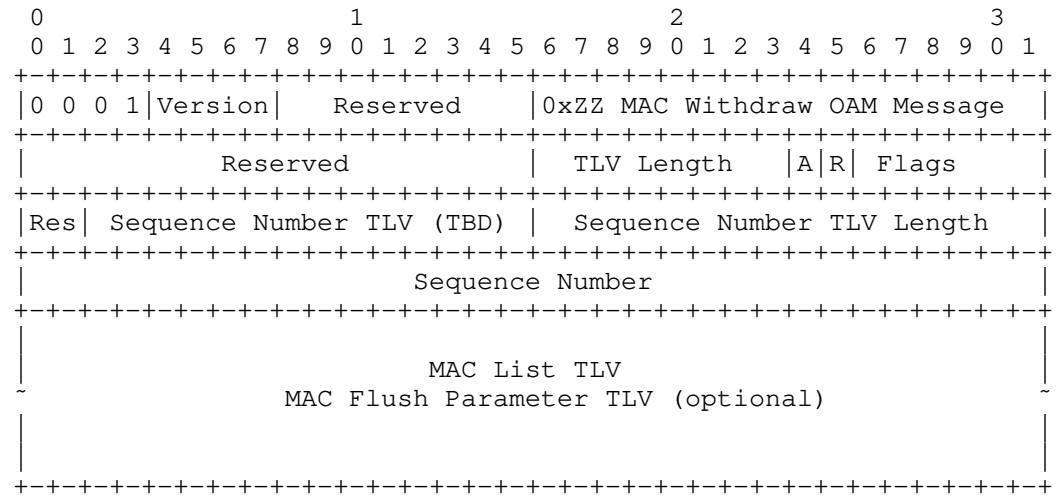


Figure 1: MAC Address Withdraw PW OAM Packet Format

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: June 1, 2014

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

November 28, 2013

LDP extensions for Pseudowire Binding to LSP Tunnels
draft-ietf-pwe3-mpls-tp-pw-over-bidir-lsp-02.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 June 1, 2014.

Copyright Notice

Table of Contents

1. Introduction

Pseudowire (PW) Emulation Edge-to-Edge (PWE3) [RFC3985] is a mechanism to emulate layer 2 services, such as Ethernet Point-to-Point (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 PWs 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 LSPs, 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 a single bidirectional co-routed LSP or two LSPs that share the same route.

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.

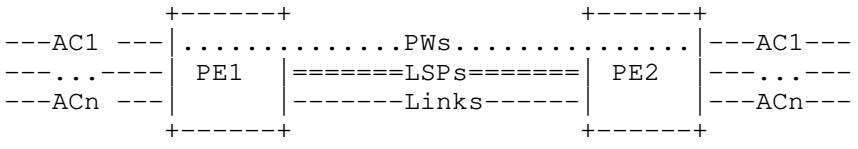


Figure 1: Explicit PW-to-LSP binding scenario

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 PWs 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 PWs 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.

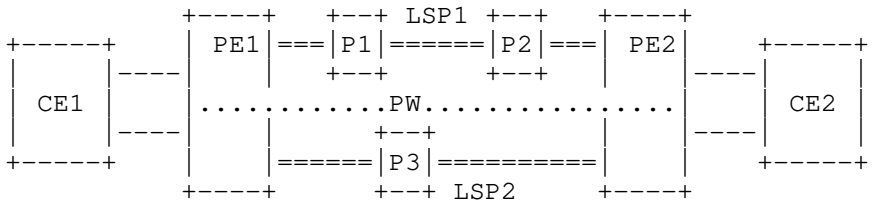


Figure 2: Inconsistent SS-PW to LSP binding scenario

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 optional 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 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. Co-routed binding: the 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:

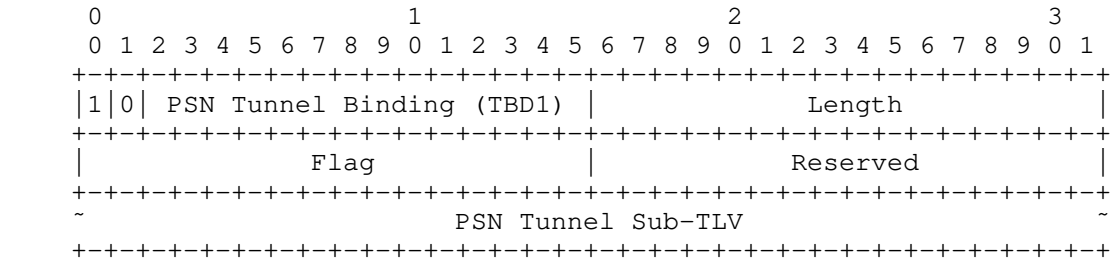
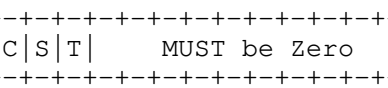


Figure 3: PSN Tunnel Binding TLV

The PSN Tunnel Binding TLV type is TBD1.

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 (Co-routed 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 co-routed LSP as the PSN tunnel. If there is no such tunnel available, it 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/LSP 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.

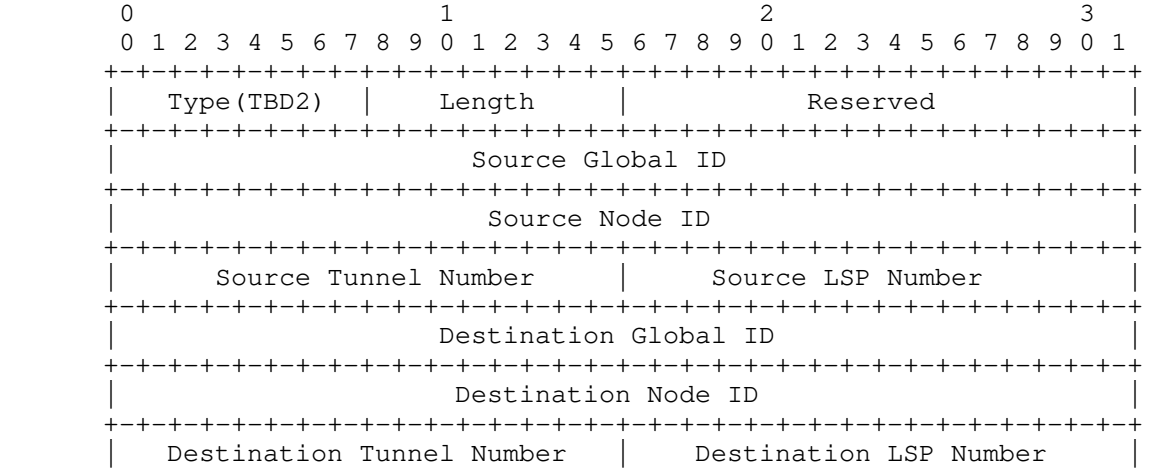
T (Tunnel Representation) bit: This indicates the format of the LSP tunnels. When the bit is set, the tunnel uses the tunnel information to identify itself, and the LSP Number fields in the PSN Tunnel sub-TLV (Section 2.1.1) MUST be set to zero. Otherwise, both tunnel and LSP information of the PSN tunnel 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. Otherwise, a Label Release message with status code set to "The C-bit and S-bit can not both be set" MUST be replied, and the PW will not be established.

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.



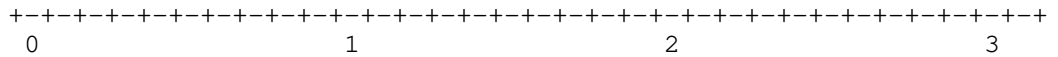


Figure 4: IPv4 PSN Tunnel sub-TLV format

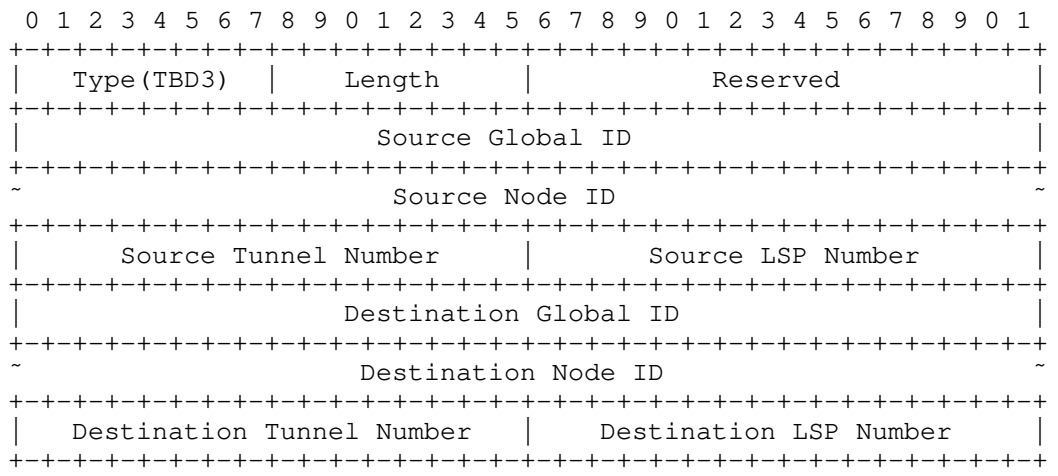


Figure 5: IPv6 PSN Tunnel sub-TLV format

The definition of Source and Destination Global/Node IDs and Tunnel/ LSP Numbers are derived from [RFC6370]. This is to describe the underlying LSPs. Note that the LSPs in this notation are 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 PEs for a particular PW may assume the active roles.

2. MS-PW: One PE is active, while the other is passive. The PWs 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.

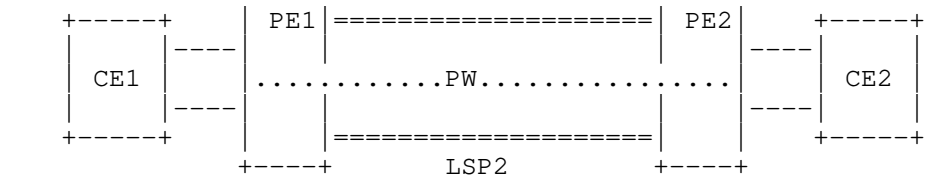


Figure 6: PSN binding operation in SS-PW environment

As outlined previously, there are two types of binding request: co-routed 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 co-routed 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 is numerically lower, PE2 needs to find/ establish a tunnel/LSP that meets the co-routed 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 co-routed 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 related PSN LSP.

After a PW established, the operators may choose to move the PWs from the current tunnel/LSPs to other tunnel/LSPs. Or, the underlying PSN tunnel 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 LSPs.

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 co-routed/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.

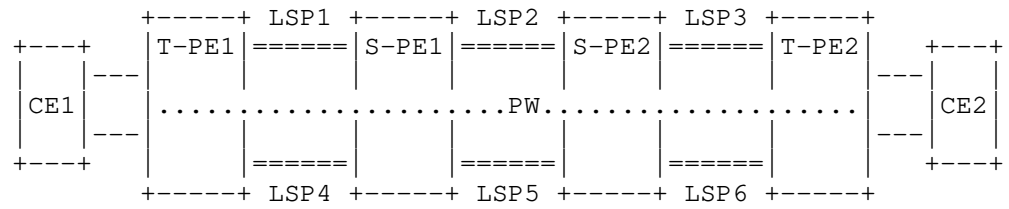


Figure 7: PSN binding operation in MS-PW environment

When an active PE (that is, T-PE1) starts to signal 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 indicate 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 co-routed 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 co-routed 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 (TBD1) 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 - TBD2

IPv6 PSN Tunnel sub-TLV - TBD3

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" - TBD4

"The C and S bit can not be both set" -TBD5

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-19 (work in progress), October 2013.

[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.chen@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

Internet Working Group
Internet Draft
Intended status: Standards Track

Y. Jiang
Y. Luo
Huawei
E. Mallette
Bright House Networks
C. Shen
China Telecom
Y. Shen
Juniper Networks
W. Cheng
China Mobile
G. Zhou
China Unicom

Expires: April 2015

October 25, 2014

Multi-chassis PON Protection in MPLS
draft-jiang-pwe3-mc-pon-03.txt

Status of this Memo

Copyright Notice

Abstract

MPLS is being deployed deeper into operator networks, often to or past the access network node. Separately network access nodes such as PON OLTs have evolved to support first-mile access protection, where one or more physical OLTs provide first-mile diversity to the customer edge. Multi-homing support is needed on the MPLS-enabled PON OLT to provide resiliency for provided services. This document describes the multi-chassis PON protection architecture in MPLS and also proposes the ICCP extension to support it.

Table of Contents

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

DSL Digital Subscriber Line

FTTx Fiber-to-the-x (FTTx, x = H for home, P for premises, C for curb)

ICCP Inter-Chassis Communication Protocol

OLT Optical Line Termination

ONU Optical Network Unit

MPLS Multi-Protocol Label Switching

PON Passive Optical Network

RG Redundancy Group

3. Introduction

MPLS is being extended to the edge of operator networks, as is described in the seamless MPLS use cases [SEAMLESS], and the MS-PW with PON access use case [RFC6456]. Combining MPLS with OLT access further facilitates a low cost multi-service convergence.

Tens of millions of FTTx lines have been deployed over the years, with many of those lines being some PON variant. PON provides operators a cost-effective solution for delivering high bandwidth (1Gbps or even 10Gbps) to a dozen or more subscribers simultaneously.

In the past, access technologies such as Passive Optical Network (PON) and Digital Subscriber Line (DSL) are usually used for subscribers, and no redundancy is provided in their deployment.

But with the rapid growth of mobile data traffic, more and more LTE small cells and Wi-Fi hotspots are deployed. PON is considered as a viable low cost backhaul solution for these mobile services. Besides its high bandwidth and scalability, PON further provides synchronization features, e.g., SyncE and IEEE1588 functionality, which can fulfill synchronization needs of mobile backhaul services.

The Broadband Forum specifies reference architecture for mobile backhaul network using MPLS transport in [TR-221] where PON can be the access technology, and is further working on PON-based mobile backhaul network architecture in [SD-331].

Unlike typical residential service where a single or handful of end-users hangs off of a single PON OLT port in a physical optical distribution network, a PON port that supports a dozen LTE small cells or Wi-Fi hotspots could be providing service to hundreds of simultaneous subscribers. Small cell backhaul often demands the economics of a PON first-mile and yet expects first-mile protection commonly available in point-to-point access portfolio.

Some optical layer of protection mechanisms, such as Trunk and Tree protection, are specified in [IEEE-1904.1] to avoid single point of failure in the access. They are called Type B and Type C protection respectively in [G983.1].

Trunk protection architecture is an economical PON resiliency mechanism, where the working OLT and the working link between the working splitter port and the working OLT (i.e., the working trunk fiber) is protected by a redundant protection OLT and a redundant trunk fiber between the protection splitter port and the protection OLT, however it only protects a portion of the optical path from OLT to ONUs. This is different from the more complex and costly Type C protection architecture where there is a working optical distribution network path from the working OLT and a complete protected optical distribution network path from the protection OLT to the ONUs. Figure 1 demonstrates a typical scenario of Trunk protection.

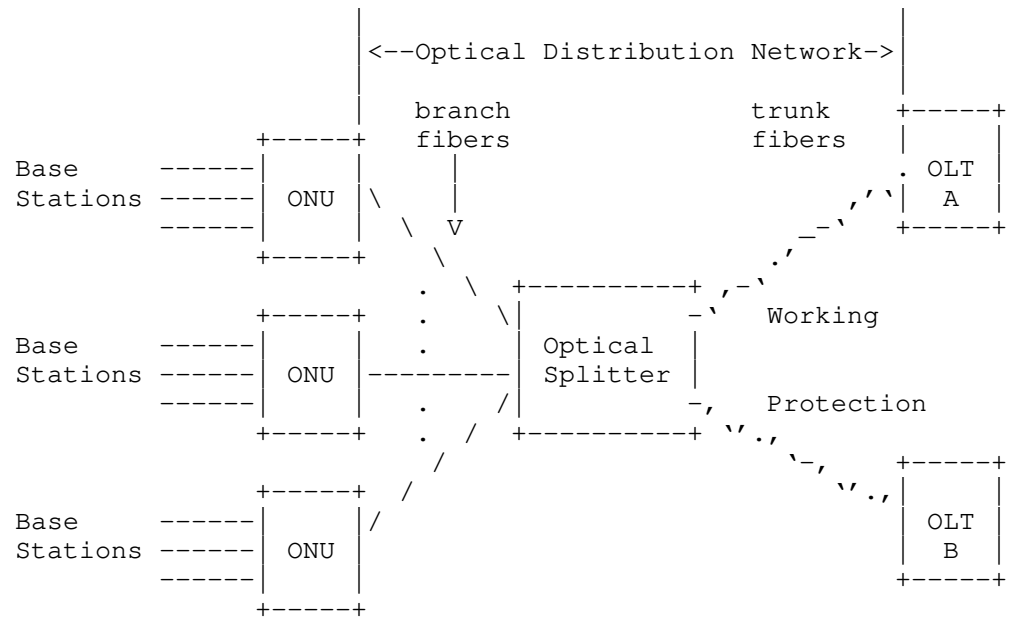


Figure 1 Trunk Protection Architecture in PON

Besides small cell backhaul, this protection architecture can also be applicable to other services, for example, DSL and Multi-System Operator (MSO) services. In that case, an ONU in Figure 1 can play the similar role as a Digital Subscriber Line Access Multiplexer (DSLAM) and dozens of Customer Premises Equipments (CPEs) or cable modems may be attached to it.

In some deployments, it is also possible that only some ONUs are needed to be protected.

The PON architecture depicted in Figure 1 can provide redundancy in its physical topology, however, all traffic including link OAM are blocked on the protection link which frustrates end to end protection mechanisms such as ITU-T G.8031. Therefore, some standard signaling mechanisms are needed between OLTs to exchange information, for example, PON link status, registered ONU information, and network status, so that protection and restoration can be done both rapidly and reliably, especially when the OLTs also support MPLS.

ICCP [ICCP] provides a framework for inter-chassis synchronization of state and configuration data between a set of two or more PEs. Currently ICCP only defines application specific messages for PW redundancy and mLACP, but it can be easily extended to support PON as an Attachment Circuit (AC) redundancy.

This document proposes the extension of ICCP to support Multi-chassis PON protection in MPLS.

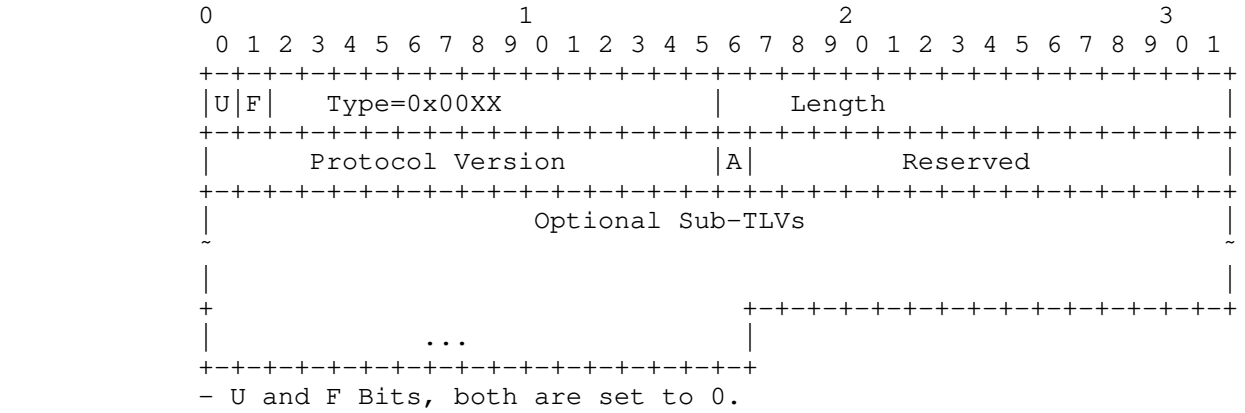
4. ICCP Protocol Extensions

4.1. Multi-chassis PON Application TLVs

A set of multi-chassis PON application TLVs are defined in the following sub-sections.

4.1.1. PON Connect TLV

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



- U and F Bits, both are set to 0.

- Type, set to 0x00XX for "PON 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 PON specific protocol for the purposes of inter-chassis communication. This is set to 0x0001.

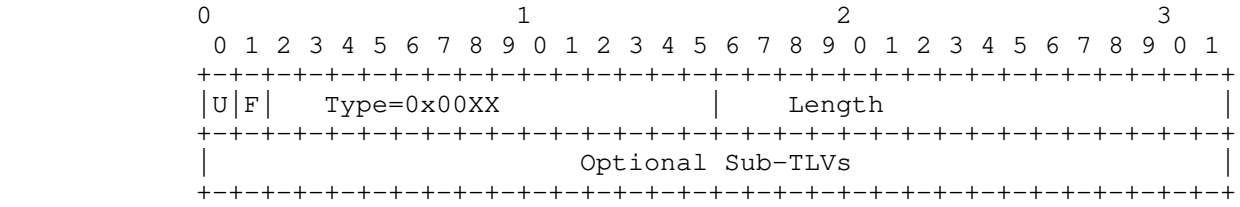
- A Bit, Acknowledgement Bit. Set to 1 if the sender has received a PON Connect TLV from the recipient. Otherwise, set to 0.

- Reserved, Reserved for future use.

- Optional Sub-TLVs, there are no optional Sub-TLVs defined for this version of the protocol.

4.1.2. PON Disconnect TLV

This TLV is included in the RG Disconnect message to indicate that the connection for the PON application is to be terminated.



- U and F Bits, both are set to 0.

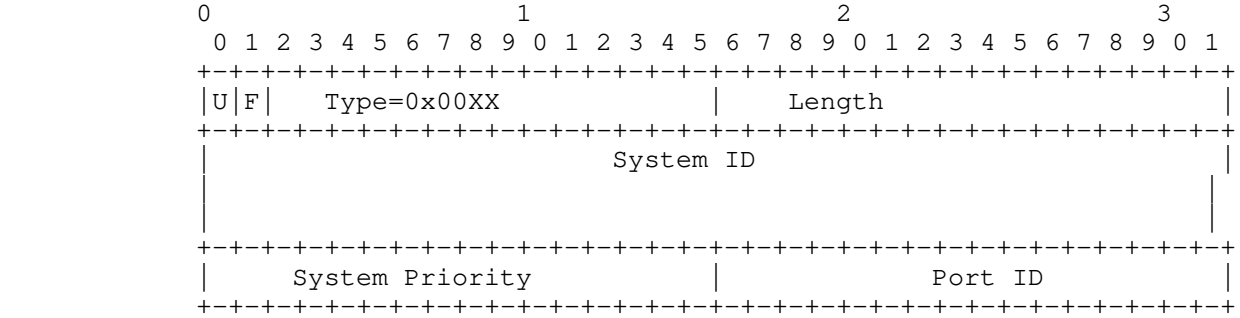
- Type, set to 0x00XX for "PON Disconnect TLV".

- 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.

4.1.3. PON Configuration TLV

The "PON Configuration TLV" is included in the "RG Application Data" message, and announces an OLT's system parameters to other members in the same RG.



- U and F Bits, both are set to 0.

- Type, set to 0x00XX for "PON Configuration TLV".

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

- System ID, 8 octets encoding the System ID used by the OLT, which is the Chassis MAC address. If a 6 octet System ID is used, the least significant 2 octets of the 8 octet field will be encoded as 0000.

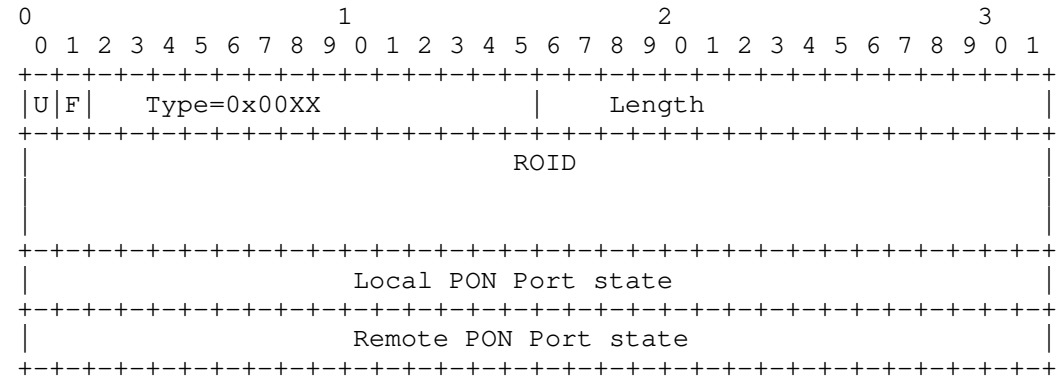
- System Priority, 2 octets encoding the System Priority.

- Port ID, 2 octets PON Port ID.

Further configuration considerations such as multicast table and ARP table for static MAC addresses will be added in a next version.

4.1.4.PON State TLV

The "PON State TLV" is included in the "RG Application Data" message, and used by an OLT to report its PON states to other members in the same RG.



- U and F Bits, both are set to 0.

- Type, set to 0x00XX for "PON State TLV"

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

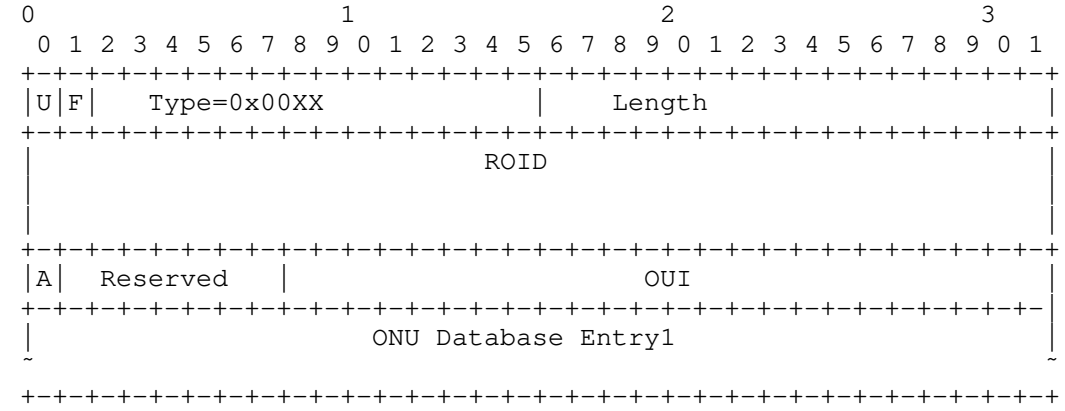
- ROID, as defined in the ROID section of [ICCP].

- Local PON Port State, the status of the local PON port as determined by the sending OLT (PE). The last bit is defined as Fault indication of the PON Port associated with this PW (1 - in fault).

- Remote PON Port State, the status of the remote PON port as determined by the remote peer of the sending OLT (PE). The last bit is defined as Fault indication of the PON Port associated with this PW (1 - in fault).

4.1.5.PON ONU Database Sync TLV

This TLV is used to communicate the registered ONU database associated with a PON port between the active and standby OLT. This message is used to both transmit the PON ONU Database from working OLT to protect OLT and to communicate the PON ONU database status between protect OLT and working OLT.



- U and F Bits, both are set to 0.

- Type, set to 0x00XX for "PON ONU Database Sync TLV"

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

- ROID, defined in the ROID section of [ICCP].

- A bit, Acknowledgement bit. Set to 1 if the receiver has received a PON ONU Database Sync. Otherwise, set to 0.

- Reserved, reserved for future use.

- OUI, the 3-byte [IEEE-802.3] organization unique identifier that uniquely identifies the format for describing the registered ONU database information. There are multiple PON standards and are varying implementations within a given PON standard which likely have

different required information, format, etc., related to the ONU Database Entry.

- ONU Database Entry, there may be one or more ONU Database Entries transmitted in the PON ONU Database Sync TLV, each of which would describe a registered ONU. The format of the ONU Database Entry is outside the scope of this document and will be defined by the relevant PON standard organization.

5. PON ONU Database Synchronization

Without an effective mechanism to communicate the registered ONUs between the working and protection OLT, all registered ONUs would be de-registered and go through re-registration during a switchover, which would significantly increase protection time. To enable faster switchover capability, the work OLT must be able to communicate the registered ONUs associated with an ROID to the protection OLT.

The PON ONU Database Synchronization would begin once the ICCP PON Application enters OPERATIONAL state. The working OLT, the one with the working link member for the ROID, would begin transmitting the database of actively registered ONUs to the protection OLT for the same ROID. Each instance of the PON ONU Database Sync TLV describes a set of ONU Database Entries. Each ONU Database Entry would describe a registered ONU.

The transmission of PON ONU Database Descriptors for a given ROID is only unidirectional - from the working OLT to the protection OLT. The protection OLT would only be responsible for acknowledging the received message to provide a reliable database synchronization mechanism. As ONUs register and deregister from the working OLT, the working OLT would transmit PON ONU Database Synchronization TLV including only the updated ONU Database Entries.

If protected ONUs and unprotected ONUs are miscellaneously attached to the same splitter, only the protected ONUs needs to be synchronized. The specific ONUs which needs to be synchronized can be policy driven and provisioned in the management plane, or by some other signaling options.

6. Multi-chassis PON application procedures

Two typical MPLS protection network architectures for PON access are depicted in Fig.2 and Fig.3 (their PON access segments are the same as in Fig.1 and thus omitted for simplification). OLTs with MPLS functionality are connected to a single PE (Fig.2) or dual home PEs (Fig.3) respectively, i.e., the working OLT to PE1 by a working PW and the protection OLT to PE1 or PE2 by a protection PW, thus these devices constitute an MPLS network which provides PW transport services between ONUs and a CE.

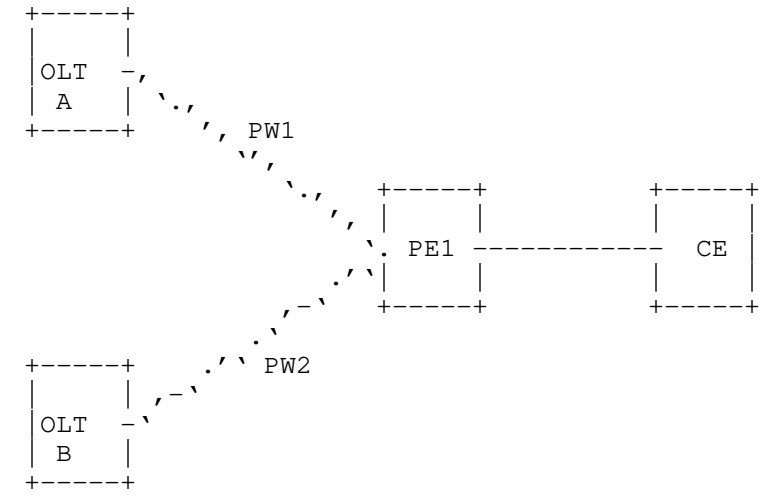


Figure 2 An MPLS Network with a Single PE

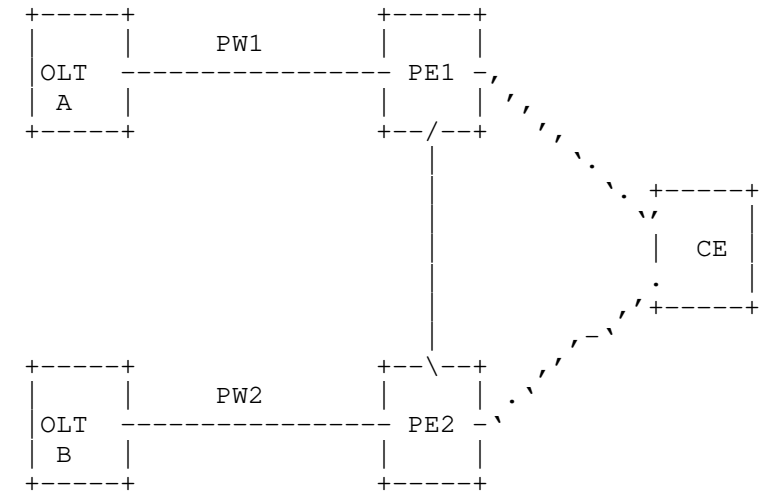


Figure 3 An MPLS Network with Dual-homing PEs

Faults may be encountered in PON access links, or in the MPLS network (including the working OLT). Procedures for these cases are described in this section (it is assumed that both OLTs and PEs are working in independent mode of PW redundancy [RFC6870]).

6.1. Protection procedure upon PON link failures

When a fault is detected on a working PON link, a working OLT MUST turn off its associated PON interface so that the protection trunk link to the protection OLT can be activated, then it MUST send an LDP fault notification message (i.e., with the status bit "Local AC (ingress) Receive Fault " being set) to its peer PE on the remote end of the PW. At the same time, the working OLT MUST send an ICCP message with PON State TLV with local PON Port State being set to notify the protection OLT of the PON fault.

Upon receiving a PON state TLV where Local PON Port state is set, a protection OLT MUST activate the protection PON link in the protection group, and advertise a notification message for the protection PW with the Preferential Forwarding status bit of active to the remote PE.

According to [RFC6870], the remote PE(s) can match the local and remote Preferential Forwarding status and select PW2 as the new active PW to which to send traffic.

6.2. Protection procedure upon PW failures

Usually MPLS networks have its own protection mechanism such as LSP protection or Fast Reroute (FRR). But in a link sparse access or aggregation network where protection for a PW is impossible in its LSP layer, the following PW layer protection procedures can be enabled.

When a fault is detected on its working PW (e.g., by VCCV BFD), a working OLT SHOULD turn off its associated PON interface and then send an ICCP message with PON State TLV with local PON Port State being set to notify the protection OLT of the PON fault.

Upon receiving a PON state TLV where Local PON Port state is set, the protection OLT MUST activate its PON interface to the protection trunk fiber. At the same time, the protection OLT MUST send a notification message for the protection PW with the Preferential Forwarding status bit of active to the remote PE, so that traffic can be switched to the protection PW.

6.3. Protection procedure upon the working OLT failure

As depicted in Fig. 2, a service is provisioned with a working PW and a protection PW, both PW terminated on PE1. If PE1 lost its

connection to the working OLT, it SHOULD send a LDP notification message on the protection PW with the Request Switchover bit set.

Upon receiving a LDP notification message from its remote PE with the Request Switchover bit set, a protection OLT MUST activate its optical interface to the protection trunk fiber and activate the associated protection PW, so that traffic can be reliably switched to the protection trunk PON link and the protection PW.

In the case of Fig.3, PW-RED State TLV [ICCP] can be used by PE1 to notify PE2 the faults in all the scenarios, and PE2 operates the same as described in Section 5.1 to 5.3.

7. Security Considerations

Security considerations as described in [ICCP] apply.

8. IANA Considerations

These values are requested from the registry of "ICC RG parameter type":

0x00X0	PON Connect TLV
0x00X1	PON Disconnect TLV
0x00X2	PON Configuration TLV
0x00X3	PON State TLV
0x00X4	PON ONU Database Sync TLV

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

[RFC6870] Muley, P., Aissaoui, M., "Pseudowire Preferential Forwarding Status Bit", RFC 6870, February 2013

[ICCP] Martini, L. and et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", RFC 7275, June 2014

9.2. Informative References

[RFC6456] Li, H., Zheng, R., and Farrel, A., "Multi-Segment Pseudowires in Passive Optical Networks", RFC 6456, November 2011

[SEAMLESS] Leymann, N., and et al, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-04, Work in progress

[G983.1] ITU-T, "Broadband optical access systems based on Passive Optical Networks (PON)", ITU-T G.983.1, January, 2005

[IEEE-1904.1] IEEE Std. 1904.1, "Standard for Service Interoperability in Ethernet Passive Optical Networks (SIEPON)", IEEE Computer Society, June, 2013

[IEEE-802] IEEE Std. 802, "IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture", IEEE Computer Society, December, 2001 with amendments

[TR-221] BBF TR-221, "Technical Specifications for MPLS in Mobile Backhaul Networks", the Broadband Forum, October, 2011

[SD-331] BBF SD-331, "Architecture and Technical Requirements for PON-Based Mobile Backhaul Networks", the Broadband Forum, Work in progress

10. Acknowledgments

The authors would like to thank Min Ye, Hongyu Li, Wei Lin, Xifeng Wan, Yannick Legoff and Shrinivas Joshi for their valuable discussions and comments.

Authors' Addresses

Yuanlong Jiang
Huawei Technologies Co., Ltd.
Bantian, Longgang district
Shenzhen 518129, China
Email: jiangyuanlong@huawei.com

Yong Luo
Huawei Technologies Co., Ltd.
Bantian, Longgang district
Shenzhen 518129, China
Email: dennis.luoyong@huawei.com

Edwin Mallette
Bright House Networks
4145 S. Falkenburg Road
Tampa, FL 33578 USA
Email: edwin.mallette@gmail.com

Chengbin Shen
China Telecom
Email: shencb@sttri.com.cn

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886, USA
Email: yshen@juniper.net

Weiqiang Cheng
China Mobile
No.32 Xuanwumen West Street
Beijing 100053, China
Email: chengweiqiang@chinamobile.com

Guangtao Zhou
China Unicom
No.9 Shouti South Road
Beijing 100048, China
Email: zhouguangtao@chinaunicom.cn