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.

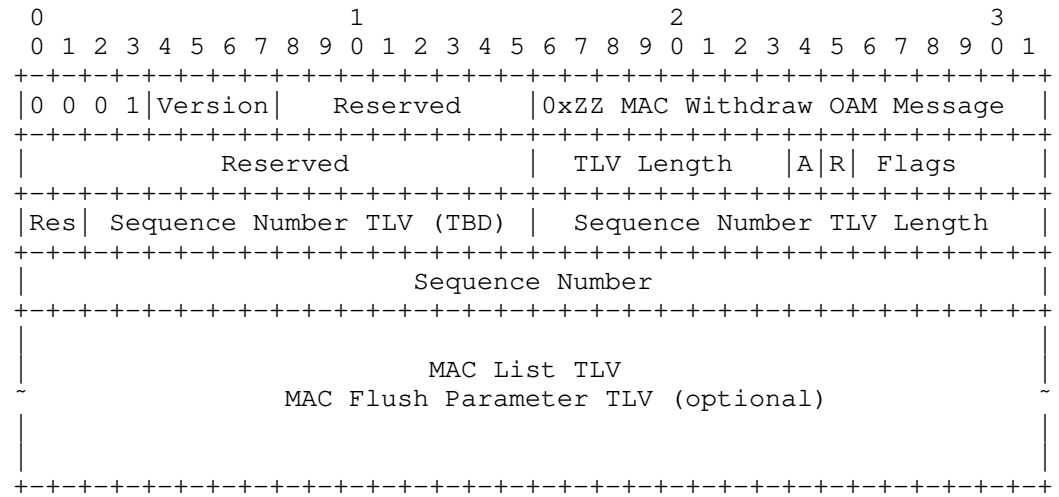


Figure 1: MAC Address Withdraw PW OAM Packet Format

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: Informational

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

Cisco Systems, Inc.

Expires: August 26, 2011

February 26, 2011

Stitching Procedures for Static PW in MPLS-TP Environment
draft-boutros-pwe3-mpls-tp-ms-pw-01.txt

Status of this Memo

Abstract

The existing procedures for concatenating static and dynamic pseudowires (PWs) do not take into account the PW status Operation, Administration, and Maintenance (OAM) messages defined for static PW. Also, these procedures do not take into account operator functions such Lock Instruct and Loopback introduced as part of MPLS Transport Profile (MPLS-TP). This informational document reiterates stitching procedures for static PW taking into account all the new proposed extensions.

This document is a product of a joint Internet Engineering Task Force(IETF) / International Telecommunication Union Telecommunication Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and PWE3 architectures to support the capabilities and functionalities of a packet transport network.

Table of Contents

1. Introduction

The PWE3 Architecture in [1] defines signaling and encapsulation techniques for establishing Single Segment PW (SS-PW) between a pair of terminating PEs. Procedures for stitching two or more static or dynamic SS-PWs to form Multi-Segment PW (MS-PW) are described in [2].

These procedures make use of PW status messages carried in LDP TLV over dynamic PW established via LDP. [3] defines a new PW status OAM message used to carry PW status in-band over static PW. This message makes it possible to exchange PW status end-to-end over a MS-PW consisting of one or more static PW.

[5] specifies operator new Operation, Administration, and Maintenance (OAM) functions Lock Instruct (LI) and Loopback (LB) for associated bi-directional circuits such as MPLS-TP LSP, SS-PW, and MS-PW in an MPLS Transport Profile (MPLS-TP) environment. These functions enable network operators to lock a circuit (LSP and PW) and operate it in loopback mode for testing/management purpose.

This informational document describes the application of the existing PW stitching procedures taking into consideration LI, LB, as well as PW status OAM messages.

This document is a product of a joint Internet Engineering Task Force (IETF) / International Telecommunication Union Telecommunication Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and PWE3 architectures to support the capabilities and functionalities of a packet transport network.

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 [1].

2. Terminology

LDP: Label Distribution Protocol.

MEP: Maintenance End Point.

MIP: Maintenance Intermediate Point.

MPLS: Multi Protocol Label Switching.

MPLS-TP: MPLS Transport Profile.

MS-PW: Multi-Segment PseudoWire.

LB: Loopback.

LI: Lock Instruct.

LSP: Label Switched Path.

OAM: MPLS Operations, Administration and Maintenance.

PE: Provide Edge Node.

PW: PseudoWire.

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

SS-PW: Single-Segment PseudoWire.

TLV: Type, Length, and Value.

T-PE: Terminating Provider Edge Node of a MS-PW.

3. Operation

In this section, we explain the use of LI/LB mechanisms referring to the MS-PW model shown in Figure 1. The SS-PW segments PW1 and PW2 can be either static or dynamic. We assume that PWs are carried over MPLS-TP LSPs (transport LSPs) so that LI/LB mechanisms can be applied at the transport LSP level, as well we consider the application of LI/LB at PW level.

PW status is sent via LDP message and PW OAM message respectively over dynamic and static PW segments. Note that even though only two PW segments are considered in the examples below, the described procedures are applicable to MS-PWs with more than two segments.

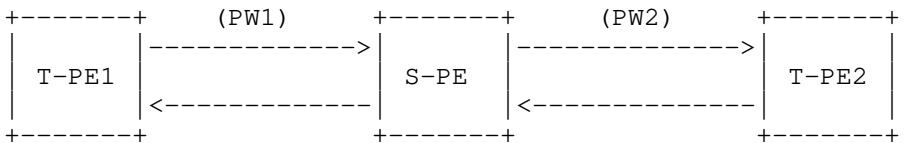


Figure 1. Reference Model for LI/LB Mechanism

3.1. Lock Operation

3.1.1. Locking MPLS-TP LSP

An MPLS-TP LSP can be taken out of service for maintenance operation using the LI mechanism described in [5]. LI messages are exchanged between MPLS-TP Maintenance End Points (MEPs). In the case of MS-PW, each MPLS-TP LSP associated with a given PW segment can be individually locked for management purpose. This means that, in a MS-PW scenario, a T-PE is always a MEP and an S-PE is a MEP for an MPLS-TP LSP carrying PW segments. Furthermore, a T-PE (MEP) assumes that an MPLS-TP LSP is successfully locked only when the corresponding LI reply is received from the other intended receiver MEP (other T-PE or S-PE).

3.1.1.1. LI originated at T-PE

Assume that T-PE1 originates an LI request for the MPLS-TP LSP carrying PW1. The intended recipient of the message will be the S-PE. When T-PE1 receives a positive LI reply from the S-PE, it assumes that the MPLS-TP LSP is successfully locked, and takes PW1 and all other PWs associated with the MPLS-TP LSP out of service. This means that PW1 and all other impacted PWs will no longer carry user data.

When S-PE receives an LI request, if the intended MPLS-TP LSP can be locked, the S-PE finds all PWs associated with this MPLS-TP LSP and first sends the PW status code 0x00000018 (Local PSN-facing PW Receive/Transmit Faults) on all stitched PWs segments to T-PE2. PW status code is sent over PW OAM message or LDP message depending on whether the segment PW2 is static or dynamic. After sending the PW status code to T-PE2, S-PE lock the MPLS-TP LSP and sends a positive LI reply to T-PE1. If the MPLS-TP LSP cannot be locked, S-PE sends a negative LI reply with the appropriate error code to T-PE1.

When T-PE2 receives the PW status codes, it processes them as described in [3] or [4] depending on whether PW2 is dynamic or static.

If PW2 is a dynamic segment and does not support PW status, S-PE needs to withdraw its labels from T-PE2 before locking the MPLS LSP.

For better scalability, S-PE may use the notion of group ID described in [6] to send PW status or withdraw labels all impacted dynamic PWs between itself and T-PE2. Use of group ID with PW status OAM over static PW is TBD.

3.1.1.2. LI originated at S-PE

Let's assume that an operator wants to originate an LI request at S-PE for the MPLS-TP LSP carrying PW1. The intended recipient of the LI request is T-PE1. First, S-PE sends PW status code 0x00000018 (Local PSN-facing PW Receive/Transmit Fault) for PW1 as well as all other PWs pinned down to MPLS-TP LSP in question to T-PE1 and PW2 and all other stitched PWs other segments to T-PE2. PW status code is sent over PW OAM message or LDP message depending on whether the segment PW2 is static or dynamic. When T-PE2 receives the PW status codes, it processes them as described in [3] or [4] respectively depending on whether PW2 is dynamic or static. It then sends LI request message to T-PE1. If T-PE1 can successfully lock the MPLS LSP, it sends a positive LI response. Upon receiving the response, S- PE1 assumes that the MPLS-TP LSP is locked, and PW1 is no longer used for carrying regular user data.

If T-PE1 is unable to lock the MPLS-TP LSP, it sends a negative LI response with the appropriate error code. In this case, S-PE sends PW status 0x00000000 to T-PE1 and T-PE2 so that services on PW1 and PW2 and all other PWs associated with the MPLS-TP LSP in question can resume.

If PW2 is a dynamic segment and PW status, S-PE needs to withdraw its labels from T-PE1 and T-PE2 before sending LI request to T-PE1.

For better scalability, S-PE may use the notion of group ID described in [6] to send PW status or withdraw labels all impacted dynamic PWs.

Use of group ID with PW status OAM over static PW is TBD.

3.1.2. Locking PW

A given PW can also be taken out of service for maintenance operation without impacting services over other PWs using the LI mechanism described in [5].

3.1.2.1. LI originated at T-PE

In our example, let's assume that, T-PE1 sends an LI request message to lock PW1. S-PE is the intended recipient (based on the TTL value of the PW label).If S-PE is able to lock PW1, it sends a PW status message with the status code 0x00000018 (Local PSN-facing PW Receive/Transmit Fault) over PW2 to T-PE2, and locks PW1. S-PE then sends a positive LI reply to T-PE1. Upon receiving the positive LI

reply, T-PE locks PW1. If S-PE is unable to lock PW1, it sends a negative LI reply to T-PE1. PW status code is sent over PW OAM message or LDP message depending on whether the segment PW2 is static or dynamic. When T-PE2 receives the PW status codes, it processes them as described in [3] or [4] depending on whether PW2 is dynamic or static.

.

3.2. Loopback Operation

3.2.1. As described in [5], an MPLS-TP LSP or a PW can be setup to in loopback mode for management purpose, e.g., to test or verify connectivity of the LSP/PW up to a specific node on the path of the MPLS-TP tunnel/PW, and to test the LSP/PW performance with respect to delay/jitter, etc. But, prior to operating in loopback mode, an MPLS-TP LSP or PW must be successfully locked. Loopback at MPLS-TP LSP Level

Assume that an operator wants to operate an MPLS-TP LSP between T-PE1 and S-PE carrying PW1 in loopback mode such that S-PE loops all the incoming packets over the MPLS-TP LSP back to the sender (in this case T-PE1).

T-PE1 sends an LB request message which is received by S-PE. S-PE can setup the MPLS-TP LSP only if all the PWs carried over that LSP can be setup in loopback mode. If S-PE can setup the MPLS-TP tunnel in loopback mode, it sends a positive LB response. Otherwise, it sends a negative LB response to T-PE1.

If the MPLS-TP LSP is successfully setup in loopback mode, all incoming packets over PW1 will be looped back to T-PE1. This is also true for any other PW(s) between T-PE1 and S-PE pinned down to the MPLS-TP LSP in question.

Similarly, MPLS-TP LSP between S-PE and T-PE1 can be operated in loopback mode such that T-PE1 loops all incoming packets over the LSP back to S-PE. In this case, S-PE and T-PE1 respectively are sender and receiver of the LB request message.

3.2.2. Loopback at PW Level

A SS-PW or MS-PW can be operated in loopback mode.

In our example, let’s assume that PW1 is to be operated in a loopback mode such that S-PE loops all incoming packets over PW1 back to T-PE1. To setup this mode of operation, T-PE1 sends an LB request

message to S-PE. TTL value of the PW label is chosen so as to expire on the intended recipient (in our example TTL value should be 1 so that LB request can be processed at S-PE). If S-PE can successfully setup PW1 in loopback mode, it sends a positive LB response to T-PE1.

If loopback operation over the entire MS-PW (i.e., over PW1 and PW2) such that T-PE2 loops all the incoming packets over PW2 back to T-PE1, T-PE1 and T-PE2 will be the sender and receiver of LB message.

3.3. Switching Point PE TLV

Switching Point PE TLV (S-PE TLV) is used to record information about S-PE(s) that a PW traverses. An S-PE TLV contains many sub-TLVs as described in [3]. One such sub-TLV carries the FEC of the last traversed PW segment.

In the case of MS-PW containing static PW segment(s), if the last traversed PW segment is statically provisioned, a new sub-TLV containing the FEC defined for static PW in [7] can be used to represent the last traversed PW segment. The new sub-TLV type will be defined in [4].

3.4. LSP-Ping/Trace

TBD

4. Security Considerations

This document does not introduce any additional security constraints.

5. IANA Considerations

Not applicable.

6. References

6.1. Normative References

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

6.2. Informative References

[2] Stewart Bryant, et. al, "Pseudowire Emulation Edge-to-Edge (PWE3) Architecture", RFC3985, March 2005.

[3] Luca Martini, et. al, "Segmented Pseudowire", draft-ietf-pwe3-segmented-pw-15.txt (work in progress), June 2010.

[4] Luca Martini, et. al, "Pseudowire Status for Static Pseudowires", draft-ietf-pwe3-static-pw-status-00.txt (work in progress), February 2010.

[5] Sami Boutros, et. al, "MPLS Transport Profile Lock Instruct and Loopback Functions", draft-ietf-mpls-tp-li-lb-00.txt (work in progress), June 2010.

[6] Luca Martini, et. al, "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC4447, April 2006.

[7] Nitin Bahadur, et. al, "LSP-Ping extensions for MPLS-TP", draft-ietf-mpls-tp-lsp-ping-extensions-01.txt (work in progress), February 2010.

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

Full Copyright Statement

Intellectual Property Statement

Acknowledgment

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

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.

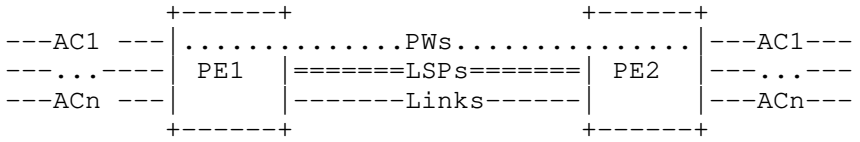


Figure 1: Explicit PW-to-LSP binding scenario

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.

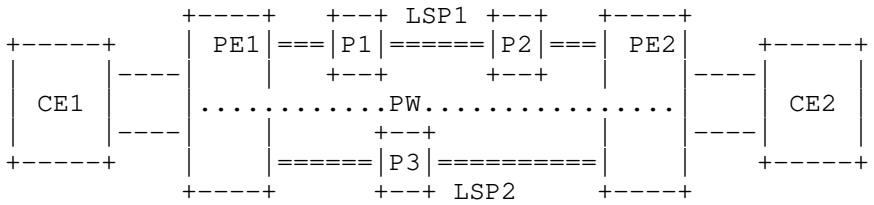


Figure 2: Inconsistent SS-PW to LSP binding scenario

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:

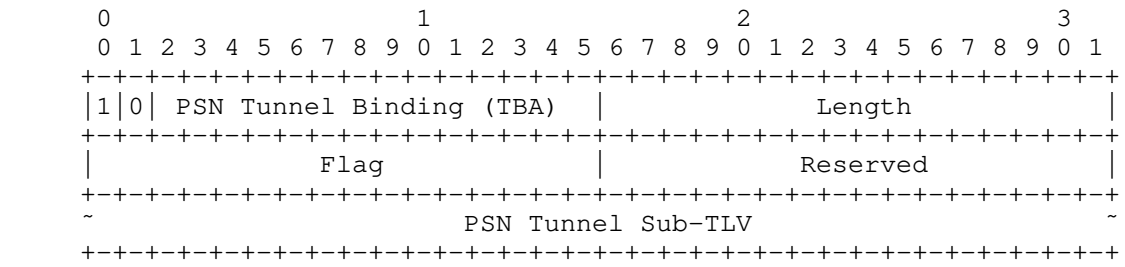
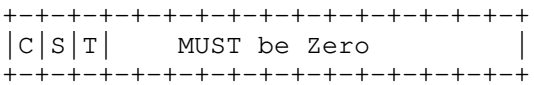


Figure 3: PSN Tunnel Binding TLV

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.

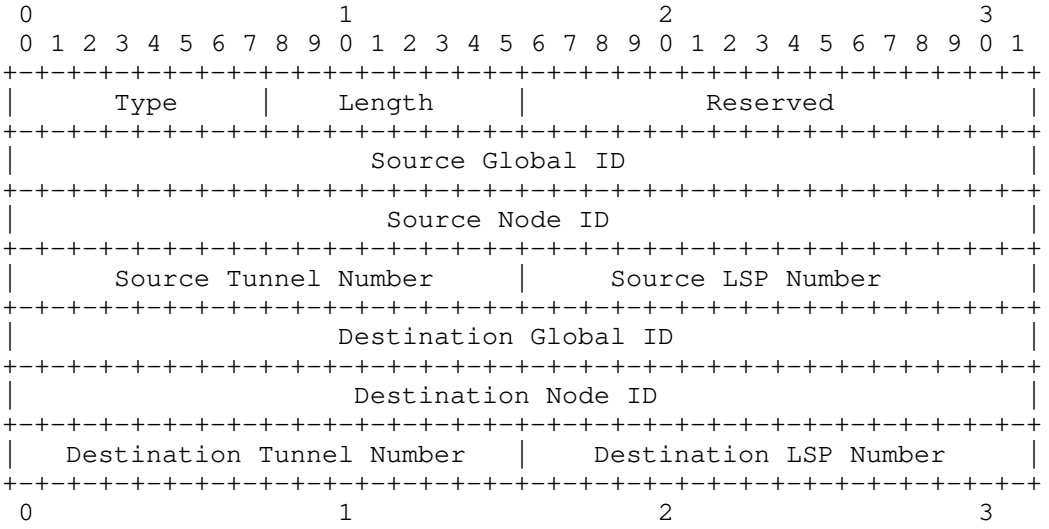


Figure 4: IPv4 PSN Tunnel sub-TLV format

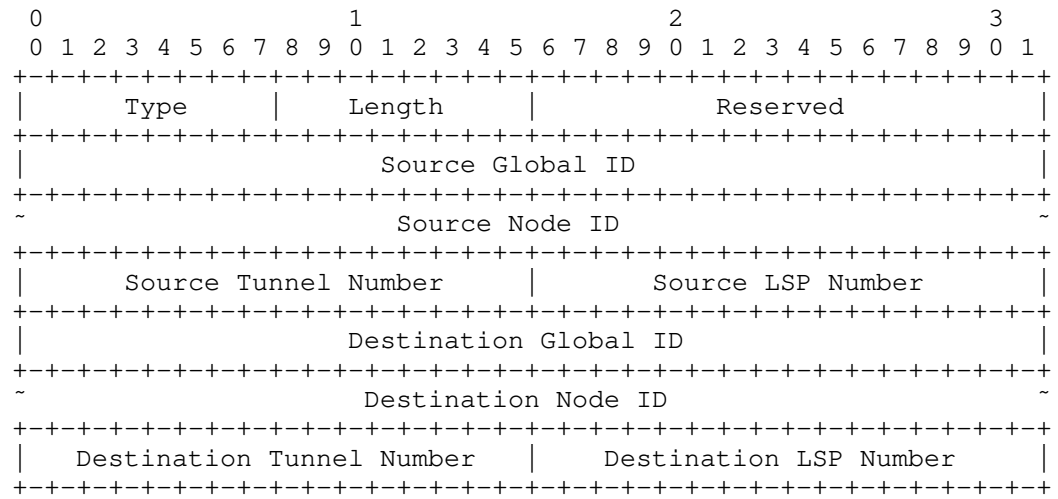


Figure 5: IPv6 PSN Tunnel sub-TLV format

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.

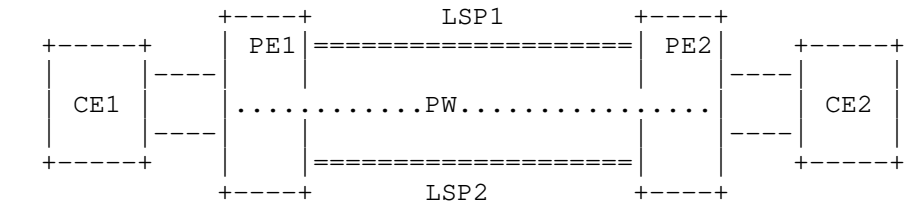


Figure 6: PSN binding operation in SS-PW environment

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.

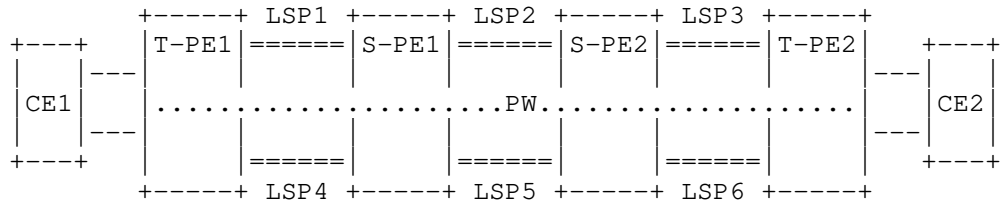


Figure 7: PSN binding operation in MS-PW environment

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
Intended status: Informational
Expires: September 8, 2011

N. Del Regno, Ed.
Verizon Communications Inc

March 7, 2011

The Pseudowire (PW) & Virtual Circuit Connectivity Verification (VCCV) Implementation Survey Results
draft-delregno-pw-vccv-impl-survey-results-00

Abstract

Most Pseudowire Emulation Edge-to-Edge (PWE3) encapsulations mandate the use of the Control Word (CW) in order to better emulate the services for which the encapsulations have been defined. However, some encapulations treat the Control Word as optional. As a result, implementations of the CW, for encapsulations for which it is optional, vary by equipment manufacturer, equipment model and service provider network. Similarly, Virtual Circuit Connectivity Verification (VCCV) supports three Control Channel (CC) types and multiple Connectivity Verification (CV) Types. This flexibility has led to reports of interoperability issues within deployed networks and associated drafts to attempt to remedy the situation. This survey of the PW/VCCV user community was conducted to determine implementation trends. The survey and results is presented herein.

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

Table of Contents

1. Introduction

The PWE3 working group has defined many encapsulations of various Layer 1 and Layer 2 links. Within these encapsulations, there are often several modes of encapsulation which have differing requirements in order to fully emulate the service. As such, the use of the PWE3 Control Word is mandated in many of the encapsulations, but not all. This can present interoperability issues related to A) Control Word use and B) VCCV Control Channel negotiation in mixed implementation environments.

The encapsulations and modes for which the Control Word is currently optional are:

- Ethernet Tagged Mode
- Ethernet Raw Mode
- PPP
- HDLC
- Frame Relay Port Mode
- ATM (N:1 Cell Mode)

[RFC5085] defines three Control Channel types for MPLS PW's: Type 1, using the Pseudowire Control Word, Type 2, using the Router Alert Label, and Type 3, using TTL Expiration (e.g. MPLS PW Label with TTL == 1). While Type 2 (RA Label) is indicated as being "the preferred mode of VCCV operation when the Control Word is not present," RFC 5085 does not indicate a mandatory Control Channel to ensure interoperable implementations. The closest it comes to mandating a control channel is the requirement to support Type 1 (Control Word) whenever the control word is present. As such, the three options yield seven implementation permutations (assuming you have to support at least one Control Channel type to provide VCCV). Due to these permuations, interoperability challenges have been identified by several VCCV users.

In order to assess the best approach to address the observed interoperability issues, the PWE3 working group decided to solicit feedback from the PW and VCCV user community regarding implementation. This document presents the survey and the information returned by the user community who participated.

1.1. PW/VCCV Survey Overview

Per the direction of the PWE3 Working Group chairs, a survey was created to sample the nature of implementations of Pseudowires, with specific emphasis on Control Word usage, and VCCV, with emphasis on Control Channel and Control Type usage. The survey consisted of a series of questions based on direction of the WG chairs and the survey opened to the public on November 4, 2010. The URL for the survey (now closed) was http://www.surveymonkey.com/pwe3/. The survey ran from November 4, 2010 until February 25, 2011.

1.2. PW/VCCV Survey Form

The PW/VCCV Implementation Survey requested the following information about user implementations:

- Responding Organziation. No provisions were made for anonymity. All responses required a valid email address in order to validate the survey response.

- Of the various encapsulations (and options therein) known at the time, including the WG draft for Fiber Channel), which were implemented b the respondent. These included:

- Ethernet Tagged Mode - RFC 4448
- Ethernet Raw Mode - RFC 4448
- SAToP - RFC 4553
- PPP - RFC 4618
- HDLC - RFC 4618
- Frame Relay (Port Mode) - RFC 4619
- Frame Relay (1:1 Mode) - RFC 4619
- ATM (N:1 Mode) - RFC 4717
- ATM (1:1 Mode) - RFC 4717
- ATM (AAL5 SDU Mode) - RFC 4717
- ATM (AAL5 PDU Mode) - RFC 4717
- CEP - RFC 4842

- CESoPSN - RFC 5086
- TDMoIP - RFC 5087
- Fiber Channel (Port Mode) - draft-ietf-pwe3-fc-encap

- Approximately how many Pseudowires of each type were deployed. Respondents could list a number, or for the sake of privacy, could just respond "In-Use" instead.

- For each encapsulation listed above, the respondent could indicated which Control Channel was in use. The options listed were:

- Control Word (Type 1)
- Router Alert Label (Type 2)
- TTL Expiry (Type 3)

- For each encapsulation listed above, the respondent could indicate which Connectivity Verification types were in use. The options were:

- ICMP Ping
- LSP Ping

- For each encapsulation type for which the use of the Control Word is optional, the respondents could indicated the encaps for which Control Word was supported by the equipment used and whether it was in use in the network. The encaps listed were:

- Ethernet (Tagged Mode)
- Ethernet (Raw Mode)
- PPP
- HDLC
- Frame Relay (Port Mode)
- ATM (N:1 Cell Mode)

- Finally, a freeform entry was provided for the respondent to provide feedback regarding PW and VCCV deployments, VCCV interoperability challenges, the survey or any network/vendor details they wished to share.

1.3. PW/VCCV Survey Highlights

There were 17 valid responses to the survey. The following companies responded.

2. Survey Results

2.1. Respondents

The following companies participated in the PW/VCCV Implementation Survey. The data provided has been aggregated. No specific company's reponse will be detailed herein.

- Time Warner Cable
- Bright House Networks
- Tinet
- AboveNet
- Telecom New Zealand
- Cox Communications
- MTN South Africa
- Wipro Technologies
- Verizon
- AMS-IX
- Superonline
- Deutsche Telekom AG
- Internet Solution
- Easynet Global Services
- Telstra Corporation
- OJSC MegaFon
- France Telecom Orange

2.2. Pseudowire Encapsulations Implemented

The following question was asked: "In your network in general, across all products, please indicate which Pseudowire encapsulations your company has implemented." Of all responses, the following list shows the percentage of responses for each encapsulation:

- Ethernet Tagged Mode - RFC 4448 = 77.8%
- Ethernet Raw Mode - RFC 4448 = 77.8%
- SAToP - RFC 4553 = 11.1%
- PPP - RFC 4618 = 11.1%
- HDLC - RFC 4618 = 5.6%
- Frame Relay (Port Mode) - RFC 4619 = 16.7%
- Frame Relay (1:1 Mode) - RFC 4619 = 44.4%
- ATM (N:1 Mode) - RFC 4717 = 5.6%
- ATM (1:1 Mode) - RFC 4717 = 22.2%
- ATM (AAL5 SDU Mode) - RFC 4717 = 5.6%
- ATM (AAL5 PDU Mode) - RFC 4717 = 0.0%
- CEP - RFC 4842 = 0.0%
- CESoPSN - RFC 5086 = 11.1%
- TDMoIP - RFC 5087 = 11.1%
- Fiber Channel (Port Mode) - draft-ietf-pwe3-fc-encap = 5.6%

2.3. Number of Pseudowires Deployed

The following question was asked: "Approximately how many Pseudowires are deployed of each encapsulation type. Note, this should be the number of pseudowires in service, carrying traffic, or pre-positioned to do so." The following list shows the number of psudowires in use for each encapsulation:

- Ethernet Tagged Mode = 93,861

- Ethernet Raw Mode = 94,231
- SAToP - RFC 4553 = 20,050
- PPP - RFC 4618 = 500
- HDLC - RFC 4618 = 0
- Frame Relay (Port Mode) - RFC 4619 = 5,002
- Frame Relay (1:1 Mode) - RFC 4619 = 50,959
- ATM (N:1 Mode) - RFC 4717 = 50,000
- ATM (1:1 Mode) - RFC 4717 = 70,103
- ATM (AAL5 SDU Mode) - RFC 4717 = 0
- ATM (AAL5 PDU Mode) - RFC 4717 = 0
- CEP - RFC 4842 = 0
- CESoPSN - RFC 5086 = 21,600
- TDMoIP - RFC 5087 = 20,000
- Fiber Channel (Port Mode) - draft-ietf-pwe3-fc-encap = 0

In the above responses, on several occasions the response was in the form of "> XXXXX" where the response indicated a number greater than the one provided. Where applicable, the number itself was used in the sums above. For example, ">20K" and "20K+" yielded 20K.

Additionally, the following encaps were listed as "In-Use" with no quantity provided:

- Ethernet Raw Mode: 2 Responses
- ATM (AAL5 SDU Mode): 1 Response
- TDMoIP: 1 Response

2.4. VCCV Control Channel In Use

The following instructions were given: "Please indicate which VCCV Control Channel is used for each encapsulation type. Understanding that users may have different networks with varying implementations, for your network in general, please select all which apply." The

numbers below indicate the number of responses. The responses were:

- Ethernet Tagged Mode - RFC 4448
  - Control Word (Type 1) = 7
  - Router Alert Label (Type 2) = 3
  - TTL Expiry (Type 3) = 3
- Ethernet Raw Mode - RFC 4448
  - Control Word (Type 1) = 8
  - Router Alert Label (Type 2) = 4
  - TTL Expiry (Type 3) = 4
- SAToP - RFC 4553
  - Control Word (Type 1) = 1
  - Router Alert Label (Type 2) = 0
  - TTL Expiry (Type 3) = 0
- PPP - RFC 4618
  - Control Word (Type 1) = 0
  - Router Alert Label (Type 2) = 0
  - TTL Expiry (Type 3) = 0
- HDLC - RFC 4618
  - Control Word (Type 1) = 0
  - Router Alert Label (Type 2) = 0
  - TTL Expiry (Type 3) = 0
- Frame Relay (Port Mode) - RFC 4619
  - Control Word (Type 1) = 1
  - Router Alert Label (Type 2) = 0

    * TTL Expiry (Type 3) = 0

  - Frame Relay (1:1 Mode) - RFC 4619
    * Control Word (Type 1) = 3
    * Router Alert Label (Type 2) = 0
    * TTL Expiry (Type 3) = 2
  - ATM (N:1 Mode) - RFC 4717
    * Control Word (Type 1) = 1
    * Router Alert Label (Type 2) = 0
    * TTL Expiry (Type 3) = 0
  - ATM (1:1 Mode) - RFC 4717
    * Control Word (Type 1) = 1
    * Router Alert Label (Type 2) = 0
    * TTL Expiry (Type 3) = 1
  - ATM (AAL5 SDU Mode) - RFC 4717
    * Control Word (Type 1) = 0
    * Router Alert Label (Type 2) = 1
    * TTL Expiry (Type 3) = 0
  - ATM (AAL5 PDU Mode) - RFC 4717
    * Control Word (Type 1) = 0
    * Router Alert Label (Type 2) = 0
    * TTL Expiry (Type 3) = 0
  - CEP - RFC 4842
    * Control Word (Type 1) = 0
    * Router Alert Label (Type 2) = 0

    * TTL Expiry (Type 3) = 0

  o CESoPSN - RFC 5086

    * Control Word (Type 1) = 0

    * Router Alert Label (Type 2) = 0

    * TTL Expiry (Type 3) = 1

  o TDMoIP - RFC 5087

    * Control Word (Type 1) = 0

    * Router Alert Label (Type 2) = 0

    * TTL Expiry (Type 3) = 0

  o Fiber Channel (Port Mode) - draft-ietf-pwe3-fc-encap

    * Control Word (Type 1) = 0

    * Router Alert Label (Type 2) = 0

    * TTL Expiry (Type 3) = 0

2.5. VCCV Connectivity Verification Types In Use

The following instructions were given: "Please indicate which VCCV Connectivity Verification types are used in your networks for each encapsulation type." Note that BFD was not one of the choices. The responses were as follows:

  o Ethernet Tagged Mode - RFC 4448

    * ICMP Ping = 5

    * LSP Ping = 11

  o Ethernet Raw Mode - RFC 4448

    * ICMP Ping = 6

    * LSP Ping = 11

  o SAToP - RFC 4553

    * ICMP Ping = 0

    * LSP Ping = 2

  o PPP - RFC 4618

    * ICMP Ping = 0

    * LSP Ping = 0

  o HDLC - RFC 4618

    * ICMP Ping = 0

    * LSP Ping = 0

  o Frame Relay (Port Mode) - RFC 4619

    * ICMP Ping = 0

    * LSP Ping = 1

  o Frame Relay (1:1 Mode) - RFC 4619

    * ICMP Ping = 2

    * LSP Ping = 5

  o ATM (N:1 Mode) - RFC 4717

    * ICMP Ping = 0

    * LSP Ping = 1

  o ATM (1:1 Mode) - RFC 4717

    * ICMP Ping = 0

    * LSP Ping = 3

  o ATM (AAL5 SDU Mode) - RFC 4717

    * ICMP Ping = 0

    * LSP Ping = 1

- ATM (AAL5 PDU Mode) - RFC 4717
  - ICMP Ping = 0
  - LSP Ping = 0
- CEP - RFC 4842
  - ICMP Ping = 0
  - LSP Ping = 0
- CESoPSN - RFC 5086
  - ICMP Ping = 0
  - LSP Ping = 1
- TDMoIP - RFC 5087
  - ICMP Ping = 0
  - LSP Ping = 1
- Fiber Channel (Port Mode) - draft-ietf-pwe3-fc-encap
  - ICMP Ping = 0
  - LSP Ping = 0

2.6. Control Word Support for Encaps for which CW is Optional

The following instructions were given: "Please indicate your network's support of and use of the Control Word for encapsulations for which the Control Word is optional." The responses were:

- Ethernet (Tagged Mode)
  - Supported by Network/Equipment = 13
  - Used in Network = 6
- Ethernet (Raw Mode)
  - Supported by Network/Equipment = 14
  - Used in Network = 7

- PPP
  - Supported by Network/Equipment = 5
  - Used in Network = 0
- HDLC
  - Supported by Network/Equipment = 4
  - Used in Network = 0
- Frame Relay (Port Mode)
  - Supported by Network/Equipment = 3
  - Used in Network = 1
- ATM (N:1 Cell Mode)
  - Supported by Network/Equipment = 5
  - Used in Network = 1

2.7. Open Ended Question

Space was provided for user feedback. The following instructions were given: "Please use this space to provide any feedback regarding PW and VCCV deployments, VCCV interoperability challenges, this survey or any network/vendor details you wish to share." Below are the responses, made anonymous.

1. BFD VCCV Control Channel is not indicated in the survey (may be required for PW redundancy purpose)
2. Using CV is not required at the moment
3. COMPANY has deployed several MPLS network elements, from multiple vendors. COMPANY is seeking a uniform implementation of VCCV Control Channel (CC) capabilities across its various vendor platforms. This will provide COMPANY with significant advantages in reduced operational overheads when handling cross-domain faults. Having a uniform VCCV feature implementation in COMPANY multi-vendor network leads to: o Reduced operational cost and complexity o Reduced OSS development to coordinate incompatible VCCV implementations. o Increased end-end service availability when handing faults. In addition, currently some of COMPANY deployed VCCV traffic flows (on some vendor platforms) are not

guaranteed to follow those of the customer's application traffic (a key operational requirement). As a result, the response from the circuit ping cannot faithfully reflect the status of the circuit. This leads to ambiguity regarding the operational status of our networks. An in-band method is highly preferred, with COMPANY having a clear preference for VCCV Circuit Ping using PWE Control Word. This preference is being pursued with each of COMPANY vendors.

4. PW VCCV is very useful tool for finding faults in each PW channel. Without this we can not find fault on a PW channel. PW VCCV using BFD is another better option. Intropерbility challences are with Ethernet OAM mechanism.

5. We are using L2PVPN AToM like-to-like models - ATMoMPLS - EoMPLS ATMoMPLS : This service offered for transporting ATM cells over IP/MPLS core with Edge ATM CE devices including BPX, Ericsson Media Gateway etc. This is purely a Port mode with cell-packing configuration on it to have best performance. QoS marking is done for getting LLQ treatment in the core for these MPLS encapsulated ATM packets. EoMPLS: This service offered for transporting 2G/3G traffic from network such as Node-B to RNC's over IP/MPLS backbone core network. QoS marking is done for getting guaranteed bandwidth treatment in the core for these MPLS encapsulated ATM packets. In addition to basic L2VPN service configuration, these traffic are routed via MPLS TE tunnels with dedicated path and bandwidth defined to avoid bandwidth related congestion.

6. EQUIPMENT MANUFACTURER does not provide options to configure VCCV control-channel and its sub options for LDP based L2Circuits. How can we achieve end-to-end management and fault detection of PW without VCCV in such cases?

7. I'm very interested in this work as we continue to experience interop challenges particularly with newer vendors to the space who are only implementing VCCV via control word. Vendors who have tailed their MPLS OAM set specifically to the cell backhaul space and mandatory CW have been known to fall into this space. That's all I've got.

3. Security Considerations

As this document is a report of the PW/VCCV User Implementation Survey results, no security considerations are introduced.

4. Acknowledgements

I would like to thank the chairs of the PWE3 Working Group for their guidance and review of the Survey questions. I would also like to sincerly thank those who took the time and effort to participate.

5. References

5.1. Normative References

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

5.2. Informative References

[RFC5085] Nadeau, T., Ed. and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", December 2007.

Author's Address

Christopher N. "Nick" Del Regno (editor)
Verizon Communications Inc
400 International Pkwy
Richardson, TX 75081
US

Email: nick.delregno@verizon.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2011

N. Del Regno, Ed.
Verizon Communications
T. Nadeau
Huawei
V. Manral
IP Infusion
D. Ward
Juniper Networks
October 15, 2010

Mandatory Use of Control Word for PWE3 Encapsulations
draft-delregno-pwe3-mandatory-control-word-00

Abstract

Of the many variations of PWE3 Encapsulations and Modes (e.g. Ethernet, Port Mode, VLAN Mode, etc), only five have the Control Word (CW) as being optional. As a result, this causes an issue with VCCV Control Channel selection. This draft endeavors to resolve the issue going forward by making the Control Word, and subsequently the CW-based VCCV Control Channel, mandatory for all PWE3 Encapsulations.

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 will expire on April 18, 2011.

Copyright Notice

Table of Contents

1. Introduction

The PWE3 working group has defined many encapsulations of various Layer 1 and Layer 2 links. Within these encapsulations, there are often several modes of encapsulation which have differing requirements in order to fully emulate the service. As such, the use of the PWE3 Control Word is mandated in many of the encapsulations, but not all. This can present interoperability issues related to A) Control Word use and B) VCCV Control Channel negotiation in mixed implementation environments.

In the various encapsulations where the Control Word is optional, the language from [RFC4385] is consistently referenced: "The features that the control word provides may not be needed for a given PW. For example, ECMP may not be present or active on a given MPLS network, strict frame sequencing may not be required, etc. If this is the case, the control word provides little value and is therefore optional." As such, early implementations may not have supported the Control Word for those encapsulations which didn't require it. However, as recent discussions have shown [CBIT], the lack of the Control Word opens up other issues related to control-word negotiation (e.g. preferred vs. not- preferred) and VCCV Contol Channel negotiation and selection [DEL].

The encapsulations and modes for which the Control Word is currently optional are:

- Ethernet Tagged Mode
- Ethernet Raw Mode
- PPP
- HDLC
- Frame Relay Port Mode
- ATM (N:1 Cell Mode)

While the encapsulation for PPP, HDLC and Frame Relay Port Mode are the same encap, the services which they emulate may have different requirements, and are therefore listed separately.

Unfortunately, some early implementations of PWE3 standard (and/or prestandard) encapsulations are limited in their support for Control Word for the above encapsulations due to A) hardware deficiencies, B) software deficiencies or C) a combination of the two. In other cases, deployed implementations support control word, but the service

provider has had no impetus to suffer the minor loss of overhead efficiency. However, this document asserts based on operational feedback of the PWE3 protocols in actual deployments, that the benefits of requiring a mandatory control word in the PWE3 standards outweigh the minor efficiencies gained when not using it.

One of the major benefits of consistent use of the Control Word pertains to the choice of the VCCV Control Channel. As identified in [DEL], Control Channel Type 1 is the only "in-band" PWE3 control channel. This provides the advantage of proper VCCV forwarding behavior in the presence of ECMP. Further, while the sequencing supported by the Control Word is not mandatory, the use of the Control Word enables the use of sequencing without forcing the renegotiation of the PW.

All increases in the amount of overhead used to provide service should be weighed versus their perceived gain, especially when that overhead is large in comparison to the data being carried. This is a common concern with the ATM N:1 encapsulation. In theory, if only a single cell is encapsulated per PSN packet, not only is the inherent overhead inacceptably large, the additon of 4 bytes only compounds the problem. However, in practice, the PDUs, or groups of PDUs, carried in encapsulations above, including ATM (N:1 Cell Mode), are sufficiently large that the additional 4-bytes of CW overhead represent a relatively minor increase in the total overhead

2. Mandatory Control Word

The Control Word SHALL be mandatory for all PWE3 encapsulations. The use of the sequence number remains OPTIONAL.

As a result of the Control Word being Mandatory, all implementations of the PWE3 encapsulations SHALL follow Section 6.1 of [RFC4447] wherein the "PWs MUST have c=1". This requirement SHALL remain until such time, if ever, RFC4447 is superceded and the support for Control Word negotiation is removed as a result of this mandate.

3. Backward Compatibility

This Control Word mandate will not support backward compatibility with implementations which cannot support Control Word. For those implementations, CW negotiation identified in [RFC4447] will result in the PW negotiation never completing since the end which cannot support CW will ignore the Label Mapping message with c=1. However, for those implementations which currently support Control Word, the Control Word mandate will be supported as long as CW is set to

PREFERRED and the subsequent c=1 is negotiated.

4. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

5. Security Considerations

This document specifies the mandatory behavior which must be supported by implementations of PWE3 encapsulations. As the Control Word is either already mandated by various encapsulations or is optional, this mandate does not introduce any security issues not already addressed by the encapsulation definitions, if any. Further, the mandate of Control Word use may improve the security of related protocol behaviors, such as VCCV Control Word (e.g. no need for Router Alert Label support).

6. Acknowledgements

7. Normative References

[CBIT] Jin, L., Key, R., Delord, S., Nadeau, T., and V. Manral, "Pseudowire Control Word Negotiation Mechanism Analysis and Update", October 2010.

[DEL] Del Regno, N., Manral, V., Kunze, R., Paul, M., and T. Nadeau, "Mandatory Features of Virtual Circuit Connectivity Verification Implementations", October 2010.

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

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", February 2006.

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

Authors' Addresses

Nick Del Regno (editor)
Verizon Communications

Phone:
Fax:
Email: nick.delregno@verizon.com
URI:

Thomas Nadeau
Huawei

Phone:
Fax:
Email: t.nadeau@lucidvision.com
URI:

Vishwas Manral
IP Infusion

Phone:
Fax:
Email: vishwas@ipinfusion.com
URI:

David Ward
Juniper Networks

Phone:
Fax:
Email: dward@juniper.net
URI:

Network Working Group
Internet-Draft
Intended status: BCP
Expires: April 18, 2011

N. Del Regno, Ed.
Verizon
V. Manral, Ed.
IPInfusion Inc.
R. Kunze
M. Paul
Deutsche Telekom
T. Nadeau
Huawei

October 15, 2010

Mandatory Features of Virtual Circuit Connectivity Verification Implementations
draft-delregno-pwe3-vccv-mandatory-features-02

Abstract

Pseudowire Virtual Circuit Connectivity Verification (VCCV) [RFC5085] defines several Control Channel (CC) Types for MPLS PW's , none of which are preferred or mandatory. As a result, independent implementations of different subsets of the three options have resulted in interoperability challenges. In RFC5085 the CV type of LSP Ping is made the default for MPLS PW's and ICMP Ping is made optional. This however, is a recommendation and not a requirement for implementations which can also lead to interoperability challenges.

To enable interoperability between implementations, this document defines a subset of control channels that is considered mandatory for VCCV implementation. This will ensure that VCCV remains the valuable tool it was designed to be in multi-vendor, multi-implementation and multi-carrier networks. This document also states requirements for the CV type too.

This draft is specific to MPLS PW's and not L2TPv3 PW. For the L2TPv3 PW only one CC and CV type are specified and the issues raised in this draft do not arise.

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 18, 2011.

Table of Contents

1. Introduction

[RFC5085] defines three Control Channel types for MPLS PW's: Type 1, using the Pseudowire Control Word, Type 2, using the Router Alert Label, and Type 3, using TTL Expiration (e.g. MPLS PW Label with TTL == 1). While Type 2 (RA Label) is indicated as being "the preferred mode of VCCV operation when the Control Word is not present," RFC 5085 does not indicate a mandatory Control Channel to ensure interoperable implementations. The closest it comes to mandating a control channel is the requirement to support Type 1 (Control Word) whenever the control word is present. As such, the three options yield seven implementation permutations (assuming you have to support at least one Control Channel type to provide VCCV). Many equipment manufacturers have gravitated to two common implementation camps: 1) Control Word and Router Alert Label support, or 2) TTL Expiration support only.

As a result, service providers are often faced with diametrically opposed support for VCCV Control Channel types when deploying mixed vendor networks. As long as operators select vendors from within a camp, VCCV can be used as the valuable fault-detection and diagnostic mechanism it was created to be. However, due to myriad other unrelated requirements associated with large router requirement specifications and related acquisitions, practice has shown it to be impractical to deploy equipment from only one camp or the other. As a result, this mismatch of support between camps often leads to a service provider's inability to use an important operational tool in networks supporting Layer 2 VPN services.

This document discusses the three Control Channel options, presents pros and cons of each approach and concludes with which Control Channel an implementation is required to implement.

This document also puts an explicit reqirement on the CV type to be supported for MPLS PW.

2. Comparison of Alternative Control Channel Types

The following section presents a review of each control channel type and the pros and cons of implementing each.

2.1. Control Channel Type 1: Control Word

As noted in [RFC5085], an in-band control channel is ideal, since this ensures that the connectivity verification messages follow the same path as the PWE3 traffic. VCCV Control Channel Type 1, also known as "PWE3 Control Word with 0001b as first nibble," is the only

"in-band" control channel specified. It uses the control word as opposed to using the label to indicate the presence of the Connectivity Verification message (CV). This ensures that the control channel follows the forwarding path of the associated traffic in all cases, including in the case of ECMP hashing.

The use of the control word is not mandatory on all PWE3 encapsulations. However based on the current hardware support the draft strongly suggest that all implementations SHOULD generically support the use of VCCV Control Channel Type 1 for all PWE3 encapsulations.

2.2. Control Channel Type 2: MPLS Router Alert Label

VCCV Control Channel Type 2 is also referred to as "MPLS Router Alert Label." In this approach, the VCCV control channel is created by using the MPLS router alert label [RFC3032] (e.g. Label Value = 1) immediately above the pseudowire label. As this label is inserted above the pseudowire label and below the PSN tunnel label, intermediate label switch routers do not act on the label. However, at the egress router, when the PSN tunnel label is popped and the next label is examined, the label value of 1 will cause the packet to be delivered to a local software module for further processing (e.g. processing of the VCCV Connectivity Verification (CV) message). Similarly, in the case of penultimate hop-popping, the labeled packet arrives with it's top-most label having a label value = 1, causing it to be delivered to a local software module for further processing.

As the processing behavior associated with Router Alert labels is germane to all MPLS implementations, VCCV Control Channel Type 2 should be supported by all implementations. However, there are issues with using Router Alert labels in operational networks. First, there are known issues related to the use of the Router Alert label and possible security risks associated with DoS attacks. While this is less of a risk in closed networks, this becomes a larger potential issue in inter-provider networks. Second, unlike use of the Control Word, inserting a label between the PSN tunnel label and the pseudowire label has ECMP implications, resulting in the very real possibility of the VCCV Control Channel diverging from the path of the associated traffic. While ECMP issues arise from both non-control-word control channels, given the security risks of using the Router Alert label, the VCCV Control Channel Type 2 cannot be mandatory for VCCV implementations.

All implementations MAY support VCCV Control Channel Type 2 so that operators who choose to use this approach can do so in mixed-implementation environments. Further, Router Alert Label MUST contain an appropriate TTL value, such that the TTL value does not

cause the CPU exception in any intermediate device in case of PHP.

2.3. Control Channel Type 3: MPLS PW Label with TTL == 1

VCCV Control Channel Type 3 is also known as "MPLS PW Label with TTL == 1." Unlike VCCV Control Channel Type 2, this approach uses the existing pseudowire label to indicate the need for further processing. Upon receiving the labeled packet, whether accompanied by a PSN tunnel label or alone (in the case of penultimate hop popping), the egress router makes a forwarding decision based on the Label Value, assuming the TTL is greater than or equal to 2. However, as part of this process and prior to forwarding the contents of the labeled packet to the attachment circuit (AC), the TTL is decremented. If the TTL value of the received packet was equal to 1, the TTL is decremented to 0, causing the packet to be sent to the control plane for processing.

Unlike the Router Alert Label (Label Value == 1), there has been no standardization of the pseudowire label TTL to this point. For example, [RFC3985], one of the only PWE3 RFCs to address TTL at all, states that "when a MPLS label is used as a PW Demultiplexer, setting of the TTL value in the PW label is application specific." However, no subsequent RFCs have defined the default value of the TTL field within the PW demultiplexer. With the advent of VCCV, it became clear that a TTL value greater than 1 was needed. Many implementations have settled on a default value of 2 for single-segment pseudowires, as evidenced by subsequent MIB drafts in which the default value of 2 is alluded to, if not explicitly defined. Consequently, implementations vary widely with regard to the default value of the TTL field and the subsequent behavior when the TTL is decremented to 0, if it is decremented at all.

Similar to VCCV CC Type 2, changing the value of the TTL in the existing PW demultiplexer label to something different from the value of the labels accompanying the associated traffic, can result in the VCCV Control Channel messages diverging from the path of the associated traffic when ECMP is employed.

Implementations MUST support the use of this option.

3. Mandatory Control Channels

Implementations of VCCV, at a minimum, MUST support VCCV Control Channel Type 3: MPLS PW Label with TTL == 1. Implementations of VCCV MUST also set the default TTL value of the PW demultiplexer label to 2 for single-segment pseudowires. Further, implementations of VCCV MUST decrement the TTL of the PW demultiplexer label in the egress

PE, and upon reaching a TTL==0, MUST pass the packet to the control plane for further processing of the VCCV message contained therein. This provides a basic level of interoperability across all implementations of VCCV without mandating the use of the control word for all VCCV-enabled pseudowire applications. Further, as VCCV is applied to multi-segment pseudowires, using Control Channel Type 3 enables PW traceroute to be implemented in a manner similar to that of MPLS and IP traceroute, through the incrementing of the TTL value in subsequent probes.

As noted previously, this baseline level of VCCV support does not address the aforementioned ECMP issues. Consequently, implementations of VCCV SHOULD support VCCV Control Channel Type 1 for pseudowire encapsulations for which a control word is not mandatory.

Implementations of VCCV MUST support VCCV Control Channel Type 1: Control Word for all implemented pseudowire encapsulations where use of the Control Word is mandatory. Implementations SHOULD support VCCV Control Channel Type 1 for implemented pseudowire encapsulations where, although optional, use of the control word is elected, on a pseudowire-by-pseudowire basis.

Implementations of VCCV MUST support the appropriate signaling of VCCV Control Channel Type support in the pseudowire setup signaling. In order to avoid interoperability issues, implementations should negotiate VCCV Control Channel Type, in decreasing priority: Type 1 (Control Word), Type 3 (TTL Expiration) and Type 2 (Router Alert), when all, or any permutation of the three CC Types are supported.

4. Mandatory CV Types

For MPLS PWs, the CV Type of LSP Ping (0x02) MUST be supported, and the CV Type of ICMP Ping (0x01) MAY be supported.

5. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

6. Security Considerations

This document describes the VCCV Control Channels which MUST be

implemented to ensure interoperability in a mixed-implementation environment. This document does not change the basic functionality associated with VCCV. As a result, no additional security issues are raised by this document over those already identified in [RFC5085].

7. Acknowledgements

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

[RFC3032] Rosen, E., "MPLS Label Stack Encoding", January 2001.

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

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

Authors' Addresses

Nick Del Regno (editor)
Verizon
400 International Pkwy
Richardson, TX 75081
US

Phone: 972-729-3411
Fax:
Email: nick.delregno@verizon.com
URI:

Vishwas Manral (editor)
IPInfusion Inc.
1188 E. Arques Ave.
Sunnyvale, CA 94085
US

Phone: 408-400-1900
Fax:
Email: vishwas@ipinfusion.com
URI:

Ruediger Kunze
Deutsche Telekom

Phone:
Fax:
Email: Ruediger.Kunze@telekom.de
URI:

Manuel Paul
Deutsche Telekom

Phone:
Fax:
Email: Manuel.Paul@telekom.de
URI:

Thomas Nadeau
Huawei

Phone:
Fax:
Email: tnadeau@lucidvision.com
URI:

Network Working Group
Internet-Draft
Updates: 4447 (if approved)
Category: Standards Track
Expires: September 11, 2011

Lizhong Jin (ed.), ZTE
Raymond Key (ed.), Telstra
Simon Delord, Alcatel-Lucent
Thomas Nadeau, Huawei
Vishwas Manral, IPInfusion
Sami Boutros, Cisco
Reshad Rahman, Cisco

March 11, 2011

Pseudowire Control Word Negotiation Mechanism Update
draft-jin-pwe3-cbit-negotiation-04

Abstract

This document describes the problem of control word negotiation mechanism specified in [RFC4447]. Based on the problem analysis, a message exchanging mechanism is introduced to solve this control word negotiation issue. This document is to update [RFC4447] control word negotiation mechanism.

Status of this Memo

This Internet-Draft is submitted to IETF 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 September 11, 2011.

Table of Contents

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].

1. Introduction

This document describes the problem of control word negotiation mechanism specified in [RFC4447]. Based on the problem analysis, a message exchanging mechanism is introduced to solve this negotiation issue. The control word negotiation mechanism in this document is to update [RFC4447] section 6.2 "PW Types for Which the Control Word is NOT Mandatory".

2. Problem Statement

[RFC4447] section 6 describes the control word negotiation mechanism. Each PW endpoint has the capability of being configurable with a parameter that specifies whether the use of the control word is PREFERRED or NOT PREFERRED. While in some case of control word negotiation, PE will advertise C-bit=0 in label mapping message with its locally configured control word PREFERRED. This kind of behavior will cause some problem when peer PE changes its control word from NOT PREFERRED to PREFERRED.

This following case will describe the negotiation problem in detail:

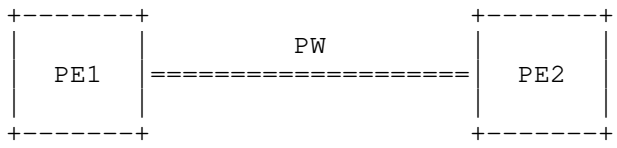


Figure 1

1. Initially, the control word on PE1 is configured to PREFERRED, and on PE2 to NOT PREFERRED.

2. The negotiation result for the control word for this PW is "not supported", and PE1 send label mapping with C-bit=0 finally.

3. PE2 then changes its control word configuration to PREFERRED.

4. PE2 will then send label withdraw message to PE1.

5. According to the control word negotiation mechanism, the received label mapping on PE2 from PE1 indicates C-bit=0, therefore PE2 will still send label mapping with C-bit=0.

The negotiation result for the PW control word is still "not supported", even though the control word configuration on both PE1 and PE2 is set to PREFERRED.

3. Control word re-negotiation by label request message

In order to solve this problem, the control word re-negotiation is operated by adding label request message. The control word negotiation mechanism can still follow the procedure described in [RFC4447] section 6.

When Local PE changes its control word from NOT PREFERRED to PREFERRED and only if it already received the remote label mapping message with C-bit=0, additional procedure will be added as follow:

-i. Local PE MUST send a label withdraw message to remote PE if it has previously sent a label mapping, and wait until receiving a label release from the remote PE.

-ii. Local PE MUST send a label request message to remote PE, and wait until receiving a label mapping message containing the remote PE configured control word setting.

-iii. After receiving remote PE label mapping with control word setting, Local PE MUST follow procedures defined in [RFC4447] section 6 when sending its label mapping message.

When the peer PE successfully processed the label withdraw and removed the remote label binding, it MUST send label mapping as a response of label request with locally configured control word parameter.

Note: the FEC element in label request message should be the PE's local FEC element. Only if FEC element in label request message could bind together with peer PE's local FEC element, the peer PE sends label mapping with its bound local FEC element and label as a response. The label request message format and procedure is described in [RFC5036].

The multi-segment PW case for T-PE is same, and the request message MUST be processed in ordered mode. When S-PE receives a label request message from a remote peer, it MUST advertise the request message to the other remote PE. This is necessary since S-PE does not have full information of interface parameter field in the FEC advertisement.

As local T-PE will send label withdraw before sending label request to remote peer, the S-PE MUST send the label withdraw upstream before it advertises the label request. When S-PE receives the label withdraw, it should process this message to send a label release as a response and a label withdraw to upstream S-PE/T-PE, then process the next LDP message, e.g. the label request message.

When Local PE changes its control word from PREFERRED to NOT PREFERRED, Local PE would be able to re-negotiate the Control Word to be NOT PREFERRED following the procedures defined in [RFC4447], and no label request message to peer PE will be needed. In that case, Local PE will always send new label mapping with C-bit reflecting the local Control Word configuration.

The procedure of PE1 and PE2 for the case in figure 1 should be as follows:

1. PE2 changes locally configured control word to PREFERRED.

2. PE2 will then send label withdraw message to PE1.

3. PE1 will send label release in reply to label withdraw message from PE2.

4. Upon receipt of Label release message from PE1, PE2 MUST send label request messages to PE1 although it already received the label mapping with C-bit=0.

5. PE1 MUST send label mapping message with C-bit=1 again to PE2 (Note: PE1 MUST send label mapping with locally configured CW parameter).

6. PE2 receives the label mapping from PE1 and updates the remote label binding information. PE2 MUST wait for PE1 label binding before sending its label binding with C-bit set, only if it previously had a label binding with C-bit=0 from PE1.

7. PE2 will send label mapping to PE1 with C-bit=1.

It is to be noted that the above assume that PE1 is configured to support CW, however in step 5 if PE1 doesn't support CW, PE1 would send the label mapping message with C-bit=0, this would result in PE2 in step 7 sending a label mapping with C-bit=0 as per [RFC4447] CW negotiation procedure.

By sending label request message, PE2 will get the configured CW parameter of peer PE1 in the received label mapping message. By using the new CW parameter from label mapping message received from peer PE1 and locally configured CW, PE2 should determine the PW CW parameter according to [RFC4447] section 6.

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

4. Backward Compatibility

Since control word re-negotiation is operated by adding label request message, and still follows the procedure described in [RFC4447] section 6, it is fully compatible with existing implementations. The remote PE (PE1 in figure 1) which already implement label request message could be compatible with the PE (PE2 in figure 1) following the mechanism of this document.

5. Security Considerations

This document does not introduce any additional security constraints.

6. IANA Considerations

This document does not require IANA assignment.

7. 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.

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

[RFC4447] Martini, L., and al, Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP), RFC 4447, April 2006

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

Authors' Addresses

Lizhong Jin (editor)
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China
Email: lizhong.jin@zte.com.cn

Raymond Key (editor)
Telstra
242 Exhibition Street, Melbourne
VIC 3000, Australia
Email: raymond.key@ieee.org

Simon Delord
Alcatel-Lucent
Email: simon.delord@gmail.com

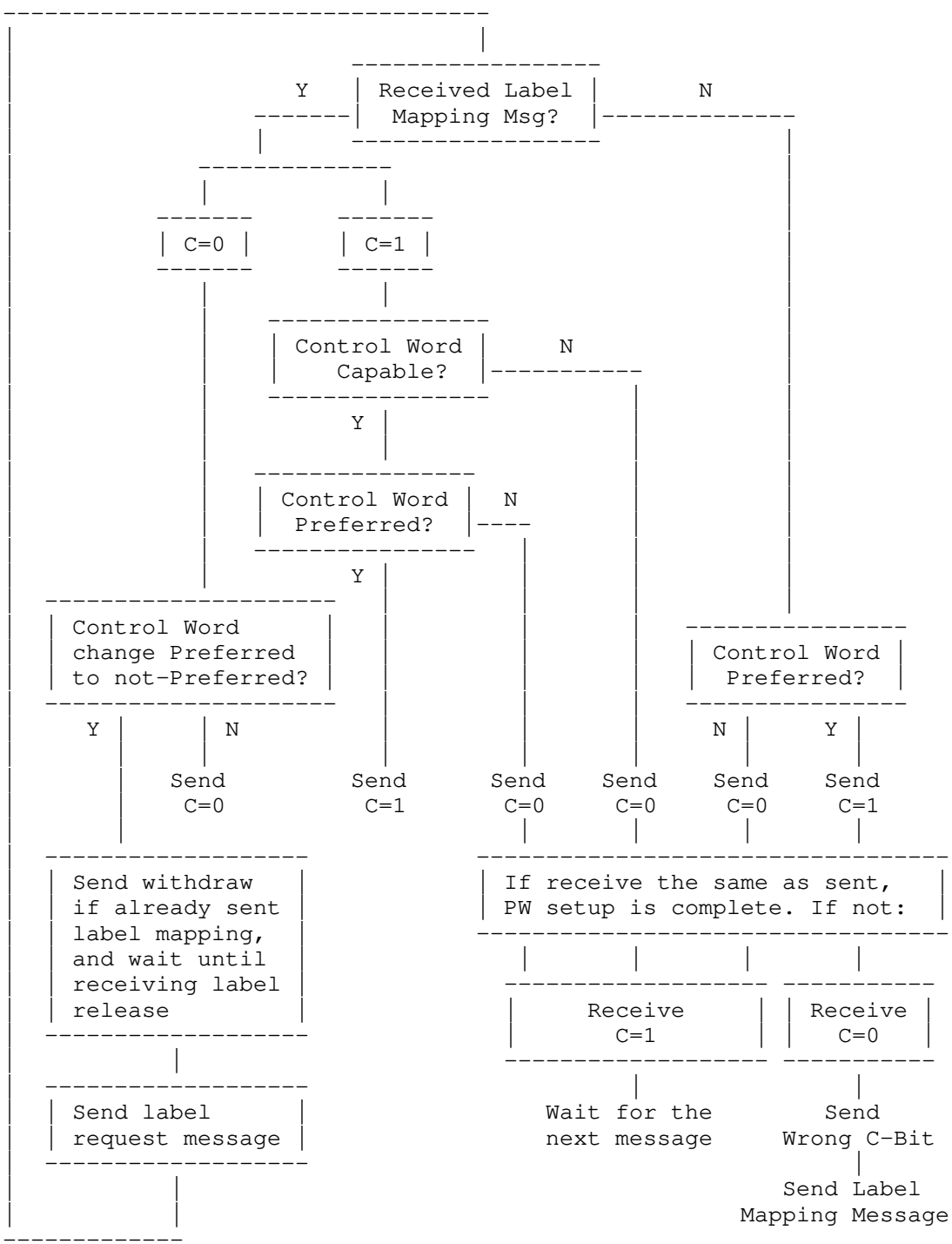
Thomas Nadeau
Huawei
Email: tnadeau@lucidvision.com

Vishwas Manral
IPInfusion
Email: vishwas@ipinfusion.com

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

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

Appendix A. Updated C-bit Handling Procedures Diagram



PWE3 Working Group S. Kini
Internet-Draft D. Sinicrope
Intended Status: Standards Track Ericsson
Expires: September 2011 March 14, 2011

Pseudowire Virtual Circuit Connectivity Verification (VCCV): An Inband Control Channel using offset draft-kini-pwe3-inband-cc-offset-01.txt

Status of this Memo

Copyright Notice

Abstract

Pseudowires need an inband control channel (CC) to do VCCV such that OAM and data packets follow the same path. PWs without a Control Word (CW) do not have an inband CC as defined in RFC5085. This document defines a simple extension to the TTL expiry CC (Type 3) to do inband VCCV. This can be used even without a CW.

Table of Contents

1. Introduction

OAM functions such as connectivity verification (CV) need an inband channel to do their operations. Only an inband control channel ensures that packets carrying OAM messages follow the same path as the data packets that they are doing OAM operations for. Most PW deployments today do not have CW enabled. However the control channels defined in [VCCV] provide an inband CC only when CW is enabled. Moreover enabling CW prevents from looking beyond the label stack to do multipath decisions. At an intermediate LSR, looking at an IP header beyond the label stack to do multipath is desirable since it is a commonly available capability in current implementations and also helps to do multipath load sharing based on a true end to end flow (e.g. [ID.PPW-EIM]), rather than rely on additional mechanisms such as [FAT-PW]. This document briefly describes the problem with the TTL Expiry CC (Type 3) in section 3. A simple extension to this CC to solve this problem is described in section 4.

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

A VCCV control channel (CC) that uses TTL expiry is not inband when the intermediate nodes along a LSP look beyond the label stack to do multipath forwarding decisions. However it is mandatory ([ID.VCCV-MF]) and is widely used especially in the commonly deployed scenario of PWs that do not use a CW (Control Word). A PW that uses CW is also unable to take advantage of the presences of multipath in the server layer. Multipath is considered useful for both redundancy as well as load sharing.

4. Solution

This document defines a new VCCV CC. It is an extension of the TTL Expiry VCCV (Type 3) defined in [VCCV]. In this CC the associated channel starts at a fixed offset after the PW label. This CC is henceforth referred to as Inband-offset VCCV (Type TBA). A fixed number of bytes between the PW label and the start of the associated channel can be used to emulate flow header information and are henceforth referred to as a "pseudo flow header". A VCCV message with a pseudo flow header will follow the same path as that taken by a data packet of the flow, as long as any multipath forwarding decision taken by the intermediate LSRs do not look beyond the pseudo flow

header. A pseudo flow header length of 64 bytes is expected to meet the requirements of all current implementations and also meet the requirements of deployments (both current and in the foreseeable future). If a size other than 64 is needed then it can be configured or signaled as an attribute of the PW. The content of the pseudo flow header is set according to the flow that needs an OAM function such as connectivity verification (CV). E.g. if the encapsulation consists of an IP packet following the PW label, then the pseudo flow header would be the IP header of a flow.

5. Security Considerations

This document does not introduce any new security considerations beyond those already listed in [VCCV].

6. IANA Considerations

IANA needs to allocate a value for Inband-offset VCCV in the registry "MPLS VCCV Control Channel Type". Recommend next available bitfield 0x8.

7. Future work

1. Define signaling extensions to convey the size of the offset.
2. Authenticate VCCV messages.

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.

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

[ID.VCCV-MF] Del Regno, N., et al, "Mandatory Features of Virtual Circuit Connectivity Verification Implementations", draft-delregno-pwe3-vccv-mandatory-features-01 (work in progress), April 2010.

[ID.PPW-EIM] Kini, S., et al, "Encapsulation Methods for Transport of packets over an MPLS PSN - efficient for IP/MPLS", draft-kini-pwe3-pkt-encap-efficient-ip-mpls-01 (work in progress), March 2011.

8.2. Informative References

[FAT-PW] Bryant, S., et al, "Flow Aware Transport of Pseudowires over an MPLS PSN", draft-ietf-pwe3-fat-pw-04 (work in progress), July 2010.

9. Acknowledgements

The authors would like to thank Joel Halpern, Luca Martini and Raymond Key for their comments.

Appendix A: Examples

Consider a PPW-EIM that is setup and an OAM CV has to be done for a specific flow. Say the flow is between IPv4 endpoints with addresses 209.85.153.104 and 209.191.122.70 that have a TCP session between ports 80 and 53950 respectively. To perform OAM operations for this flow e.g. to do a VCCV ping for this flow, the encoding of the VCCV packet is as follows.

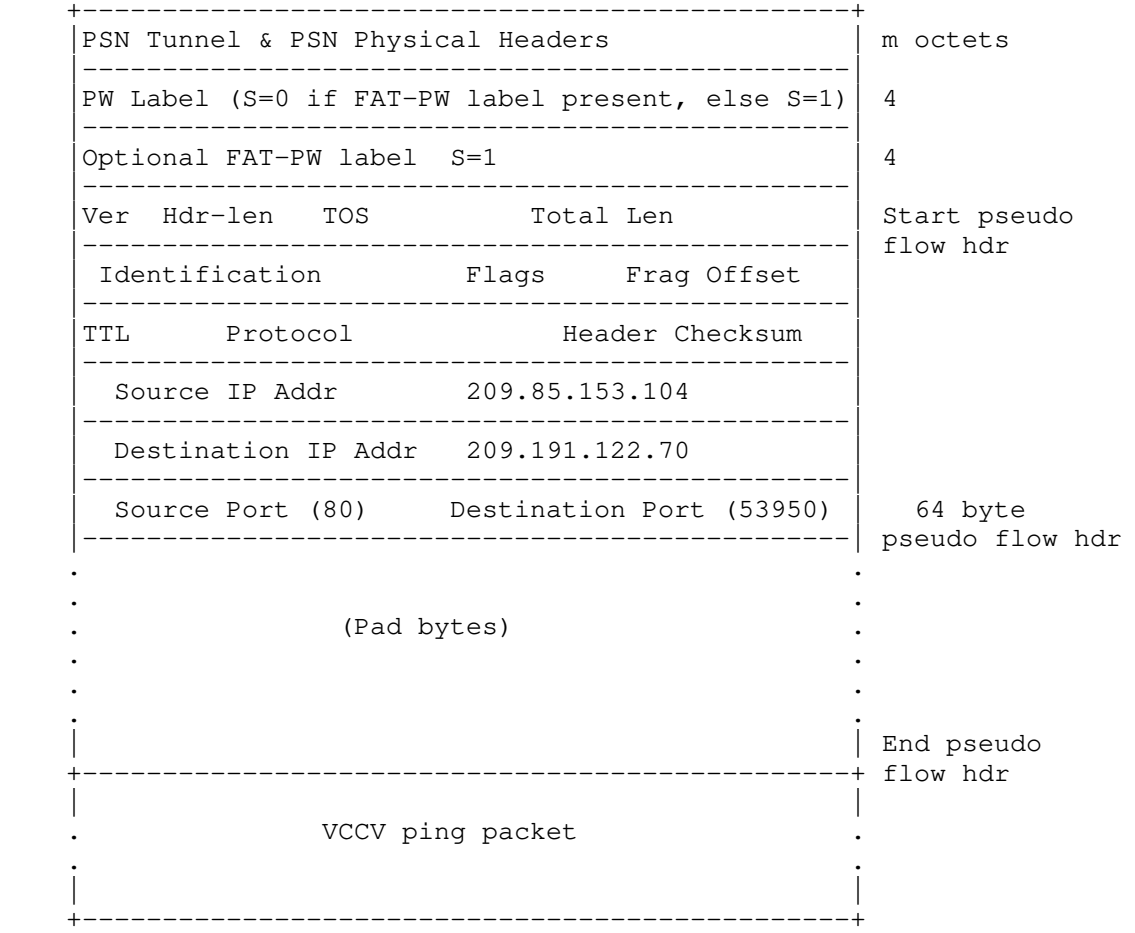


Figure 1 IPv4/v6 packet encapsulated into PPW without CW

Authors' Addresses

Sriganesh Kini
Ericsson
300 Holger Way, San Jose, CA 95134
EMail: sriganesh.kini@ericsson.com

David Sinicrope
Ericsson
8001 Development Dr, Research Triangle Park, NC 27709
EMail: david.sinicrope@ericsson.com

MPLS Working Group S. Kini
Internet-Draft D. Sinicrope
Intended Status: Standards Track Ericsson
Expires: September 2011 March 14, 2011

Encapsulation Methods for Transport of packets over an MPLS PSN - efficient for IP/MPLS

draft-kini-pwe3-pkt-encap-efficient-ip-mpls-02.txt

Status of this Memo

Distribution of this memo is unlimited.

This Internet-Draft is submitted to IETF 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 September 15, 2011.

Copyright Notice

Abstract

A Packet Pseudowire (PPW) must be able to carry a packet of any protocol that can be carried over Ethernet. In many cases IP and MPLS are the pre-dominant protocols on a PPW transported over an MPLS PSN. Other protocols are used mainly for control purposes. In such a scenario it is highly beneficial to make IP/MPLS encapsulation efficient. This document defines such an encapsulation while retaining the ability to exchange packets of any other protocol over the PPW.

Table of Contents

1. Introduction

A packet transport service modeled along [PWE3-ARCH] is considered useful. Such a service is also referred to as a packet pseudowire (PPW). The server network is a Packet Switched Network (PSN) and could be a MPLS (or a MPLS-TP) network. The client requires a generic packet transport service that is isolated from the underlying PSN.

It must be possible to carry any number and type of client protocols on the PPW, similar to Ethernet. Some of these may be purely control protocols such as [ARP] or [LLDP]. Such protocols may not take up the majority of the bandwidth of the service. On the other hand client protocols such as IP and MPLS can take up the majority of the bandwidth and it is very useful for the PPW to encapsulate them efficiently.

This document defines an encapsulation for a PPW over a MPLS PSN that efficiently encapsulates IP and MPLS. However it is still possible to carry all client protocols on the PPW. It is useful when IP and/or MPLS are the pre-dominant protocols on the PPW. The encapsulation defined in this document is referred to as PPW-EIM (where EIM stands for Efficient IP MPLS). The efficiency is realized by minimizing any extra headers that would be needed to transport an IP or MPLS packet when compared to a solution such as [PWE3-ETH]. The benefits of this efficiency include increased bandwidth available for user traffic due to lesser overhead, better throughput due to reduced possibility of fragmentation and also more efficient use of ECMP paths.

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. Scope

This document covers a PPW as a point-to-point (p2p) service. Multi-access service is considered outside the scope of this version of the document.

The encapsulation scheme PPW-EIM is useful when IP/MPLS packets are the majority of the packets on the PPW. The method to determine this is considered outside the scope of this document.

4. Network Reference Model

The solution in this document addresses the following two cases of the reference model in Figure 2 of [PWE3-ARCH]

1. The native service is an ethernet virtual circuit (EVC). The EVC may either be untagged or tagged. The untagged traffic is treated as a unique EVC. The stack of VLAN Identifiers (VIDs) in the VLAN tags stack of an Ethernet frame uniquely identifies an EVC. The number of VIDs in the stack identifying the circuit may be one (as in [802.1q], e.g. a customer tag C-tag) or more (similar to [802.1ad] e.g. a customer and service tag C-tag and S-tag). Typically the physical interface between CE and PE will be an Ethernet interface. Note that if another VLAN tag is stacked on an EVC it MUST be treated as a separate EVC to apply PPW-EIM. This is a subset of the reference model in [PWE3-ETH] and is henceforth referred to as PWE3-ETH-EVC. PPW-EIM encapsulates a single EVC into a PPW. If a packet transport service is required for multiple EVCs then a separate PPW should be used for each. The encapsulation in [PWE3-ETH] must be used instead of PPW-EIM under the following conditions:

    a. If an EVC has to be transported transparently in a single pseudowire (PW) by carrying all VLAN tags encapsulated inside the EVC.

    b. If the EVC is not pre-dominantly carrying IP or MPLS. The method to determine this is outside the scope of this document.

    c. If there are a large number of EVCs (pre-dominantly carrying IP/MPLS) that need a p2p transport service towards another PE but one of the PEs has PPW scaling limitations that prevent it from creating separate PPWs per EVC as required by PPW-EIM.

2. The CE and the corresponding PE are co-located in the same equipment. This is similar to a virtual untagged point-to-point (p2p) Ethernet interface between the two CEs. This should be treated as the case of providing p2p transport service for the untagged traffic EVC of the PWE3-ETH-EVC reference model described above.

It should be noted that the access circuit is modeled as an EVC since an EVC can carry any protocol packet. However, the technique defined in this draft can be extended to any access circuit encapsulation that encapsulates IP and MPLS packets.

5. Solution

This solution does not use a data link layer header (such as Ethernet) on the PPW to transport IP/MPLS packets. This reduces the overhead bytes for such packets. There are implementations that look

beyond the MPLS label stack for an IP packet. For non IP/MPLS packets, whenever there is a potential for such a condition, an IP encapsulation (with GRE) is used. Thus ECMP based on looking for an IP packet beyond the MPLS stack will work correctly and not re-order any flows. To prevent the GRE encapsulated packets from having IP address conflicts with the IP address space of the customer's network, a non-routable IP address (in the 127/8 range) is used. The details of the packet encapsulation are in section 5.1. The adaptation of PE-bound and CE-bound traffic is explained in section 5.2.

5.1. Encapsulation format on the PPW

The encapsulation of the packet is described below along with any control word (CW) bits that are required to be defined. A more formal definition of the CW for PPW-EIM is in section 5.5.

5.1.1. IP packets

An IPv4/v6 packet encapsulation into a PPW depends on whether CW is present. If the CW is not present, the encapsulation is as shown in Figure 1. Any ECMP implementation that looks for an IP packet beyond the label stack will not re-order flows. If the CW is present then the flags bits 6 and 7 in the CW are set to 01. The encapsulation is as shown in Figure 2. In both cases the first nibble of the IP packet is used to distinguish between an IPv4 and IPv6 packet.

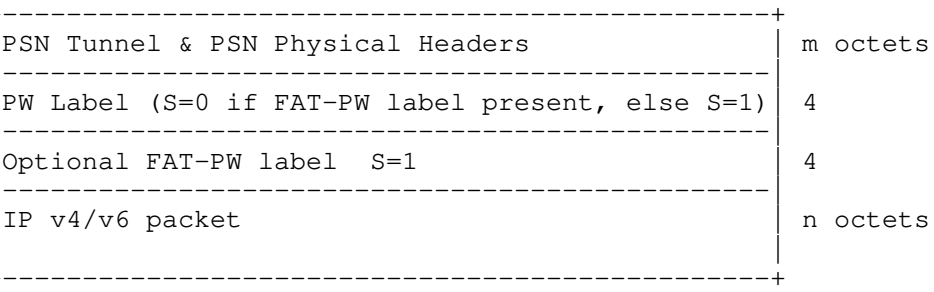


Figure 1 IPv4/v6 packet encapsulated into PPW without CW

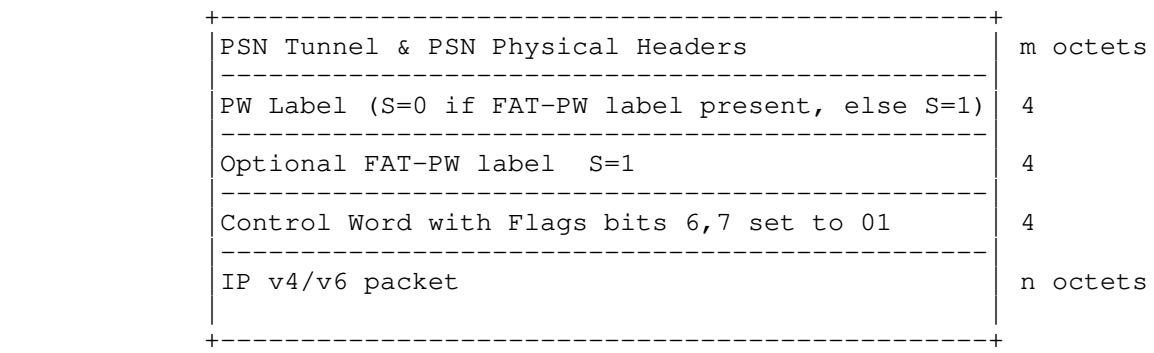


Figure 2 IPv4/v6 packet encapsulated into PPW with CW

5.1.2. MPLS packet

A MPLS packet encapsulation into a PPW depends on whether the CW is present in the packet. If the CW is present then the flags bits 6 and 7 in the CW are set to 10. The encapsulation is as shown in Figure 3. If the CW is not present, the S-bit in the bottom-most label in the pseudowire label stack is set to zero and the format is as shown in Figure 4. The pseudowire label stack (including the PSN tunnel label stack if any) along with the label stack of the payload appear as a single label stack. This is also consistent with the notion of having a single S-bit set in a labeled packet. Since the payload (MPLS) has (independently) ensured that looking beyond the label stack correctly interprets IP payloads and PWE3 payloads, the same holds true for the combined label stack. Hence flows are identified correctly.

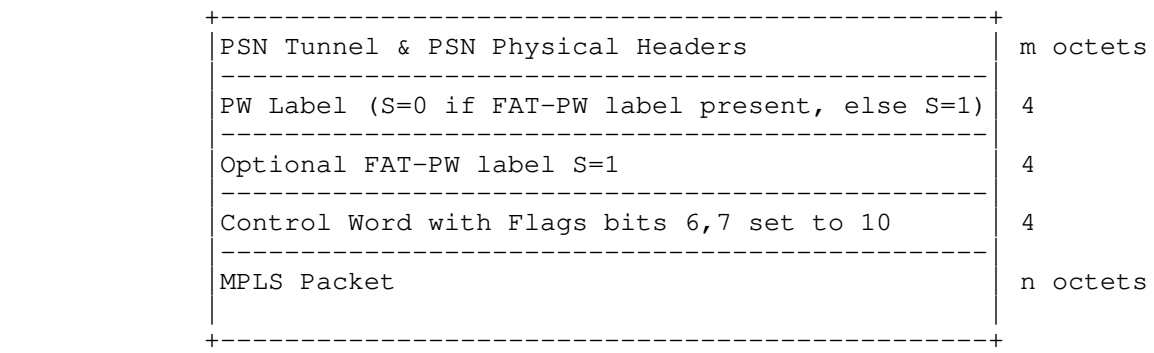


Figure 3 MPLS packet encapsulated into PPW with CW

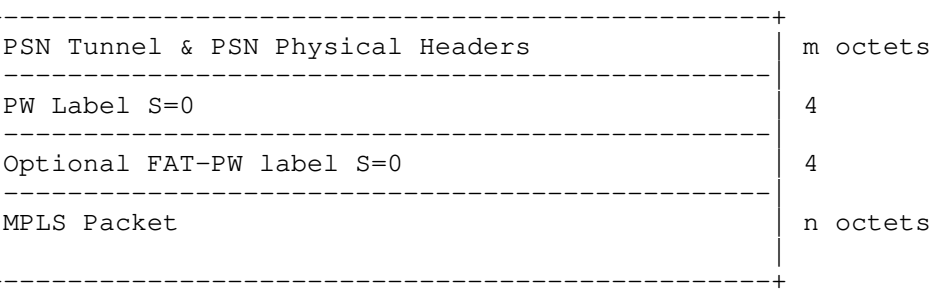


Figure 4 MPLS packet encapsulated into PPW without CW

5.1.3. An arbitrary protocol

An arbitrary protocol (other than IP and MPLS) being encapsulated into a PPW depends on whether a CW is present. If a CW is not present a GRE encapsulation MUST be used as shown in Figure 5. This extends the encapsulation for an IPv4 packet shown earlier in Figure 1 of section 5.1.1. The IP destination addresses in the GRE delivery header is a non-routable address from the 127/8 range. These are used to identify that the packet does not belong to a real GRE tunnel in the IP address space of the payload but rather is a protocol packet on the PPW. Also the protocol type in the GRE Header is according to the protocol that is being carried. The TTL in the GRE delivery header is set to 0 (or 1) to prevent this packet from being IP routed.

If the CW is present then the flags bits 6 and 7 in the CW are set to 00 and the format is as shown in Figure 6. Note that the ethernet frame carrying the arbitrary protocol packet immediately follows the CW. The GRE encapsulation is not needed in this case.

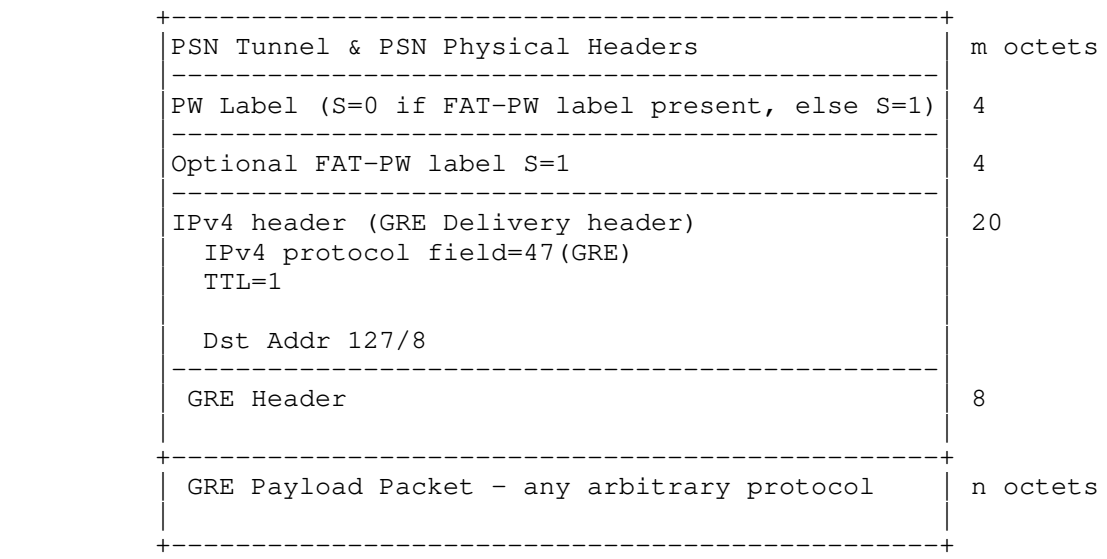


Figure 5 An arbitrary protocol packet encapsulated into PPW without CW

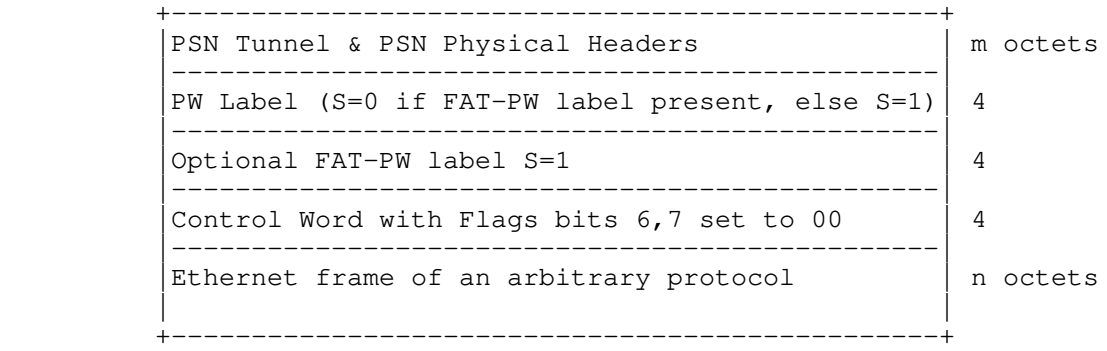


Figure 6 An arbitrary protocol packet encapsulated into PPW with CW

5.2. Traffic adaptation

5.2.1. PE-bound

After the Native service processing (NSP), the Ethernet frame (from CE) MUST be mapped into the PPW based on the value of the Ethernet type field as follows:

1. If it is IP (0x800 - IPv4 or 0x86DD - IPv6), the Ethernet header (including the VLAN tags stack) is stripped off and the encapsulation format is as described in section 5.1.1. Note

that the flags bits 6 and 7 in the CW MUST be set to 01.

2. If it is MPLS (0x8847, 0x8848), the Ethernet header (including the VLAN tags stack) is stripped off and the encapsulation format is as described in section 5.1.2. The S-bit in the bottom-most label of the pseudowire label stack is set to 1 or 0 depending whether the CW is present or not respectively. Note that the flags bits 6 and 7 in the CW MUST be set to 10.

3. For all other values of the Ethernet type field, the entire Ethernet frame is carried on the PPW. Depending on whether the CW is use, the encapsulation is as follows:

    a. If CW is not present then the frame is first encapsulated into GRE (with IP) and the encapsulation format is as described in section Figure 3. The GRE header protocol-type is set according to the protocol being carried. The IP destination address MUST be chosen from the 127/8 range. Typically the same source and destination addresses SHOULD be used for the life of the PPW. The IP header TTL SHOULD be set to 0. If there is any hardware limitation due to which TTL of zero cannot be set then a TTL of 1 MUST be used. The checksum in the GRE Header and the IP header MAY be set to 0 since the packet is not forwarded based on these headers and the protocol packet typically has its own data integrity verification mechanisms. If the IP packet (encapsulating GRE) exceeds the PW's MTU, IP fragmentation SHOULD be used provided the PW peer is capable of IP reassembly. If the PW peer is not capable of reassembly the packet must be dropped.

    b. If CW is present then the Ethernet frame immediately follows the CW. If packet exceeds MTU then [PWE3-FRAG] SHOULD be used.

5.2.2. CE-bound

The association between the EVC and the PPW has the following extra information that will be used when adapting traffic from the PPW to the EVC.

1. MAC address of the directly connected CE. This would be the source MAC address of any frame received from the CE and is henceforth referred to as PPW-EIM-SMAC. This may be configured, signaled or dynamically learnt.

2. MAC address of the remotely connected CE. This would be the source MAC address of any frame received from the remote CE and

is henceforth referred to as PPW-EIM-DMAC. This may be configured or dynamically learnt.

3. The VLAN tag stack (henceforth referred to as PPW-EIM-VSTACK). The VLAN Identifier (VID) portion of PPW-EIM-VSTACK should be known as this uniquely identifies the EVC. The Canonical Format Indicator (CFI) must always be 0.

4. A mapping function to map IP differentiated services (DS) [RFC2474] field to Ethernet PCP bits (henceforth referred to as PPW-EIM-DS-to-PCP). This is applicable only if the EVC is tagged. If there are multiple tags in the VLAN tag stack this may be a separate mapping for each tag. It is recommended that the same mapping be used for all tags. The mapping may be user-configurable. A default mapping of a DS field "xyzPQRCU" to a PCP of "xyz" is recommended.

When the packet is parsed the type and location of the user data is known. If the packet belongs to the G-ACh then its processing is defined in [VCCV] and remains unchanged for PPW-EIM. The processing for an IP or MPLS packet in the PW is as follows:

1. If the payload of the PPW is an MPLS packet it is mapped into an Ethernet frame as follows:

   a. PPW-EIM-SMAC as the source MAC address.

   b. PPW-EIM-DMAC as the destination MAC address.

   c. PPW-EIM-VSTACK as the VLAN tag stack. The PCP bits for each tag in the stack are mapped from the Traffic Class (TC) bits of the first MPLS label in the payload.

   d. The Ethernet type field is set to 0x8847 (MPLS).

2. If the payload of the PPW is an IP packet, the first nibble of the IP header and the Protocol-type then determine further processing.

   a. If the first nibble is 0x6 then the payload of the PPW is an IPv6 packet. The IPv6 packet is mapped into an Ethernet frame as follows:

      i. PPW-EIM-SMAC as the source MAC address.

      ii. If the destination IPv6 address is broadcast/multicast then the destination MAC address of the Ethernet frame is determined

accordingly. Else if the destination IPv6 address is unicast then PPW-EIM-DMAC is used.

iii. PPW-EIM-VSTACK as the VLAN tag stack. The PCP bits for each tag in the stack are mapped from the DS field in the IPv6 header using PPW-EIM-DS-to-PCP mapping.

iv. The Ethernet type field is set to 0x86DD (IPv6)

b. If the first nibble is 0x4 then the payload of the PPW is an IPv4 packet. The IP destination address together with protocol field determines further processing:

i. If the destination IP address is in the 127/8 range and the protocol field is 47 (GRE) then the GRE payload packet is an arbitrary protocol packet on the PPW. It should be noted that comparing 3 fields that start at fixed offsets in the header and require a comparison of a fixed number of bits from those offsets is sufficient to shunt the packet off the IP/MPLS de-capsulation path. These three fields are the first nibble (starting offset 0, field size 1 nibble), IP header protocol field (starting offset 10, field size 2), IP destination address (starting offset 16, compare just first byte). Moreover these comparisons are against fixed values and should be easily implementable in hardware. Further validation of the GRE Delivery header for checksum, TTL, etc as well as the GRE header validation can be done after the packet is shunted off the IP/MPLS de-capsulation path. The VLAN tag stack in the Ethernet frame is validated against PPW-EIM-VSTACK and if the VLAN IDs match, the frame is passed to the NSP. If the IP packet was fragmented it SHOULD be reassembled. If the node is not capable of IP reassembly, the packet is dropped.

ii. For all other values it is an IPv4 packet and the processing is similar to that of an IPv6 packet except that the Ethernet type field on the CE-bound frame is set to 0x800 (IPv4).

3. If the payload of the PPW is any protocol packet, then it is an Ethernet frame.

5.3. QoS considerations

The QoS considerations in [PWE3-ETH] are applicable in this document.

5.4. PW Types

Depending on the requirements of a particular deployment the packet transport service may be required to carry only a subset of the packet types that are carried on a PPW. The following deployment scenarios of the client network on the p2p link (that is emulated by the PPW) are considered useful:

1. IP only - In this deployment scenario the client network uses the p2p link to exchange exclusively IP packets. This would be especially true when the PE and CE co-exist on the same device at both ends of the PPW and the CE's exchange only IP packets on that p2p link. A MAC address is not needed in this case. This deployment scenario would also be the case when the PE and CE are on separate devices, the CE's exchange only IP packets on the p2p link and the MAC address mapping for the IP is configured on the CE (e.g. static ARP entry). IP encapsulated control protocols (such as RIP, OSPF, etc) could run on the link.

2. IP and ARP only - In this deployment scenario the client network uses the p2p link to exchange exclusively IP packets but additionally uses ARP for layer-2 address resolution.

3. MPLS only - In this deployment scenario the client network uses the p2p link to exchange exclusively MPLS packets. Typically the client network would be purely a MPLS (or MPLS-TP) network and would not even use an IP based control plane. This deployment scenario would be especially true when the PE and CE co-exist on the same device at both ends of the PPW and the CE's exchange only MPLS packets on the p2p link. A MAC address is not needed in this case. This deployment scenario would also be the case when the PE and CE are on separate devices, the client network uses the p2p link to exchange MPLS (or MPLS-TP) packets and the mapping of MPLS-label to MAC address is configured on the CE. The MAC address may be from an assigned range (as defined in MPLS-TP).

4. IP/MPLS only - In this deployment scenario the client network uses the p2p link to exchange exclusively IP/MPLS packets. This would be the typical case when the PE and CE co-exist on the same device at both ends of the PPW and the CE sends only IP/MPLS packets on the p2p link. A MAC address is not needed in this case. This would also be the case when the PE and CE are

on separate devices but the MAC address mapping for IP and MPLS is configured on the CE (e.g. static ARP entry). IP encapsulated control protocols (such as RIP, OSPF, BGP, LDP, RSVP-TE, etc) could run on the link.

5. IP/MPLS and ARP only - In this deployment scenario the client network uses the p2p link to exchange exclusively IP/MPLS packets but additionally uses ARP for layer-2 address resolution. This is the typical case when the client network uses that p2p link exclusively with the IP protocol for layer-3 routing and MPLS protocol for switching but uses ARP for layer-2 address resolution.

6. Generic packet service - In this deployment scenario the client network can use the p2p link to exchange any type of packet that can be sent over an EVC. Even MAC address configuration is not necessary since ARP can be run on this link.

For many of these scenarios a subset of the encapsulation and traffic adaptation that has been defined for PPW-EIM is relevant. The following pseudowire types are additionally defined that perform a subset of the full functionality of PPW-EIM.

1. IP-only-PPW-EIM - Only IP traffic is transported in PPW-EIM. The relevant encapsulations are in section 5.1.1. Only the adaptations for IP traffic are relevant from section 5.2. This PW would not implement the [GRE] encapsulation. It would optionally implement the CW. When the CW is not used the encapsulation format of this PW is similar to L3VPN.

2. MPLS-only-PPW-EIM - Only MPLS traffic is transported in PPW-EIM. The relevant encapsulations are in 5.1.2. Only the adaptations for MPLS traffic are relevant from section 5.2. This PW would not implement the [GRE] encapsulation. It would optionally implement the CW. When the CW is not used, the encapsulation (label-stack) of this PW is similar to a MPLS-TP LSP that has MPLS as a client.

3. IPMPLS-only-PPW-EIM - Only IP and MPLS traffic is transported in PPW-EIM. The relevant encapsulations are in sections 5.1.1. and 5.1.2. Only the adaptations for IP and MPLS traffic are relevant from section 5.2. This PW would not implement the [GRE] encapsulation. It would optionally implement the CW.

Each deployment scenario described earlier can be realized by the generic PPW-EIM. However many deployment scenarios can also be realized by a PPW that implements a subset of PPW-EIM. The method and choice of PPW to do this for each deployment scenario is as follows:

1. IP only - A PW can be realized with an IP-only-PPW-EIM.

2. IP and ARP only - The straightforward way to realize this is by the generic PPW-EIM. It is also possible to realize it using an IP-only-PPW-EIM if the PE acts as a proxy ARP ([PXY-ARP]) gateway to its directly connected CE.

3. MPLS only - A PW can be realized with a MPLS-only-PPW-EIM.

4. IP/MPLS only - A PW can be realized with an IPMPLS-only-PPW-EIM.

5. IP/MPLS and ARP only - The straightforward way to realize this is by the generic PPW-EIM. It is also possible to realize it using an IPMPLS-only-PPW-EIM if the PE acts as a proxy ARP gateway to its directly connected CE.

6. Generic packet service - This of course should be realized using PPW-EIM.

5.5. Control Word

One of the primary purposes of the CW ([PWE3-CW]) is to prevent re-ordering within a flow if there are implementations that look beyond the label stack for an IP flow. PPW-EIM has different characteristics due to the use of IP for encapsulating non IP/MPLS packets. Hence a CW is considered optional and the characteristics of PPW-EIM without a CW are analyzed in section 5.5.1. A CW that meets the requirements in [PWE3-CW] is described in section 5.5.2. This should be used in cases where a CW is required for reasons other than preventing flow re-ordering.

5.5.1. Characteristics without CW

PPW-EIM (without CW) is not susceptible to re-ordering flows within the PPW. It can also take advantage of ECMP implementations that examine the first nibble after the MPLS label stack to determine whether the labeled packet is an IP packet. Such implementations are widely available today and will correctly identify the IP flow in the PPW. Even the flows of non IP/MPLS protocols will not be re-ordered as long as the same source and destination IP addresses are used in the GRE Delivery header for the life of the PPW. Hence a CW is not necessary for PPW-EIM to prevent flow re-ordering. This can also obviate the need for [FAT-PW] within PPW-EIM and thereby save on processing power at ingress to identify the flow (through packet classification) and add the flow-label. When an ECMP based on the label stack is required (and available), then [FAT-PW] must be used with PPW-EIM. An important benefit of not adding a CW and/or flow-

label is that the difference in packet size between the access network and the PSN is further reduced by up to 8 bytes (compared with [PWE3-ETH]) and hence there is less chance for fragmentation of jumbo IP/MPLS packets.

5.5.2. PPW-EIM-CW

If a CW is needed for PPW-EIM, then the one defined in [PWE3-ETH] must be used with the following extension. In accordance with the preferred CW format in [PWE3-CW] that specifies the flags field for per-payload signaling, the bits 6 and 7 are defined as follows:

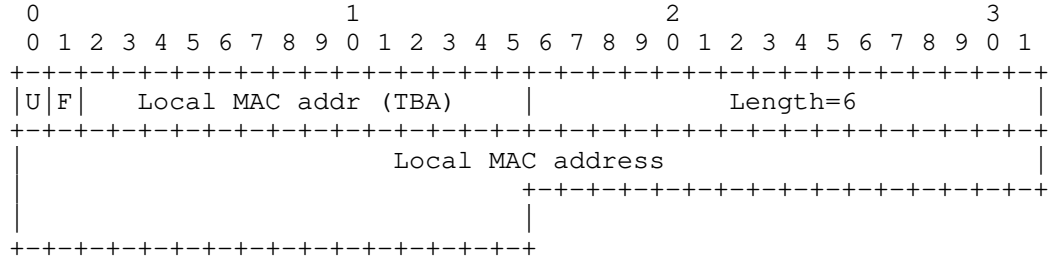
- 00 indicates payload is any protocol encapsulated in an Ethernet frame
- 01 indicates payload is IP
- 10 indicates payload is MPLS

This CW is also applicable to IP-only-PPW-EIM, MPLS-only-PPW-EIM and IPMPLS-only-PPW-EIM.

5.6. Signaling extensions

New values for the "PW type" field should be defined for the pseudowire encapsulations as "Packet - Efficient IP/MPLS", "Packet - IP only Efficient IP/MPLS", "Packet - MPLS only Efficient IP/MPLS", "Packet - IPMPLS only Efficient IP/MPLS" (values to be allocated by IANA).

An LDP optional parameter TLV "Local MAC Address" may be used to indicate the local MAC address to the remote peer. This TLV should be used in the LDP Notification message. The MAC address may have been configured or dynamically learnt. The format of the Local MAC address TLV is:



U bit: Unknown bit. This bit MUST be set to 1. If the MAC address

format is not understood, then the TLV is not understood and MUST be ignored.

F bit: Forward bit. This bit MUST be set to 1. In a MS-PW the S-PE should not interpret this TLV and it MUST be forwarded.

5.7. Implementation considerations

It is worthwhile noting that IP-only-PPW-EIM without the CW has an encapsulation format similar to that used in L3VPN. Also, MPLS-only-PPW-EIM without the CW has a packet format similar to that of a MPLS-TP LSP that has MPLS as a client. The action of pop and forward of the packet is in-line with the MPLS architecture. The capability to handle these formats should exist in most of the currently used hardware. The PPW-EIM with CW, has a format that is in line with the format in [PWE3-CW] and existing hardware should be capable of handling it. It is important to note that even with the GRE encapsulation, the PE does not have to do any of the typical GRE processing such as IP lookups. A capability to match a few nibbles/bytes in the header is sufficient to correctly identify and process the packet. Alternatively, an implementation may make CW mandatory for PPW-EIM, in which case the GRE encapsulation is not needed.

6. PSN MTU requirements

The MPLS PSN MUST be configured with an MTU that is large enough to transport a maximum-sized Ethernet frame that has been encapsulated with a control word, a flow label (if ECMP is desired), a pseudowire demultiplexer, and a tunnel encapsulation. With MPLS used as the tunneling protocol, for example, this is likely to be 12 or 16 bytes greater than the largest frame size. The methodology described in [PWE3-FRAG] MAY be used to fragment encapsulated frames that exceed the PSN MTU. However, if [PWE3-FRAG] is not used and if the ingress router determines that an encapsulated layer 2 PDU exceeds the MTU of the PSN tunnel through which it must be sent, the PDU MUST be dropped.

Note that the benefits associated with [FAT-PW] can be recognized in PPW-EIM for IP/MPLS packets without adding the flow-label, if ECMP is done by looking for an IP packet beyond the MPLS label stack when the PPW is setup without a control-word. This also reduces the MTU difference to only 8 bytes for IP/MPLS packets since both the control-word and the flow-label are not needed. In the scenario where the EVC is [802.1q] and the PE's interface into the PSN is Ethernet but not virtualized, the MTU difference is further reduced to 4. For the extreme case where PSN tunnel is a MPLS LSP with a single hop and has PHP, there is no difference in the MTU. Alternately, if the EVC

has two or more tags (similar to [802.1ad]) no fragmentation is needed for IP/MPLS packets even if the PSN tunnel LSP has multiple hops and there is no PHP.

7. Security Considerations

The security considerations in [PWE3-ETH] are applicable to this document.

8. IANA Considerations

IANA needs to allocate values for the following:

1. 'PW Type' field for "Packet - Efficient IP/MPLS", "Packet - IP only Efficient IP/MPLS", "Packet - MPLS only Efficient IP/MPLS" and "Packet - IPMPLS only Efficient IP/MPLS". Recommend next available values 0x0020, 0x0021, 0x0022 and 0x0023.

2. LDP 'TLV type' for 'Local MAC address'. Recommend available value 0x0405.

9. Conclusion

PPW-EIM has the following useful advantages:

1. Reduces the number of bytes on the wire. This translates into a significant reduction in bandwidth (as a percentage of packet size) for smaller packets.

2. Reduces the possibility of fragmentation (and reassembly) of jumbo IP/MPLS packets. This improves the throughput of the network.

3. Helps multi-layer networks by reducing the overhead required to stack each layer. This also reduces the possibility of fragmentation for jumbo packets in such networks.

4. Utilizes ECMP based on IP, a capability that exists in many current implementations.

5. Reduces the requirement to implement [FAT-PW] by taking advantage of existing implementations of ECMP based on IP.

6. Makes ECMP more efficient in multi-layer networks by enabling existing implementations (at any layer) to examine the label stack through all higher layers. In addition it enables existing implementations (at any layer) to easily examine the end-host's IP packet and simplifies deep-packet-

inspection/flow-based applications (including ECMP).

10. References

10.1. Normative References

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

[GRE] Farinacci, D., et al, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

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

[PWE3-CW] Bryant, S., et al, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.

[PWE3-FRAG] Malis, A., et al, "Pseudowire Emulation Edge-to-Edge (PWE3) Fragmentation and Reassembly", RFC 4623, August 2006.

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

10.2. Informative References

[ARP] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.

[PXY-ARP] Carl-Mitchell, S., et al, "Using ARP to Implement Transparent Subnet Gateways", RFC 1027, October 1987.

[ISIS] International Organization for Standardization, "Intermediate system to intermediate system intra-domain-routing routine information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO Standard 10589, 1992.

[RFC2474] Nichols, K., et al, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[PWE3-ETH] Martini, L., et al, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[FAT-PW] Bryant, S., et al, "Flow Aware Transport of Pseudowires over an MPLS PSN ", draft-ietf-pwe3-fat-pw-05 (Work in progress), October 2010.

[802.1q] "Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2005, 2005.

[802.1ad] "Virtual Bridged Local Area Networks - Amendment 4: Provider Bridges", IEEE Std 802.1ad-2005, 2005.

[LLDP] "IEEE Standard for Local and Metropolitan Area Networks - Station and Media Access Control Connectivity Discovery", IEEE Std 802.1AB-2005, 2005.

[MS-PW-ARCH] Bocci, M., et al, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.

[CAIDA-PKT-SIZE] CAIDA, "Packet size distribution comparison between Internet links in 1998 and 2008", http://www.caida.org/research/traffic-analysis/pkt_size_distribution/graphs.xml

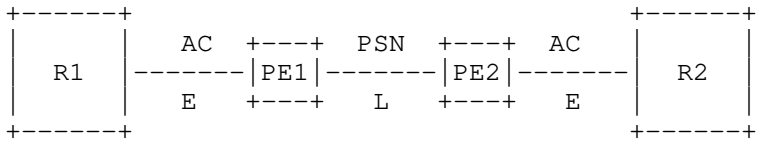
11. Acknowledgments

The authors would like to thank Joel Halpern, Loa Andersson, Andy Malis, Stewart Bryant and Edwin Mallette for their comments.

Appendix A: Example

Two examples are provided, one each for the two cases of the reference model described in section 4.

A.1. PWE3-ETH-EVC to connect routers



R1, R2 - IP routers
PE1, PE2 - PPW(PPW-EIM) capable PEs
AC - Attachment Circuit
E - Ethernet Frame, L - MPLS packet

Figure 7 Router inter-connect using PPW

R1 has an p2p IP interface to R2. This interface is created on VLAN 5 and runs ISIS level-2 ([ISIS]) as a routing protocol.

MAC addr - R1: 00-01-02-03-04-05, R2: 10-11-12-13-14-15
IP address - R1: 198.0.2.1/24, R2: 198.0.2.2/24

The VLAN 5 is emulated with a PPW (using encapsulation PPW-EIM) from PE1 to PE2 for EVC 5. Neither a control-word nor a flow-label is used on the PPW. PE2 has allocated a MPLS label 0x4321 as the PW demultiplexer. The PPW is encapsulated in a MPLS PSN and the PSN tunnel is a 1-hop LSP tunnel from PE1 to PE2 setup with PHP.

Using a typical encapsulation on an Ethernet port for an ISIS protocol packet, the level-2 LAN ISIS hello packet (LAN-IIH) from R1 to R2 is formatted by R1 into an ethernet frame E as shown below:

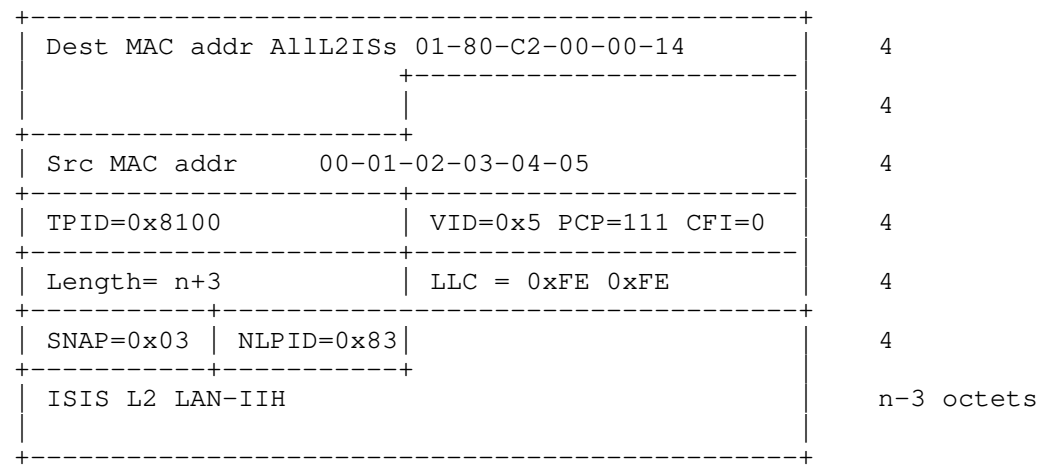


Figure 8 ISIS L2 LAN-IIH from R1 to R2 on AC

When the IIH is carried over the PPW it is encapsulated by PE1 as shown below:

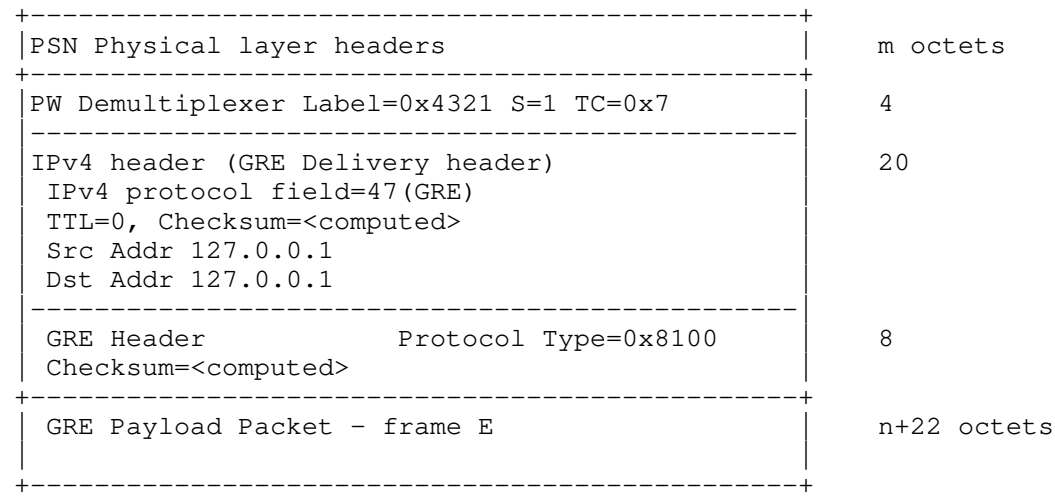


Figure 9 ISIS L2 LAN-IIH from R1 to R2 on PPW-EIM

A unicast IP packet routed by R1 that has 198.0.2.2 as next-hop is formatted by R1 as shown below:

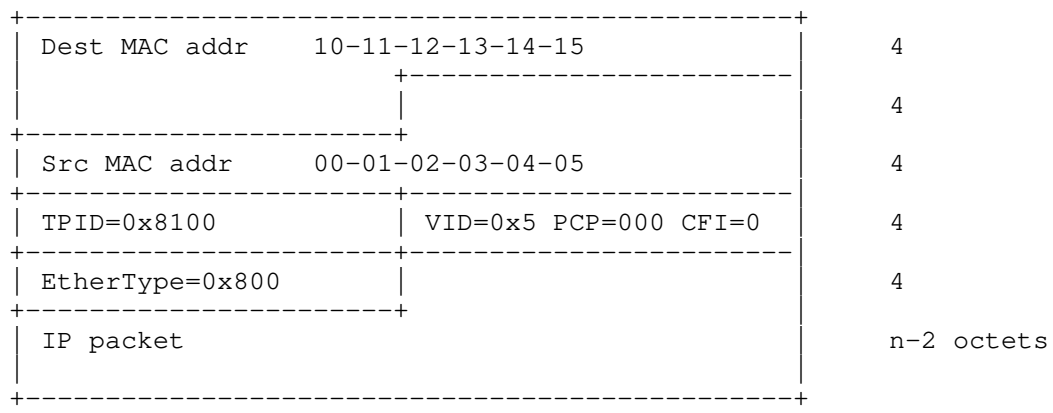


Figure 10 IP packet from R1 to R2 on AC

When this IP packet is carried over the PPW it is encapsulated by PE1 as shown below:

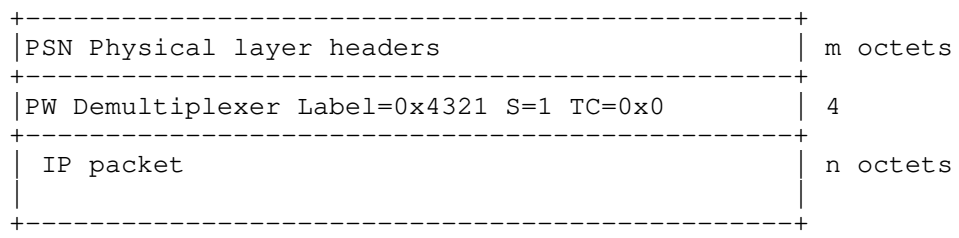
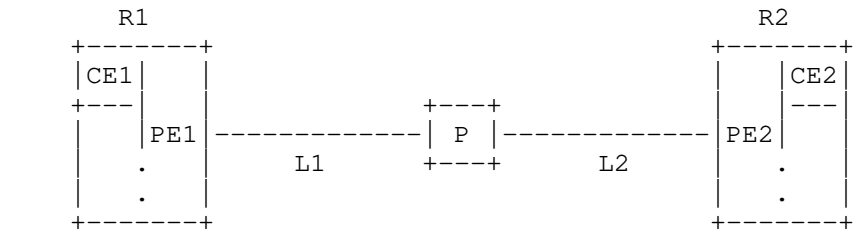


Figure 11 IP packet from R1 to R2 on PPW-EIM

A.2. CE co-existing with PE - interconnect



R1, R2 - IP/MPLS routers with co-existing PE and CE
PE1, PE2 - PPW(PPW-EIM) capable PEs
CE1, CE2 - IP/MPLS routers with a p2p IP/MPLS interface
P - MPLS P router
L1, L2 - MPLS packets

Figure 12 CE interconnect when co-existing with PE

CE1 has a p2p unnumbered IP interface to CE2. This interface runs ISIS level-2 as a routing protocol.

The IP interface is emulated with a PPW (using encapsulation PPW-EIM) from PE1 to PE2. Neither a control-word nor a flow-label is used on the PPW. PE2 has allocated a MPLS label 0x4321 as the PW demultiplexer. The PPW is encapsulated in a MPLS PSN tunnel that is a 2-hop bi-directional LSP TE tunnel from PE1 to PE2 setup without PHP.

The level-2 p2p ISIS hello packet (IIH) from CE1 to CE2 is encapsulated by PE1 as shown below:

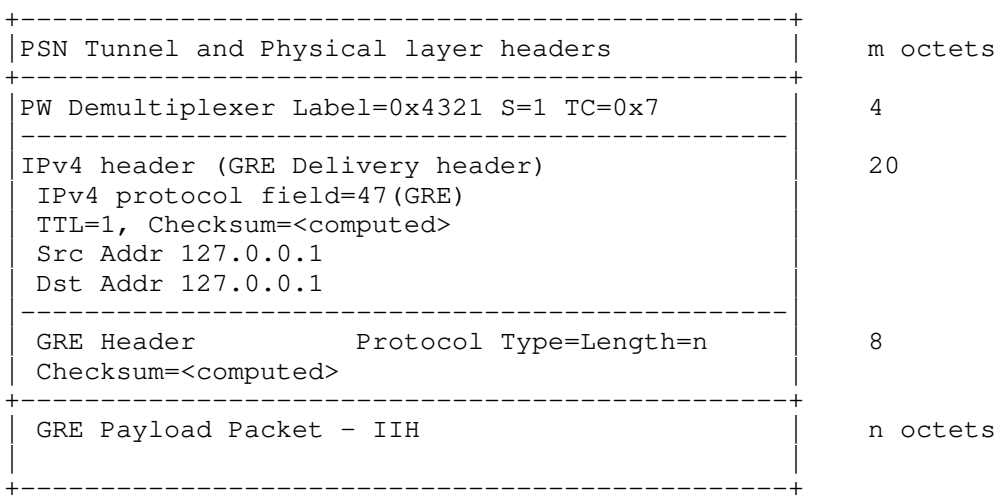


Figure 13 ISIS IIH from CE1 to CE2 on PPW-EIM

An IP packet routed by CE1 that has the unnumbered interface to CE2 as the next-hop is encapsulated by PE1 as shown below:

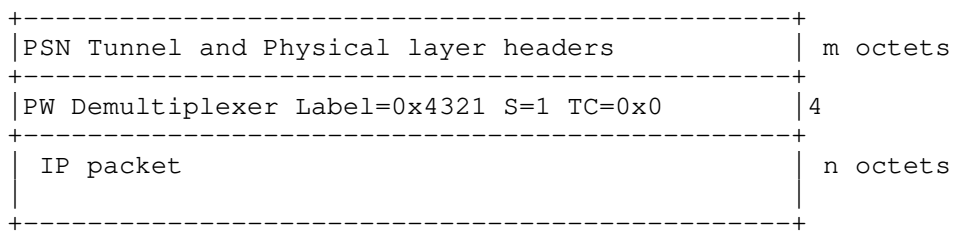


Figure 14 IP packet from CE1 to CE2 on PPW-EIM

An MPLS packet switched by CE1 that has the unnumbered interface to CE2 as the next-hop is encapsulated by PE1 as shown below:

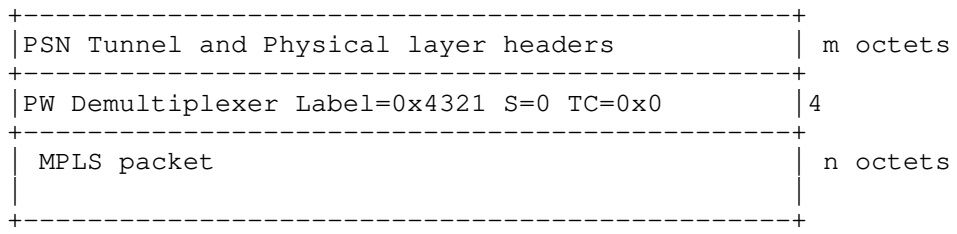


Figure 15 MPLS packet from R1 to R2 on PPW-EIM

Authors' Addresses

Sriganesh Kini
Ericsson
300 Holger Way, San Jose, CA 95134
EMail: sriganesh.kini@ericsson.com

David Sinicrope
Ericsson
8001 Development Dr, Research Triangle Park, NC 27709
EMail: david.sinicrope@ericsson.com

Internet Engineering Task Force
Internet Draft
Intended status: Standards Track
Expires: January 2012

Luca Martini
George Swallow
Cisco

Elisa Bellagamba
Ericsson

July 7, 2011

MPLS LSP PW status refresh reduction for Static Pseudowires

draft-martini-pwe3-status-aggregation-protocol-03.txt

Status of this Memo

Abstract

This document describes a method for generating an aggregated pseudowire status message on Multi-Protocol Label Switching (MPLS) network Label Switched Path (LSP).

The method for transmitting the pseudowire (PW) status information is not new, however this protocol extension allows a Service Provider (SP) to reliably monitor the individual PW status while not overwhelming the network of multiple periodic status messages. This

is achieved by sending a single cumulative status message for all the PWs grouped in the same LSP.

Table of Contents

1. Introduction

When PWs use a Multi Protocol Label Switched (MPLS) network as the Packet Switched Network (PSN), they are setup according to [RFC4447] static configuration mode and the PW status information is propagated using the method described in [PW-STATUS]. There are 2 basic modes of operation described in [PW-STATUS] section 5.3: Periodic retransmission of non-zero status messages, and a simple acknowledge of PW status (sec 5.3.1 of [PW-STATUS]). The LSP level protocol described below applies to the case when PW status is acknowledged immediately with a requested refresh value of zero (no refresh). In this case the PW status refresh reduction protocol is necessary for several reasons, such as:

- i. Greatly increase the scalability of the PW status protocol by reducing the amount of messages that a PE needs to periodically send to it's neighbors.
- ii. Detect a remote PE restart.
- iii. If the local state is lost for some reason, the PE needs to be able to request a status refresh reduction from the remote PE
- iv. Optionally detect a remote PE provisioning change.

1.1. 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].

1.2. Terminology

FEC: Forwarding Equivalence Class

LDP: Label Distribution Protocol

LSP: Label Switching Path

MS-PW: Multi-Segment Pseudowire

PE: Provider Edge

PW: Pseudowire

SS-PW: Single-Segment Pseudowire

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

T-PE: Terminating Provider Edge Node of MS-PW

1.3. Notational Conventions in Backus-Naur Form

All multiple-word atomic identifiers use underscores (_) between the words to join the words. Many of the identifiers are composed of a concatenation of other identifiers. These are expressed using Backus-Naur Form (using double-colon - "::" - notation).

Where the same identifier type is used multiple times in a concatenation, they are qualified by a prefix joined to the identifier by a dash (-). For example Src-Node_ID is the Node_ID of a node referred to as Src (where "Src" is short for "source" in this example).

The notation does not define an implicit ordering of the information elements involved in a concatenated identifier.

2. PW status refresh reduction protocol

PW status refresh reduction protocol consists of a simple message that is sent at the LSP level using the MPLS Generic Associated Channel.

A PE using the PW status refresh reduction protocol MUST send the PW status refresh reduction Message as soon as a PW is configured on a particular LSP. The message is then re-transmitted at a locally configured interval indicated in the refresh timer field. If no acknowledgment is received, the protocol does not reach active state, and the PE SHOULD NOT send any PW status messages with a refresh timer of zero as described in [PW-STATUS] section 5.3.1.

It is worth noting that no relationship is existing between the locally configured timer for the refresh reduction protocol and the PW individual status refresh timers.

2.1. Protocol states

The protocol can be in 3 possible states: INACTIVE, STARTUP, and ACTIVE.

2.1.1. INACTIVE

This state is entered when the protocol is turned off. This state is also entered if all PW on a specific LSP are unprovisioned, or the feature is unprovisioned.

2.1.2. STARTUP

In this state the PE transmits periodic PW status refresh reduction messages, with the Ack Session ID set to 0. The PE remains in this state until a PW status refresh message is received with the correct local session ID in the Ack Session ID Field. This state can be exited to the ACTIVE or INACTIVE state.

2.1.3. ACTIVE

This state is entered once the PE receives a PW status refresh reduction message with the correct local session ID in the Ack Session ID Field within 3.5 times the refresh timer field value of the last PW status refresh reduction message transmitted. This state is immediately exited as follows:

- -i. A valid PW status refresh reduction message is not received within 3.5 times the current refresh timer field value. (assuming a timer transition procedure is not in progress) New state: STARTUP
- -ii. A PW status refresh reduction message is received with the wrong, or a zero, Ack Session ID field value. New state: STARTUP
- -iii. All PWs using the particular LSP are unprovisioned, or the protocol is disabled. New state: INACTIVE

2.2. Timer value change transition procedure

If a PE needs to change the refresh timer value field while the PW refresh reduction protocol is in the ACTIVE state, the following procedure must be followed:

- -i. A PW status refresh reduction message is transmitted with the new timer value.
- -ii. If the new value is greater then the original one the PE will operate on the new timer value immediately.
- -iii. If the new value is smaller then the original one, the PE will operate according to the original timer value for a period 3.5 times the original timer value, or until the first valid PW status refresh reduction message is received.

A PE receiving a PW status refresh reduction message with a new timer value, will immediately transmit an acknowledge PW status refresh reduction message, and start operating according to the new timer value.

3. PW status refresh reduction procedure

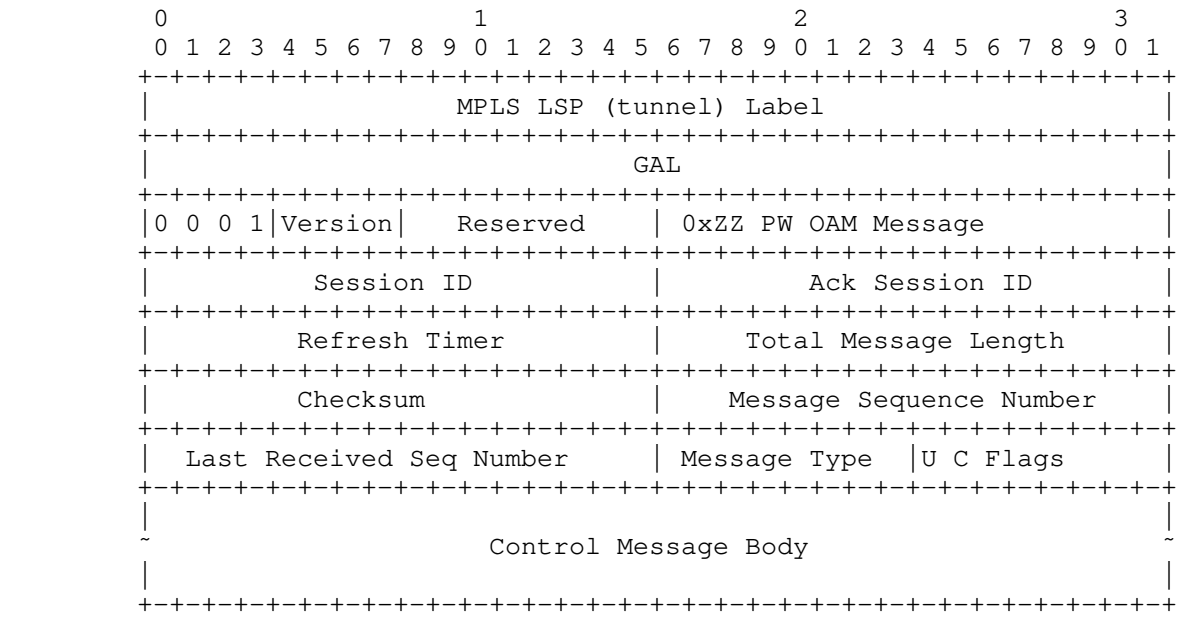
When the refresh reduction protocol, on a particular LSP, is in the ACTIVE state, the PE can send all PW status messages, for PWs on that LSP, with a refresh timer value of zero. This greatly decreases the amount of messages that the PE needs to transmit to the remote PE because once the PW status message for a particular PW is acknowledged, further repetitions of that message are no longer necessary.

To further mitigate the amount of possible messages when an LSP starts forwarding traffic, care should be taken to permit the PW refresh reduction protocol to reach the ACTIVE state quickly, and before the the first PW status refresh timer expires. This can be achieved by using a PW status refresh reduction Message refresh timer value that is much smaller then the PW status message refresh timer value in use. (sec 5.3.1 of [PW-STATUS])

If the refresh reduction protocol session is terminated by entering the INACTIVE or STARTUP states, the PE MUST immediately re-send all the previously sent PW status messages for that particular LSP for which the session terminated. In this case the refresh timer value MUST NOT be set to zero, and MUST be set according to the local policy of the PE router.

4. PW status refresh reduction Message Encoding

The packet containing the refresh reduction message is encoded as follows: (omitting link layer information)



This message contains the following fields:

* PW OAM Message.

  This field indicates the generic associated channel type in the GACH header as defined in [RFC5586].

  Note: Channel type 0xZZ pending IANA allocation.

* Session ID

  A non-zero, locally selected session number that is not preserved if the local PE restarts.

  In order to get a locally unique session ID, the recommended choice is to perform a CRC-16 giving as input the following data

  |Y|Y|M|M|D|D|H|H|M|M|S|S|L|L|L|

  Where: YY: are the decimal two last digit of the current year MM: are the decimal two digit of the current month DD: are the decimal two digit of the current day HHMMSSLLL: are the decimal digits of the current time expressed in (hour, minutes, seconds, milliseconds)

* Ack Session ID

  The Acknowledgment Session ID received from the remote PE.

* Refresh Timer.

  A non zero unsigned 16 bit integer value greater or equal to 10, in milliseconds, that indicates the desired refresh interval. The default value of 30000 is RECOMENDED.

* Total Message Length

  Total length in octets of the Checksum, Message Type, Flags, Message Sequence Number, and control message body. A value of zero means that no control message is present, and therefore that no Checksum, and following fields are present either.

* Checksum

  A 16 bit field containing the one's complement of the one's complement sum of the entire message (including the GACH header), with the checksum field replaced by zero for the purpose of computing the checksum. An all-zero value means that no checksum was transmitted. Note that when the checksum is not computed, the header of the bundle message will not be covered by any checksum.

* Message Sequence Number

  A unsigned 16 bit integer number that is started from 1 when the protocol enters ACTIVE state. The sequence numbers wraps back to 1 when the maximum value is reached. The value of zero is reserved and MUST NOT be used.

* Last Received Message Sequence Number

  The sequence number of the last message received. In no message has yet been received during this session, this field is set to zero.

* Message Type

  The Type of the control message that follows. Control message types are allocated in this document, and by IANA.

* (U) Unknown flag bit.

  Upon receipt of an unknown message, if U is clear (=0), the keepalive session MUST be terminated by entering STARTUP state;

if U is set (=1), the unknown message MUST be acknowledge and silently ignored and the following messages, if any, processed as if the unknown message did not exist.

* (C) Configuration flag bit. The C Bit is used to signal the end of PW configuration transmission. If it is set, the sending PE has finished sending all it's current configuration information.

* Flags (Reserved)

  7 bits of flags reserved for future use, they MUST be set to 0 on transmission, and ignored on reception.

* Control Message Body

  The Control Message body is defined in a section below, and is specific to the type of message.

It should be noted that the Checksum, Message Sequence Number, Last Received Message Sequence Number, Message Type, Flags, and control message body are OPTIONAL.

5. PW status refresh reduction Control Messages

PW status refresh reduction Control messages consist of the Checksum, Message Sequence Number, Last Received Message Sequence Number, Message Type, Flags, and control message body.

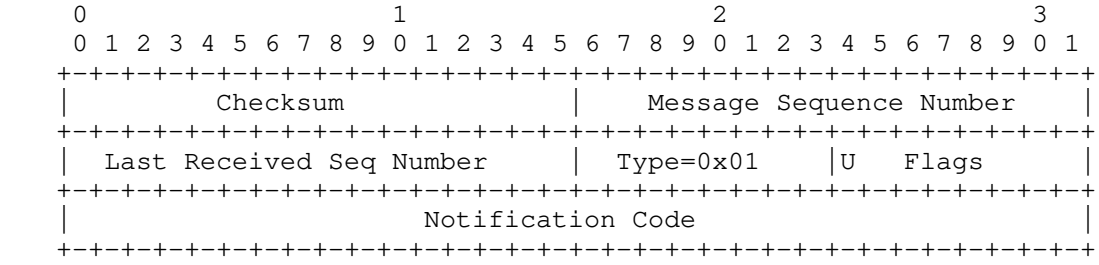
When there is the need to send a PW status refresh reduction Control Messages, the system can attach it to a scheduled PW status refresh reduction or send one ahead of time. In any case PW status refresh reduction Control Messages always piggy back on normal messages.

There can only be one control message construct per PW status refresh reduction Message. If the U bit is set, and a PE receiving the PW status refresh reduction Message does not understand the control message, the control message MUST be silently ignored. However the control message sequence number MUST still be acknowledged by sending a null message back with the appropriate value in the Last Message Received Field. If a control message is not acknowledged, after 3.5 times the value of the Refresh Timer, a fatal notification "unacknowledged control message" MUST be sent, and the PW refresh reduction session MUST be terminated.

If a PE does not want or need to send a control message, the Checksum, and all following fields MUST NOT be sent, and the Total Message Length field is then set to zero.

5.0.1. Notification message

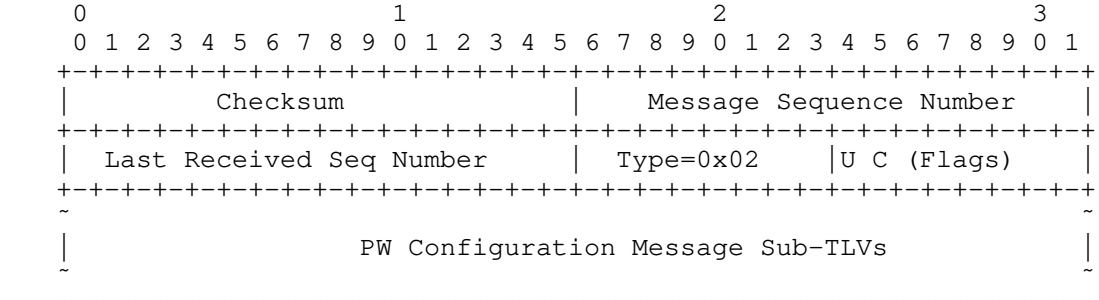
The most common use of the Notification Message is to acknowledge the reception of a message by indicating the received message sequence number in the "Last Received Sequence Number" field. The notification message is encoded as follows:



The message type is set to 0x01, and the U bit is treated as described in the above section. The Notification Codes are a 32 bit quantity assigned by IANA. (see IANA consideration section) Notification codes are either considered "Error codes" or simple notifications. If the Notification code is an Error code as indicated in the IANA allocation registry, the keepalive session MUST be terminated by entering STARTUP state.

5.0.2. PW Configuration Message

The PW status refresh reduction TLVs are informational TLVs, that allow the remote PE to verify certain provisioning information. This message contain a series of sub-TLVs in no particular order, that contain PW and LSP configuration information. The message has no preset length limit, however its total length will be limited by the transport network Maximum Transmit Unit (MTU).



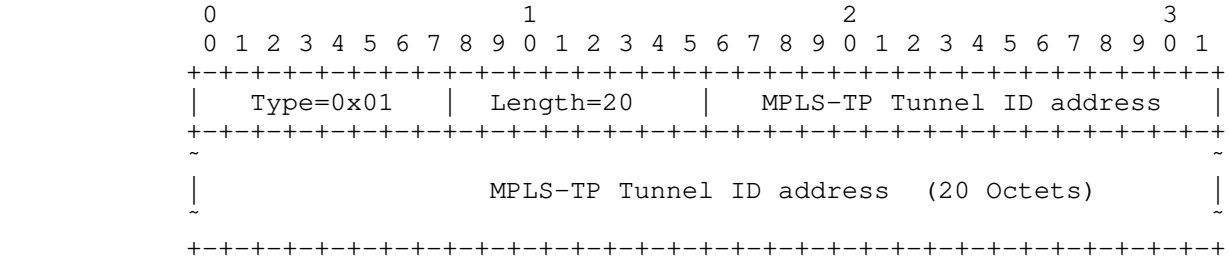
The PW Configuration Message type is set to 0x02. For this message the U-bit is set to 1 as processing of these messages is OPTIONAL.

The C Bit is used to signal the end of PW configuration transmission. If it is set, the sending PE has finished sending all its current configuration information. The PE transmitting the configuration MUST set the C bit on the last PW configuration message when all current PW configuration has been sent.

5.0.2.1. MPLS-TP Tunnel ID

This TLV contains the address of the MPLS-TP tunnel ID. When the configuration message is used for a particular keepalive session the MPLS-TP Tunnel ID sub-TLV MUST be sent at least once.

The MPLS-TP Tunnel ID address is encoded as follows:



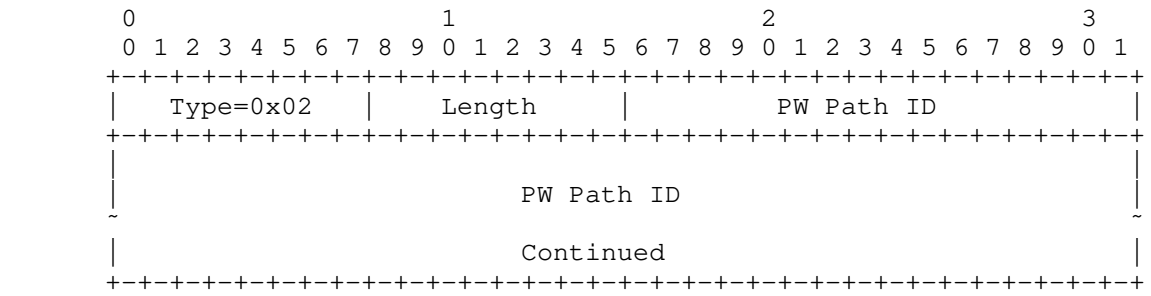
The MPLS-TP point to point tunnel ID is defined in [IDENTIFIER] as follows:

Src-Global_Node_ID::Src-Tunnel_Num::Dst-Global_Node_ID::Dst-Tunnel_Num

Note that a single address is enough to identify the tunnel, and the source end of the message.

5.0.2.2. PW ID configured List

This OPTIONAL TLV contains a list of the provisioned PWs on the LSP.

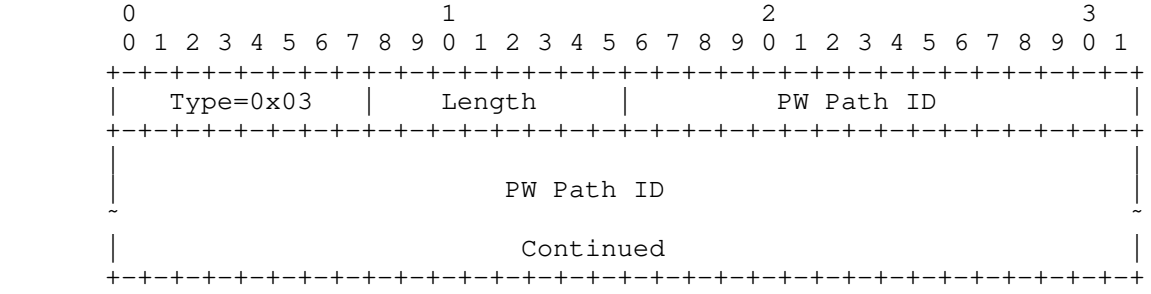


The PW Path ID is a 32 octet pseudowire path identifier specified in [IDENTIFIER] as follows: AGI::Src-Global_ID::Src-Node_ID::Src-AC_ID:: Dst-Global_ID::Dst-Node_ID::Dst-AC_ID

The number of PW Path IDs in the TLV will be inferred by the length of the TLV up to a maximum of 8. The procedure for processing this TLV will be described in a section below.

5.0.2.3. PW ID unconfigured List

This OPTIONAL TLV contains a list of the PWs that have been unprovisioned on the LSP. Note that it is a fatal session error to send the same PW address in both the configured list TLV , and the unconfigured list TLV in the same configuration message.



The PW Path ID is a 32 octet pseudowire path identifier specified in [IDENTIFIER] as follows: AGI::Src-Global_ID::Src-Node_ID::Src-AC_ID:: Dst-Global_ID::Dst-Node_ID::Dst-AC_ID

The number of PW Path IDs in the TLV will be inferred by the length of the TLV up to a maximum of 8.

6. PW provisioning verification procedure

This procedure and the advertisement of the PW configuration message are OPTIONAL.

A PE that desires to use the PW configuration message to verify the configuration of PWs on a particular LSP, should advertise its PW configuration to the remote PE on LSPs that have active keepalive sessions. When a PE receives PW configuration information using this protocol and it not supporting or not willing to use the information, it MUST acknowledge all the PW configuration messages with a notification of "PW configuration not supported". In this case, the information in the control messages is silently ignored. If a PE receives such a notification it should stop sending PW configuration control messages for the duration of the PW refresh reduction keepalive session.

If PW configuration information is received, it is used to verify the accuracy of the local configuration information against the remote PE's configuration information. If a configuration mismatch is detected, where a particular PW is configured locally but not on the remote PE, the following action SHOULD be taken:

-i. The local PW MUST be considered in "Not Forwarding" State.

-ii. The PW Attachment Circuit status is set to reflect the PW fault.

-iii. An Alarm MAY be raised to a network management system.

6.1. PW ID List advertising and processing

When configuration messages are advertised along a particular LSP, the PE sending the messages needs to check point the configuration information sent by setting the C bit when all currently known configuration information has been sent. This process allows the receiving PE to immediately proceed to verify all the currently configured PWs on that LSP, eliminating the need for a long waiting period.

If a new PW is added to a particular LSP, the PE MUST place the configuration verification of this PW on hold for a period of at least 10 seconds. This is necessary to prevent false positive events of mis-configuration due to the ends of the PW being slightly out of sync.

7. Security Considerations

Section to be completed in a later version of the document.

8. IANA Considerations

8.1. PW Status Refresh Reduction Message Types

IANA needs to set up a registry of "PW status refresh reduction Control Messages". These are 8-bit values. Type value 1 through 2 are defined in this document. Type values 3 through 64 are to be assigned by IANA using the "Expert Review" policy defined in RFC5226. Type values 65 through 127, 0 and 255 are to be allocated using the IETF consensus policy defined in [RFC5226]. Type values 128 through 254 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC5226.

The Type Values are assigned as follows:

Type	Message Description
0x01	Notification message
0x02	PW Configuration Message

8.2. PW Configuration Message Sub-TLVs

IANA needs to set up a registry of "PW status refresh reduction Configuration Message Sub-TLVs". These are 8-bit values. Type value 1 through 2 are defined in this document. Type values 3 through 64 are to be assigned by IANA using the "Expert Review" policy defined in RFC5226. Type values 65 through 127, 0 and 255 are to be allocated using the IETF consensus policy defined in [RFC5226]. Type values 128 through 254 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC5226.

The Type Values are assigned as follows:

sub-TLV type	Description
0x01	MPLS-TP Tunnel ID address.
0x02	PW ID configured List.
0x03	PW ID unconfigured List.

8.3. PW Status Refresh Reduction Notification Codes

IANA needs to set up a registry of "PW status refresh reduction Notification Codes". These are 32-bit values. Type value 1 through 7 are defined in this document. Type values 8 through 65536 are to be assigned by IANA using the "Expert Review" policy defined in RFC5226. Type values 65536 through 134,217,728, 0 and 4,294,967,295 are to be allocated using the IETF consensus policy defined in [RFC5226]. Type values 134,217,729 through 4,294,967,294 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC5226.

The Type Values are assigned as follows:

Code	Error?	Description
0x00000000	No	Null Notification.
0x00000001	No	PW configuration rejected.
0x00000002	Yes	PW Configuration TLV conflict.
0x00000003	No	Unknown TLV (U-bit=1)
0x00000004	Yes	Unknown TLV (U-bit=0)
0x00000005	No	Unknown Message Type
0x00000006	No	PW configuration not supported.
0x00000007	Yes	Unacknowledged control message.

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., rfc4447 April 2006.

[PW-STATUS] L. Martini, G. Swallow, G. Heron, M. Bocci "Pseudowire Status for Static Pseudowires", draft-ietf-pwe3-static-pw-status-06.txt, (work in progress), July 2011

[IDENTIFIER] M. Bocci, G. Swallow, E. Gray "MPLS-TP Identifiers" draft-ietf-mpls-tp-identifiers-06.txt, IETF Work in Progress, june 2011

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations section in RFCs", BCP 26, RFC 5226, May 2008

9.2. Informative References

[RFC5586] M. Bocci, Ed., M. Vigoureux, Ed., S. Bryant, Ed., "MPLS Generic Associated Channel", rfc5586, June 2009

10. Author's Addresses

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

George Swallow
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough, Massachusetts 01719
United States
e-mail: swallow@cisco.com

Elisa Bellagamba
Ericsson EAB
Torshamnsgatan 48
16480, Stockholm
Sweden
e-mail: elisa.bellagamba@ericsson.com

Copyright Notice

Expiration Date: January 2012

Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 24, 2011

T. Nadeau
CA Technologies

L. Martini
Cisco Systems, Inc.

June 24, 2011

A Unified Control Channel for Pseudowires
draft-nadeau-pwe3-vccv-2-02.txt

Abstract

This document describes a unified mode for Virtual Circuit Connectivity Verification (VCCV), which provides a control channel that is associated with a pseudowire (PW). VCCV applies to all supported access circuit and transport types currently defined for PWs, as well as those being transported by The MPLS Transport Profile. This new mode is intended to augment those described in RFC5085, but this document describes new rules requiring this mode to be used as the default/mandatory mode of operation for VCCV. The older types will remain optional.

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 September 6, 2011.

Copyright Notice

Table of Contents

1. Introduction

There is a need for fault detection and diagnostic mechanisms that can be used for end-to-end fault detection and diagnostics for a Pseudowire, as a means of determining the PW's true operational state. Operators have indicated in [RFC4377], [RFC3916]. that such a tool is required for PW operation and maintenance. To this end, the IETF's PWE3 Working Group defined The Virtual Circuit Connectivity Verification Protocol (VCCV) in [RFC5085]. Since then a number of interoperability issues have arisen with the protocol as it is defined.

The variety of VCCV options or "modes" have been created to support legacy hardware, the use of the control word in some cases, while in others not, among others. The difficulty of operating these different combinations of "modes" have been detailed in an implementation survey the PWE3 Working Group conducted. Many of the motivations of this survey are detailed in [MAN-CW]. This document

and the implementation survey concluded that operators have had difficulty deploying the protocol given the number of combinations and options for its use.

In addition to the implementation issues just described, the ITU-T and IETF have set out to enhance MPLS to make it suitable as an optical transport protocol. The requirements for this protocol are defined as the MPLS Transport Profile (MPLS-TP). The requirements for this protocol can be found in [RFC5654]. In order to support VCCV when an MPLS-TP PSN is in use, the GAL-ACH had to be created; this effectively resulted in another mode of operation.

This document seeks to simplify the modes of operation of VCCV down to a single mode of operation we refer to as type 4 for the moment. This mode simply defines two ways to run VCCV: 1) with a control word or 2) without a control word, but with a ACH encapsulation making it easy to handle all of the other cases handled by the other modes of VCCV. In either case, it will be mandatory to implement and use that mode, thus simplifying the implementation and operation of the protocol.

Figure 1 depicts the architecture of a pseudowire as defined in [RFC3985]. It further depicts where the VCCV control channel resides within this architecture, which will be discussed in detail shortly.

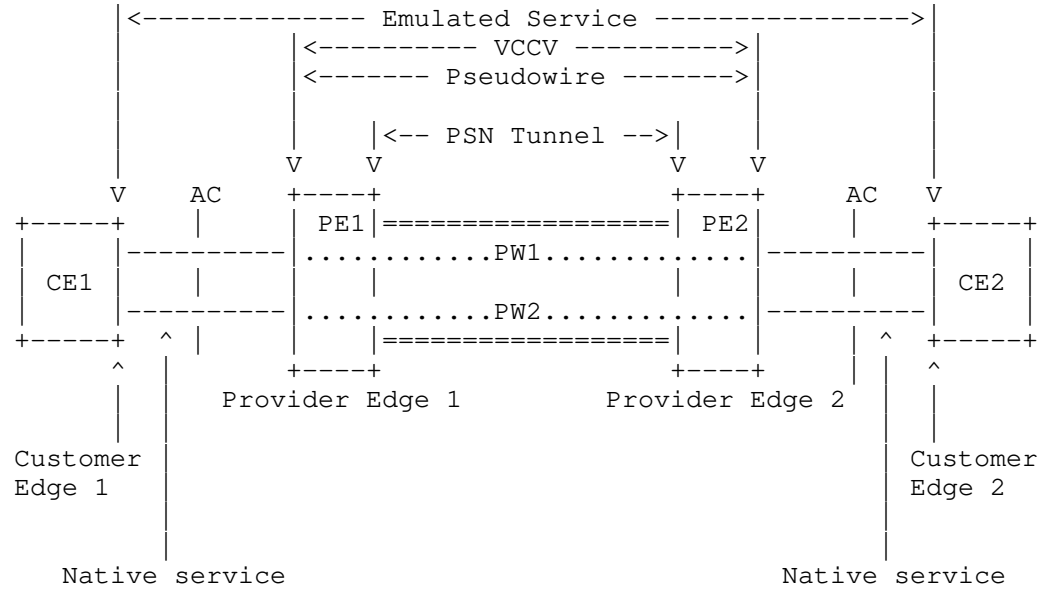


Figure 1: PWE3 VCCV Operation Reference Model

From Figure 1, Customer Edge (CE) routers CE1 and CE2 are attached to the emulated service via Attachment Circuits (ACs), and to each of the Provider Edge (PE) routers (PE1 and PE2, respectively). An AC can be a Frame Relay Data Link Connection Identifier (DLCI), an ATM Virtual Path Identifier / Virtual Channel Identifier (VPI/VCI), an Ethernet port, etc. The PE devices provide pseudowire emulation, enabling the CEs to communicate over the PSN. A pseudowire exists between these PEs traversing the provider network. VCCV provides several means of creating a control channel over the PW, between the PE routers that attach the PW.

Figure 2 depicts how the VCCV control channel is associated with the pseudowire protocol stack.

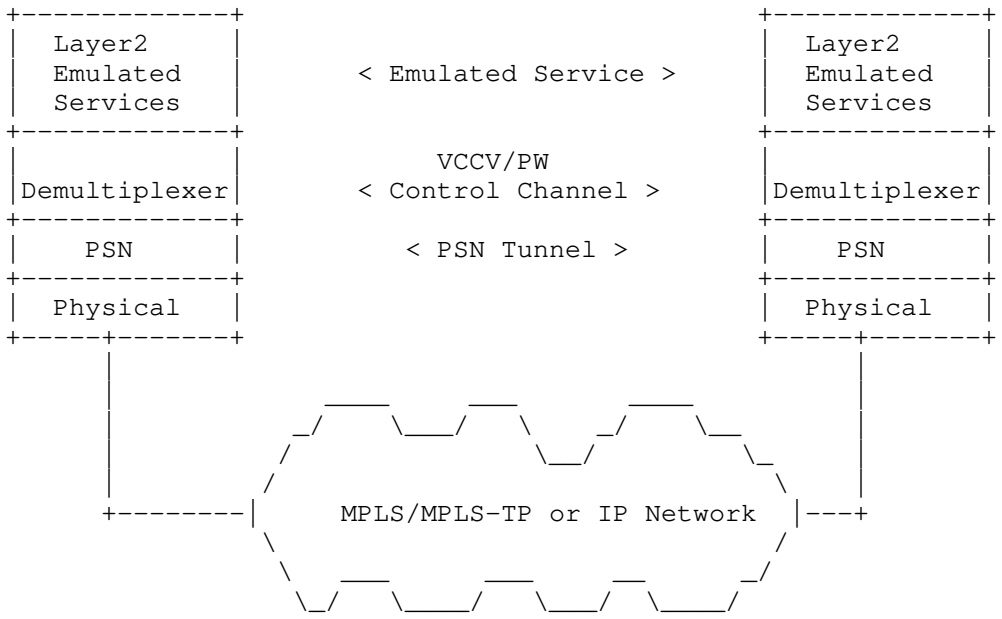


Figure 2: PWE3 Protocol Stack Reference Model including the VCCV Control Channel

VCCV messages are encapsulated using the PWE3 encapsulation as described in Sections 2 and 3, so that they are handled and processed in the same manner (or in some cases, a similar manner) as the PW PDUs for which they provide a control channel. These VCCV messages are exchanged only after the capability (expressed as two VCCV type spaces, namely the VCCV Control Channel and Connectivity Verification Types) and desire to exchange such traffic has been advertised between the PEs (see Sections 5.3 and 6.3), and VCCV types chosen.

1.2. Acronyms

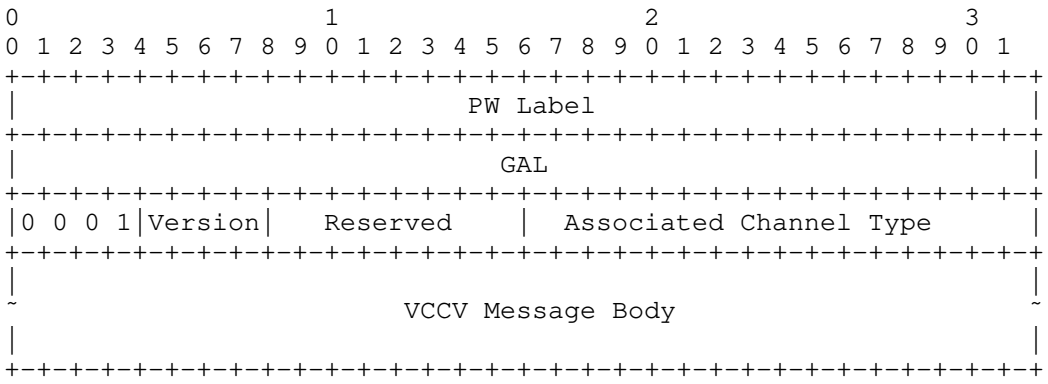
AC	Attachment Circuit [RFC3985].
AVP	Attribute Value Pair [RFC3931].
CC	Control Channel (used as CC Type).
CE	Customer Edge.
CV	Connectivity Verification (used as CV Type).
CW	Control Word [RFC3985].
L2SS	L2-Specific Sublayer [RFC3931].
LCCE	L2TP Control Connection Endpoint [RFC3931].
OAM	Operation and Maintenance.
PE	Provider Edge.
PSN	Packet Switched Network [RFC3985].
PW	Pseudowire [RFC3985].
PW-ACH	PW Associated Channel Header [RFC4385].
VCCV	Virtual Circuit Connectivity Verification [RFC5085].

2. VCCV Control Channel When The Control Word is Used

When the PWE3 Control Word is used to encapsulate pseudowire traffic, the rules described for encapsulating VCCV CC Type 1 as specified in section 9.5.1 [RFC6073] and section 5.1.1 of [RFC5085] MUST be used. In this case the advertised CC Type is 1, and Associated Channel Types of 21, 07, or 57 are allowed.

3. VCCV Control Channel When The Control Word is Not Used

When the PWE3 Control Word is not used a new CC Type 4 is defined as follows.



The PW Label must set the TTL field to 1. In the case of multi-segment pseudo-wires, the PW Label TTL MUST be set to the correct value to reach the intended destination PE as described in [RFC6073].

The GAL field MUST contain the reserved label as defined in [RFC5586].

The first nibble of the next field is set to 0001b to indicate an ACH associated with a pseudowire (see Section 5 of [RFC4385] and Section 3.6 of [RFC4446]) instead of PW data. The Version and the Reserved fields MUST be set to 0, and the Channel Type is set to 0x0021 for IPv4, 0x0057 for IPv6 payloads [RFC5085] or 0x0007 for BFD payloads [RFC5885].

The "VCCV Messag Body" field is defined based on the Associated Channel Type and defined therein.

4. VCCV Capability Advertisement

The capability advertisement MUST match that c-bit setting that is advertised in the PW FEC element. If the c-bit is set, indicating the use of the control word, type 1 MUST be advertised and type 4 MUST NOT be advertised. If the c-bit is not set, indicating that the control word is not in use, type 4 MUST be advertised, and type 1 MUST NOT be advertised.

A PE supporting Type 4 MAY advertise other CC types as defined in RFC5085. If the remote PE also supports Type 4, then Type 4 MUST be used superceding the Capability Advertisement Selection rules of section 7 from RFC5085. If a remote PE does not support Type 4, then the rules

from section 7 of RFC5085 apply. If a CW is in use, then Type 4 is not applicable, and therefore the normal capability advertisement selection rules of section 7 from RFC5085 apply.

4. IANA Considerations

4.1. VCCV Interface Parameters Sub-TLV

The VCCV Interface Parameters Sub-TLV codepoint is defined in [RFC4446]. IANA has created and will maintain registries for the CC Types and CV Types (bitmasks in the VCCV Parameter ID). The CC Type and CV Type new registries (see Sections 8.1.1 and 8.1.2, respectively) have been created in the Pseudo Wires Name Spaces, reachable from [IANA.pwe3-parameters]. The allocations must be done using the "IETF Consensus" policy defined in [RFC5226].

4.1.1. MPLS VCCV Control Channel (CC) Type 4

IANA is requested to augment the registry of "MPLS VCCV Control Channel Types" with the new type defined below. As defined in RFC5058, this new bitfield is to be assigned by IANA using the "IETF Consensus" policy defined in [RFC5226]. A VCCV Control Channel Type description and a reference to an RFC approved by the IESG are required for any assignment from this registry.

MPLS Control Channel (CC) Types:

Bit (Value)	Description
Bit 3 (0x08)	Type 4

The most significant (high order) bit is labeled Bit 7, and the least significant (low order) bit is labeled Bit 0, see parenthetical "Value".

5. Security Considerations

This document does not by itself raise any particular security considerations that differ from those described in RFC5085.

6. Acknowledgements

7. References

7.1. Normative References

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

[RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.

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

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

[RFC5586] Bocci, M., Ed., Vigoureux, M., Ed., and S. Bryant, Ed., "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5885] Nadeau, T., Ed., and C. Pignataro, Ed., "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.

[RFC5654] Niven-Jenkins, B., Brungard, D., and M. Betts, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009

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

12.2. Informative References

[IANA.l2tp-parameters]
Internet Assigned Numbers Authority, "Layer Two Tunneling Protocol "L2TP"", April 2007, <http://www.iana.org/assignments/l2tp-parameters>.

[IANA.pwe3-parameters]
Internet Assigned Numbers Authority, "Pseudo Wires Name Spaces", June 2007, <http://www.iana.org/assignments/pwe3-parameters>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC3916] Xiao, X., McPherson, D., and P. Pate, "Requirements for Pseudo-Wire Emulation Edge-to-Edge (PWE3)", RFC 3916, September 2004.

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

[RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.

[MAN-CW] Del Regno, N., Nadeau, T., Manral, V., Ward, D., "Mandatory Use of Control Word for PWE3 Encapsulations", "Work in progress", October 2010.

8. Authors' Addresses

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA

Email: thomas.nadeau@ca.com

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO, 80112 USA

EMail: lmartini@cisco.com

Network Working Group
Internet Draft
Category: Standard Track

R. Ram, Orckit-Corrigent
D. Cohn, Orckit-Corrigent
R. Key, Telstra
P. Agarwal, Broadcom

Expires: November 18, 2011

May 18, 2011

Extension to LDP-VPLS for E-Tree Using Two PW
draft-ram-l2vpn-ldp-vpls-etree-2pw-02.txt

Status of this Memo

Copyright Notice

Abstract

This document proposes a solution for Metro Ethernet Forum (MEF) Ethernet Tree (E-Tree) support in Virtual Private LAN Service using LDP Signaling (LDP-VPLS) [RFC4762]. The proposed solution is characterized by the use of two PWs between a pair of PEs. This solution is applicable for both VPLS and H-VPLS.

Table of Contents

1. Introduction

This document proposes a solution for Metro Ethernet Forum (MEF) Tree (E-Tree) support in Virtual Private LAN Service using LDP Signaling (LDP-VPLS) [RFC4762].

[Draft ETree VPLS Req] is used as requirement specification.

The proposed solution is characterized by the use of two PWs between a pair of PEs, which requires extension to the current VPLS standard [RFC4762].

This solution is applicable for both VPLS and H-VPLS.

The proposed solution is composed of three main components:

- Current standard LDP-VPLS [RFC4762]
- Extension to LDP-VPLS specified in this document
- PE local split horizon mechanism

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. The Problem

[Draft ETree VPLS Req] identifies the problem when there are two or more PEs with both Root AC and Leaf AC.

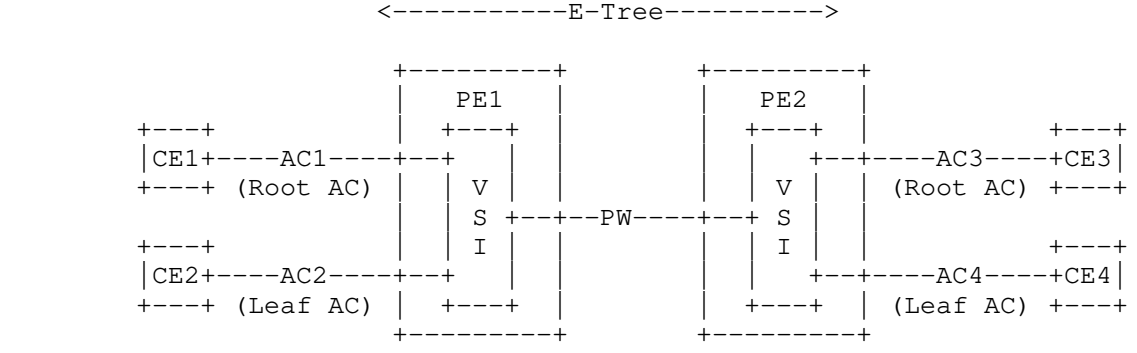


Figure 1: Problem Scenario for Leaf-to-Leaf Communication Restriction

When PE2 receives a frame from PE1 via the Ethernet PW,
- PE2 does not know whether the ingress AC is a Leaf AC or not
- PE2 does not have sufficient information to enforce the Leaf-to-Leaf communication restriction

4. The 2-PW Solution

A simple fix is to carry additional information with each frame on the PW, indicating whether the frame is originated from a Leaf AC or a Root AC on the ingress PE.

The proposed solution uses a pair of PWs to interconnect two VPLS PEs:

- o First PW is used for frames originated from Root ACs
- o Second PW is used for frames originated from Leaf ACs

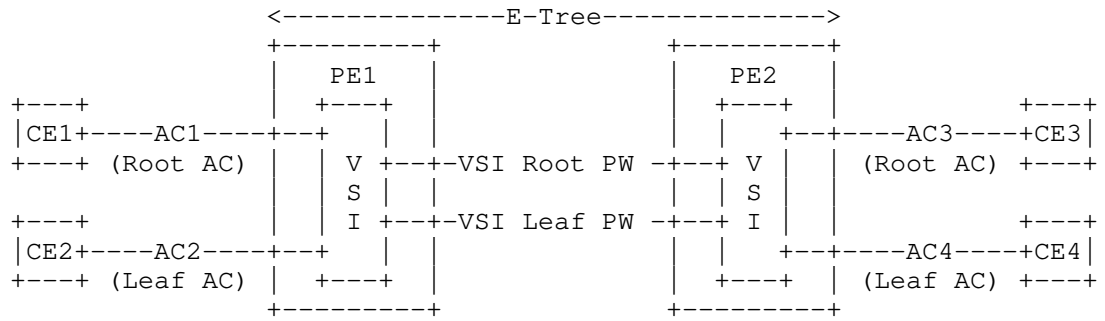


Figure 2: Two-PW Solution for Leaf-to-Leaf Communication Restriction

Extension to current VPLS standard [RFC4762] is required.

5. Extension to VPLS for E-Tree

5.1. AC E-Tree Type

Each AC connected to a specific VPLS instance on a PE MUST have an AC E-Tree Type attribute, either Leaf AC or Root AC. For backward compatibility, the default AC E-Tree Type MUST be Root.

This AC E-Tree Type is locally configured on a PE and no signaling is required between PEs.

5.2. VSI E-Tree Type and Identifier

Two new PW interface parameters (as defined in section 5.5 of [RFC4447]) are defined for use in E-Tree VPLS: VSI E-Tree type and VSI E-Tree identifier.

VSI E-Tree type can be either root or leaf and identifies VSI root PW and VSI leaf PW respectively, as defined in section 4.

VSI E-tree identifier is a number that is used to identify a pair of root and leaf PW as part of the same logical VSI interface.

On reception, the two PWs SHALL be handled as the same logical VSI interface with respect to MAC address learning/forwarding, e.g. traffic SHALL NOT be forwarded between such PWs and MAC addresses arriving at one of the PWs SHALL be learned with a common logical VSI interface.

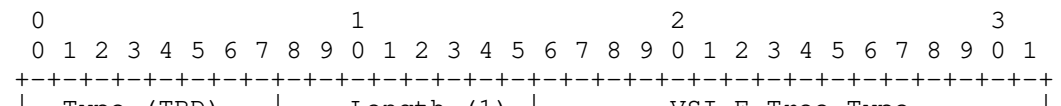
On transmission, the VPLS processing entity SHALL send root-originated traffic via the root PW, and SHALL send leaf-originated traffic via the leaf PW.

The <VSI E-Tree type, VSI E-Tree identifier> pair SHALL be unique in PWs connecting a pair of VPLS PEs.

5.2.1. VSI E-Tree Type Encoding

The VSI E-Tree type field is encoded as an interface parameters sub-TLV (as defined in section 5.5 of [RFC4447]).

The field structure is defined as follows:



VSI E-tree Type can take the following values:

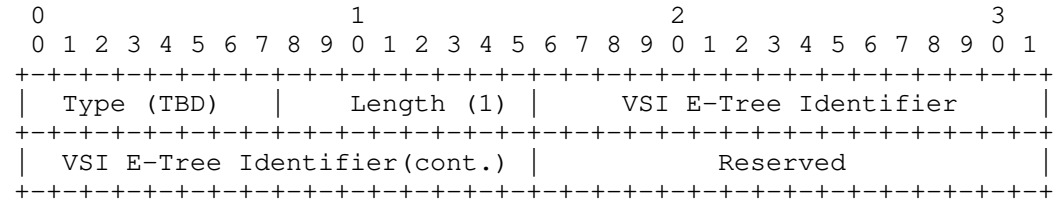
0 E-Tree Root VSI

1 E-Tree Leaf VSI

5.2.2. VSI E-Tree Identifier Encoding

The VSI E-Tree identifier field is encoded as an interface parameters sub-TLV (as defined in section 5.5 of [RFC4447]).

The field structure is defined as follows:



VSI E-tree Identifier is a 32-bit number that is used to identify a pair of root and leaf PW as part of the same logical VSI interface, in the context of a pair of VPLS PEs.

The reserved field SHALL be set to zero.

5.3. Additional Filtering in Data Forwarding

An egress PE SHALL NOT deliver a frame originated at a leaf AC to another leaf AC.

The following specifies how AC E-Tree type per frame is determined:

- A frame received from a root PW indicates that the frame was originated from a root AC
- A frame received from a leaf PW indicates that the frame was originated from a leaf AC.

o For the case where both ingress AC and egress AC are on the same PE, local split horizon implementation on the PE will be sufficient, and is not further discussed in this document.

5.4. Root/Leaf PWs Signaling

Signaling of root and leaf PWs is required only when two PWs are used for interconnecting between pair of VSIs. As explained in section . 5.2. :

o Root VSI E-Tree type SHALL be used to signal a root PW.

o Leaf VSI E-Tree type be used to signal a leaf PW.

PW type signaling rules remain as defined in [RFC4447].

5.5. Supporting Remote AC

When PW is used to interconnect between VSI and a remote AC (e.g. the PW1, PW2 in Figure 3), an Ethernet Raw or Ethernet tagged PW types SHALL be used as defined in [RFC4762].

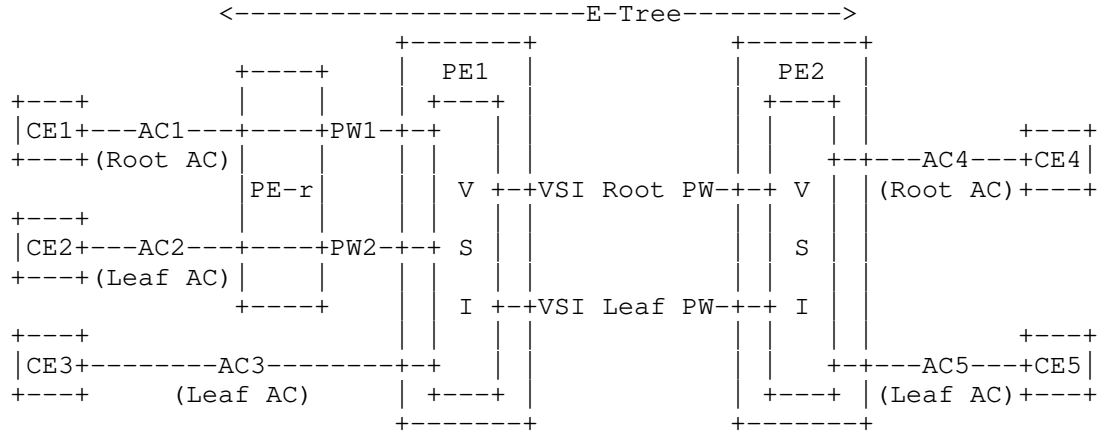


Figure 3: VPLS with Remote AC Connectivity

In addition, the AC type i.e. Root or leaf, SHALL be locally provisioned on the VSI side to specify the remote AC E-Tree Type per PW. Moreover, such PWs that are used for interconnecting between a remote AC and a VSI SHALL considered as separate logical VSI interfaces with respect to MAC address learning/forwarding e.g. traffic forwarding between such PWs is allowed as long as they are not both defined as Leaf.

In Figure 3, AC1 is remotely interconnected to the VPLS service via PW1, and AC2 is remotely interconnected to the VPLS service via PW2.

AC1 is a Root AC and therefore the local type for PW1 in PE1 SHALL be Root.

AC2 is a Leaf AC and therefore the local type for PW2 in PE1 SHALL be Leaf.

6. Backward Compatibility

Root or leaf VSI E-Tree type and identifier parameters SHALL be used only in cases where both PEs are VPLS capable and both support E-Tree root/leaf.

In a case where one of the peers do not support E-Tree, VSI E-Tree type and identifier parameters SHALL NOT be used.

7. Compliance with Requirements

This refers to [Draft ETree VPLS Req] Section 5. Requirements.

The solution prohibits communication between any two Leaf ACs in a VPLS instance.

The solution allows multiple Root ACs in a VPLS instance.

The solution allows Root AC and Leaf AC of a VPLS instance co-exist on any PE.

The solution is applicable to LDP-VPLS [RFC4762].

The solution is applicable to Case 1: Single technology "VPLS Only".

8. Security Considerations

This will be added in later version.

9. IANA Considerations

Additional assignments will be required for the new interface parameter sub-TLV types introduced in Section 4.2. Details will be added in a later version.

10. Acknowledgements

The authors wish to acknowledge the contributions of Luca Martini and Amir Halperin.

11. References

11.1. Normative References

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

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

[RFC4762] Lasserre & Kompella, Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling, January 2007

11.2. Informative References

[Draft VPLS ETree Req] Key, et al., Requirements for MEF E-Tree Support in VPLS, draft-key-l2vpn-vpls-etree-reqt-01.txt, September 2010

Authors' Addresses

Rafi Ram
Orckit-Corrigent
126 Yigal Alon st.
Tel Aviv, Israel
Email: rafir@orckit.com

Daniel Cohn
Orckit-Corrigent
126 Yigal Alon st.
Tel Aviv, Israel
Email: danielc@orckit.com

Raymond Key
Telstra
242 Exhibition Street, Melbourne
VIC 3000, Australia
Email: raymond.key@team.telstra.com

Puneet Agarwal
Broadcom
3151 Zanker Road
San Jose, CA 95134
Email: pagarwal@broadcom.com