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.

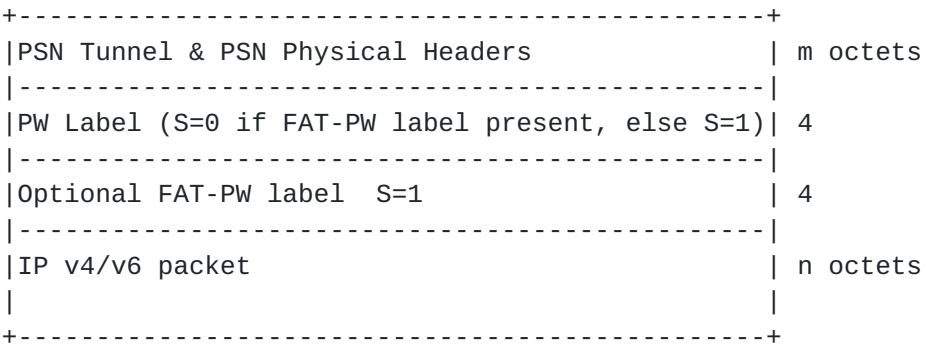


Figure 1 IPv4/v6 packet encapsulated into PPW without CW

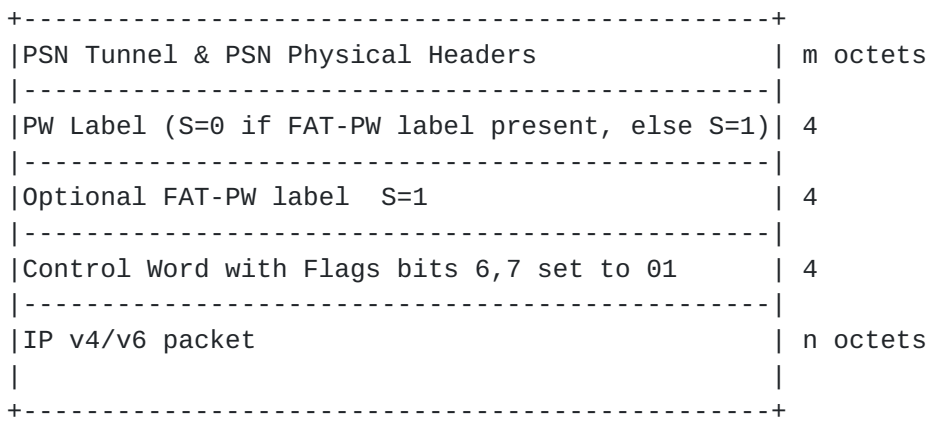


Figure 2 IPv4/v6 packet encapsulated into PPW with CW

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.

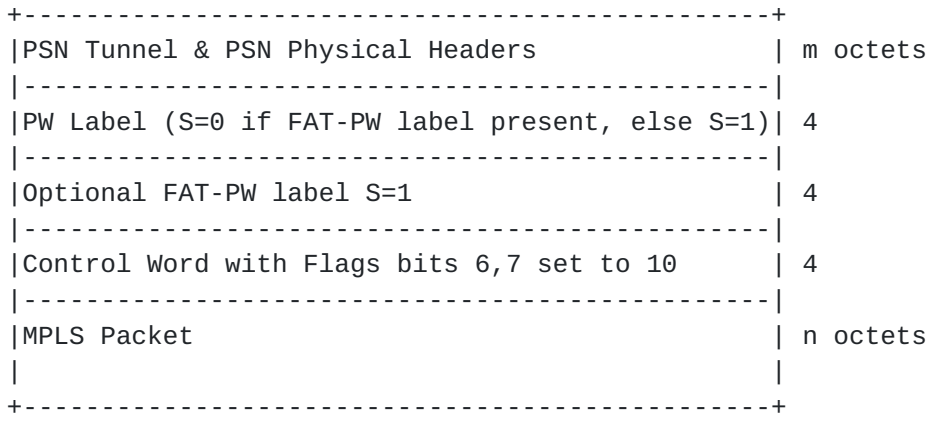


Figure 3 MPLS packet encapsulated into PPW with CW

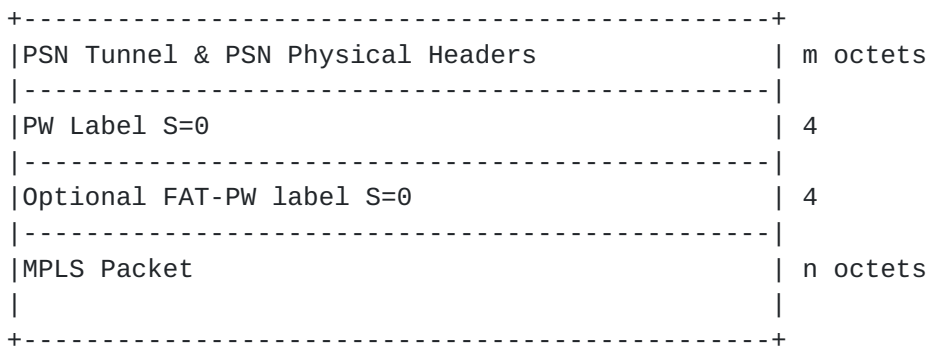


Figure 4 MPLS packet encapsulated into PPW without CW

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.

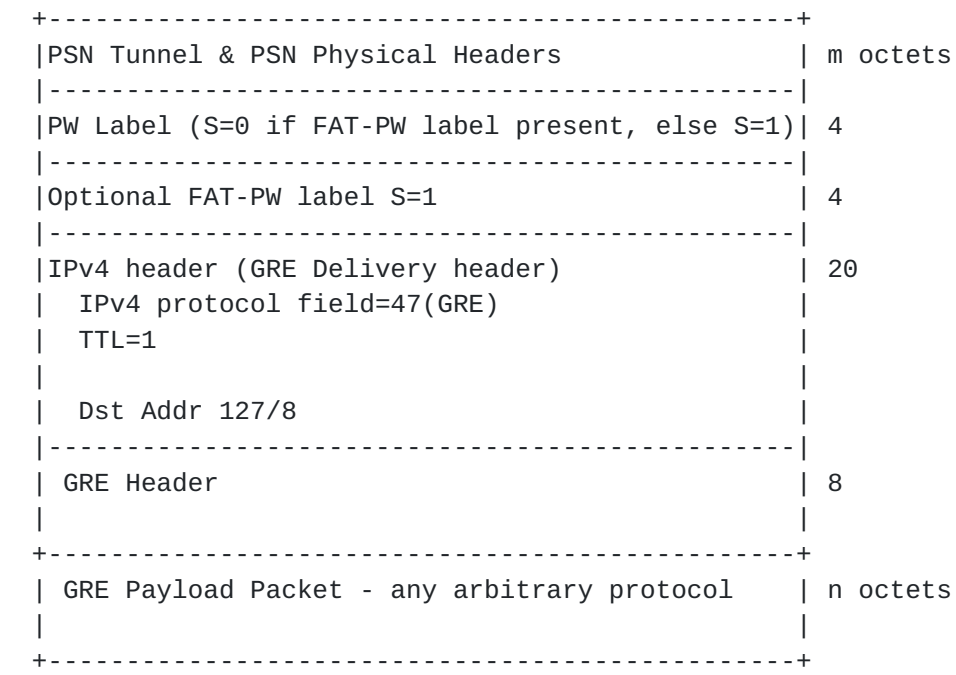


Figure 5 An arbitrary protocol packet encapsulated into PPW without CW

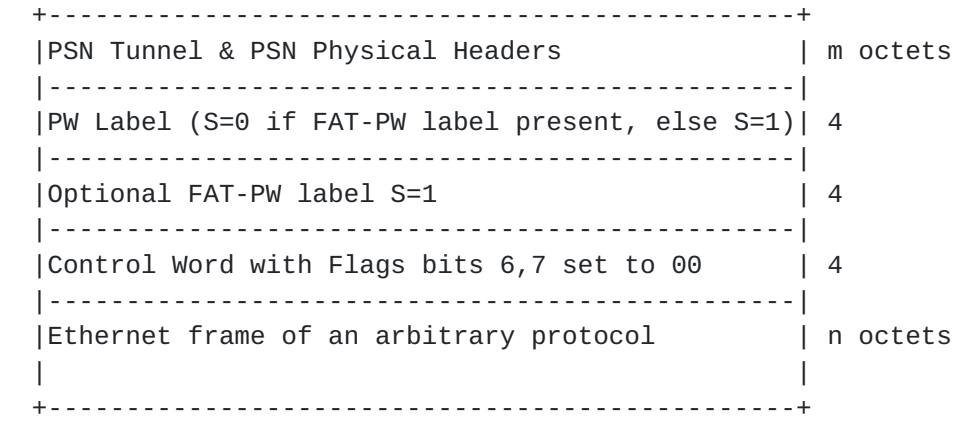


Figure 6 An arbitrary protocol packet encapsulated into PPW with CW

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:

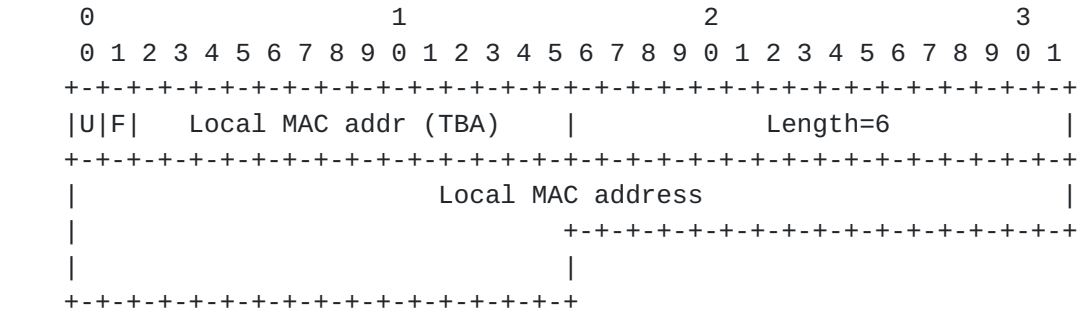
- 00 indicates payload is any protocol encapsulated in an Ethernet frame
- 01 indicates payload is IP
- 10 indicates payload is MPLS

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

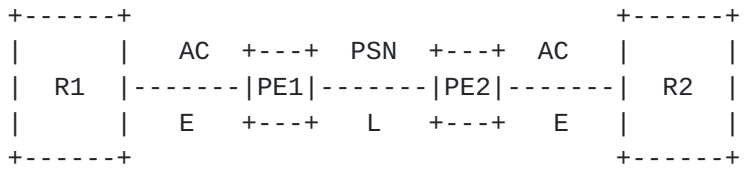
11. Acknowledgments

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:

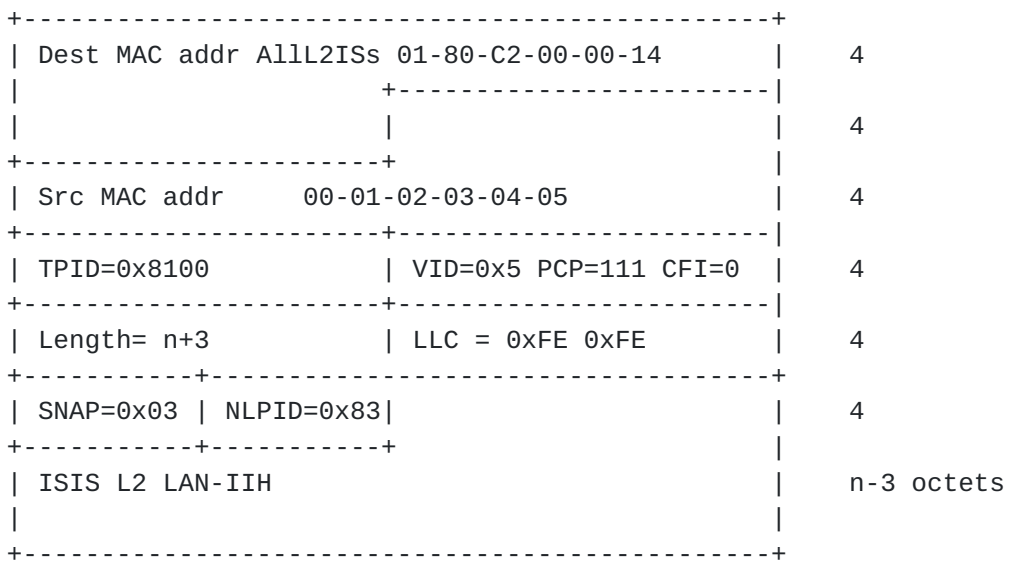


Figure 8 ISIS L2 LAN-IIH from R1 to R2 on AC

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

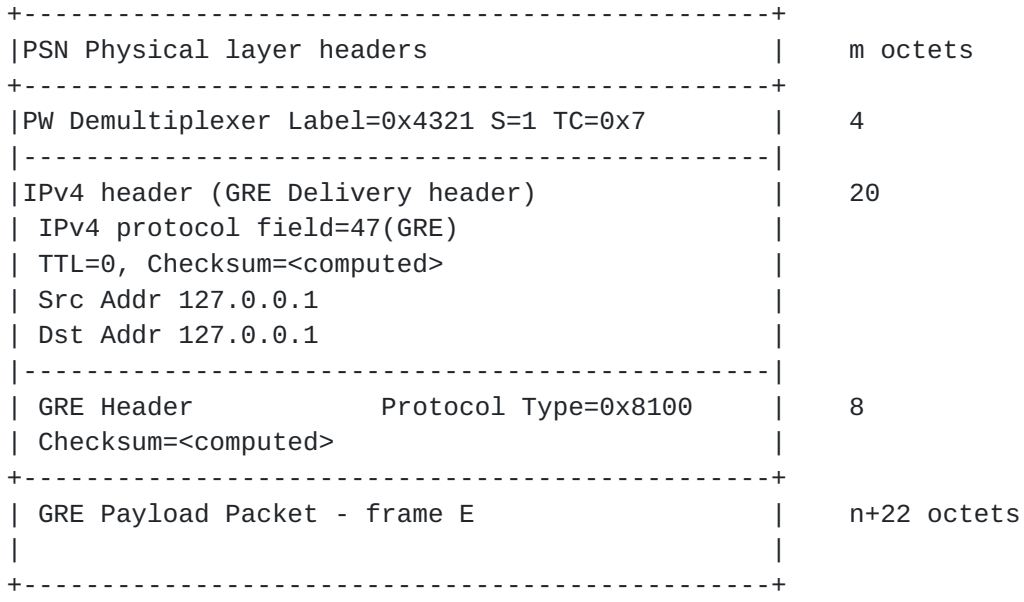


Figure 9 ISIS L2 LAN-IIH from R1 to R2 on PPW-EIM

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

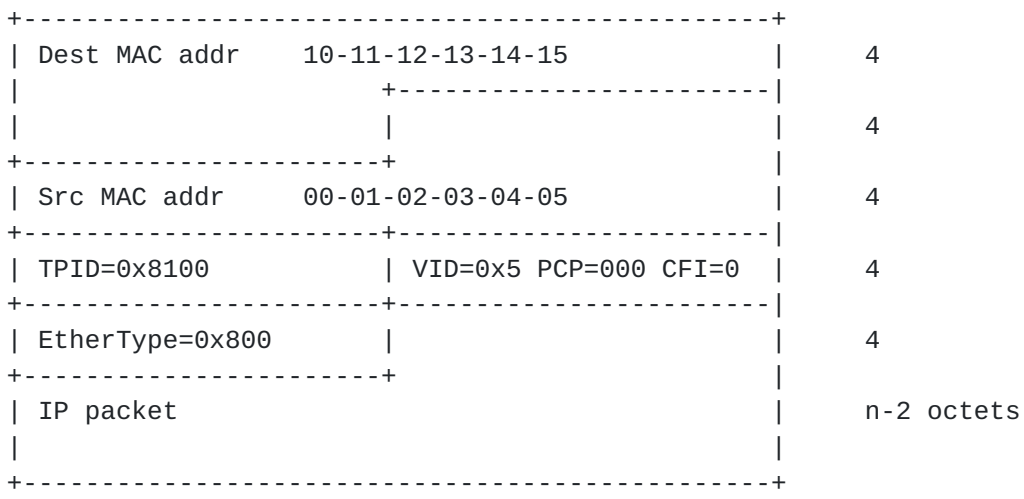


Figure 10 IP packet from R1 to R2 on AC

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

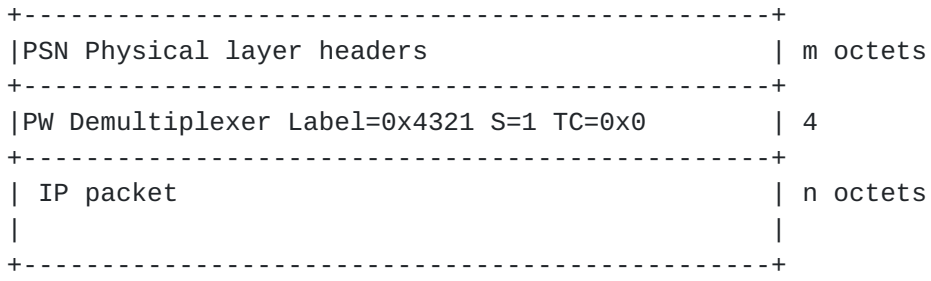
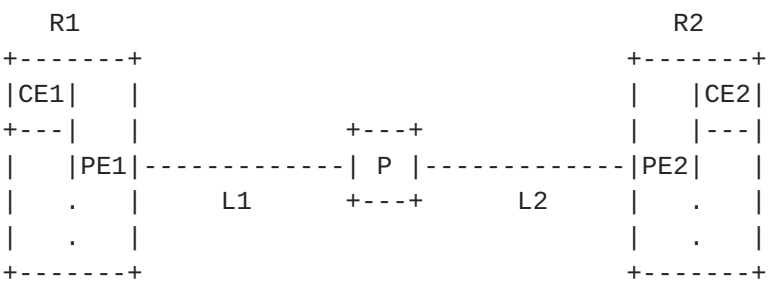


Figure 11 IP packet from R1 to R2 on PPW-EIM

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:

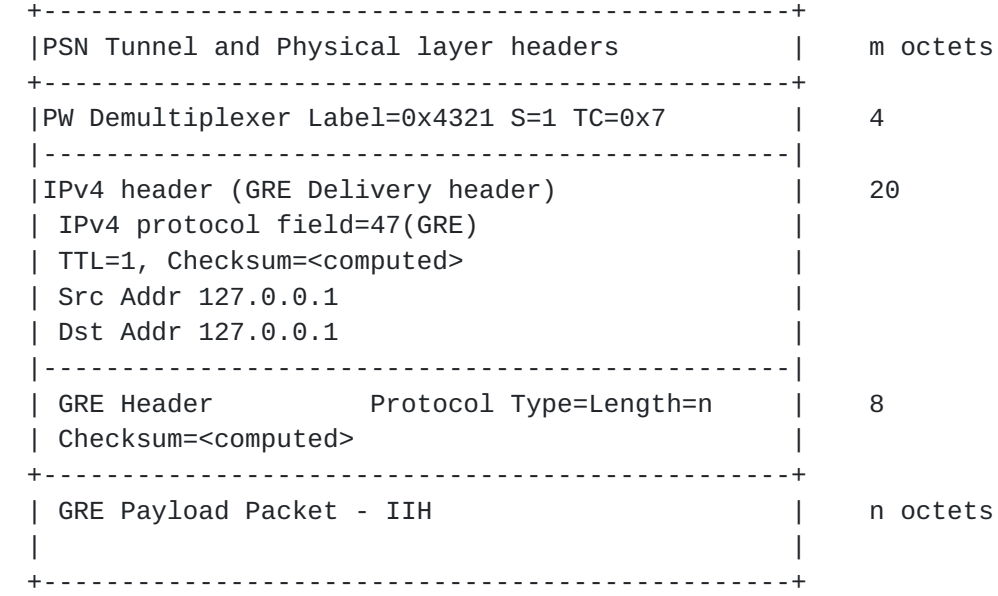


Figure 13 ISIS IIH from CE1 to CE2 on PPW-EIM

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

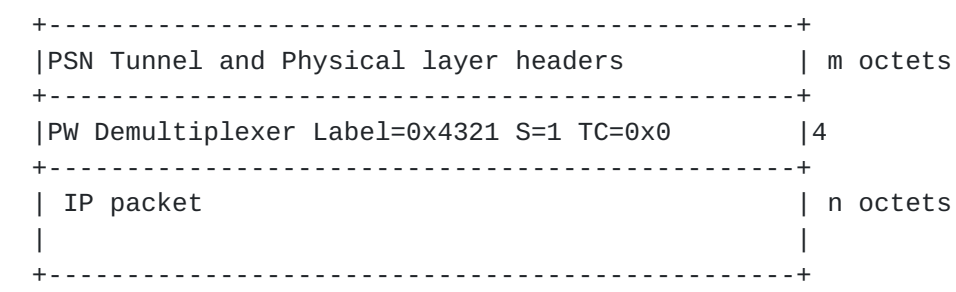


Figure 14 IP packet from CE1 to CE2 on PPW-EIM

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

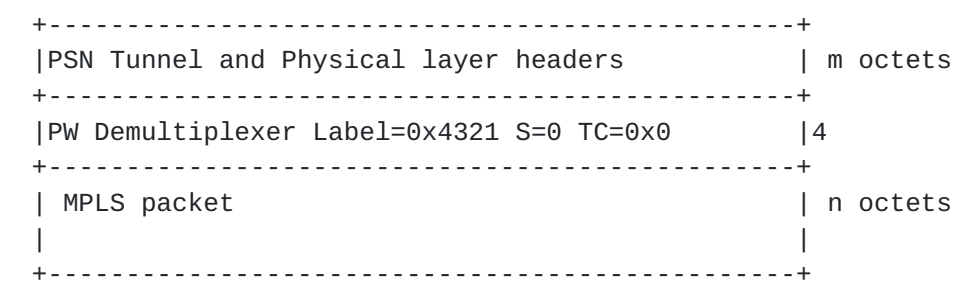


Figure 15 MPLS packet from R1 to R2 on PPW-EIM

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