

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 30, 2011

M. Chen
W. Cao
Huawei Technologies Co., Ltd
A. Takacs
Ericsson
P. Pan
Infinera
June 28, 2011

LDP extensions for Explicit Pseudowire to transport LSP mapping
draft-cao-pwe3-mpls-tp-pw-over-bidir-lsp-03.txt

Abstract

A bidirectional Pseudowire (PW) service currently uses two unidirectional PWs each carried over a unidirectional LSP. Each end point of a PW or segment of multi-segment PW (MS-PW) independently selects the LSP to use to carry the PW for which it is the head end.

Some transport services may require that bidirectional PW traffic follows the same paths through the network in both directions. Therefore, PWs may be required to use LSP with congruent paths. Bidirectional LSPs or co-routed associated unidirectional LSPs allow this service to be provided.

This document specifies an optional extension to LDP that allows both ends of a PW (or segment of a MS-PW) to select and bind to the same bidirectional LSP (or co-routed unidirectional 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 December 30, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. LDP Extensions	5
2.1. PSN Tunnel Binding TLV	6
2.1.1. PSN Tunnel Sub-TLV	7
3. Theory of Operation	8
4. PSN Binding Operation for SS-PW	9
5. PSN Binding Operation for MS-PW	11
6. Security Considerations	13
7. IANA Considerations	13
7.1. LDP TLV Types	13
7.1.1. PSN Tunnel Sub-TLVs	13
7.2. LDP Status Codes	14
8. Acknowledgements	14
9. References	14
9.1. Normative References	14
9.2. Informative References	14
Authors' Addresses	15

1. Introduction

Pseudo Wire (PW) Emulation Edge-to-Edge (PWE3) [RFC3985] is a mechanism to emulate a number of layer 2 services, such as Asynchronous Transfer Mode (ATM), Frame Relay or Ethernet. 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). Today PWE3 generally uses two reverse unidirectional Label Distribution Protocol (LDP) [RFC5036] or Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) [RFC3209] LSPs as PSN tunnels, and each of the PEs selects and binds PSN tunnel independently. There is no protocol-based provision to explicitly associate a PW with a specific PSN tunnel.

For transport applications it has been identified that many transport services may require bidirectional traffic that follows congruent paths. When bidirectional LSPs [RFC3471][RFC3473] are used as PSN tunnels, this requirement can be fulfilled if both PEs of a specific/segment PW select and bind to the same bidirectional LSPs. In the case of unidirectional LSPs, LSPs with congruent paths need to be selected to support the PW. However, current mechanisms cannot guarantee appropriate mapping of PWs to underlying LSPs.

The lack of the control over LSP-PW binding may introduce service issues in operation, as shown in Figure 1.

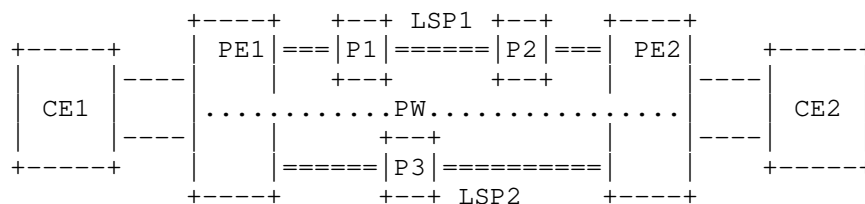


Figure 1: Inconsistent SS-PW to LSP binding scenario

There are two bidirectional LSPs: LSP1 and LSP2, along diverse paths. A bidirectional PW service is offered 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 the PE1->PE2 direction of the PW, while PE2 may select LSP2 (PE1-P3-PE2) as the PSN tunnel for the PE2->PE1 direction of the PW.

Consequently, the bidirectional PW service is delivered over two disjoint LSPs, which may have completely different service attributes in terms of bandwidth and latency. If service offering requires

consistent traffic behavior on forward and reverse direction, this may not be acceptable.

The similar problems may also exist in bidirectional 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 and the underlying PSN tunnels. However, this is prone to configuration errors and does not 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 at the bi-directional PW's setup time. The TLV carries PW's binding profile and provides both explicit and implicit information on the underlying PSN tunnels.

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.

Presently, the extension supports two types of binding requests:

1. Congruent binding: the requesting PE will ask the underlying LSPs to be congruent.
2. Strict binding: the requesting PE will choose and explicitly indicate both forwarding and reverse LSP's in the requests.

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 end point 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 explicit PW to PSN tunnel binding is required. The format of this TLV is as follows:

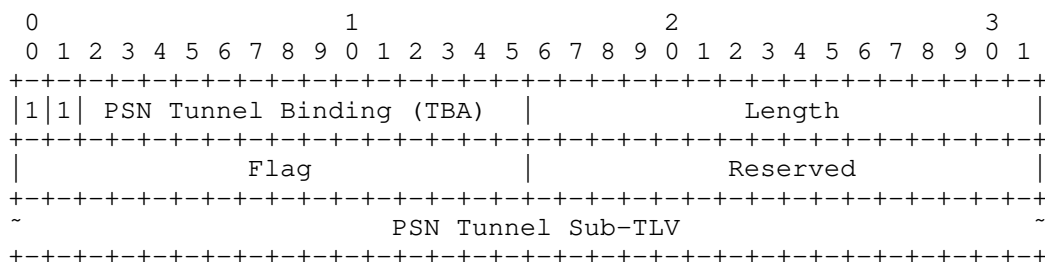
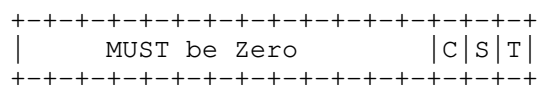


Figure 2: PSN Tunnel Binding TLV

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:



Three flags have been defined at the present time.

C (Congruent path) bit: This informs the remote T-PE/S-PEs about the properties of the underlying PSN tunnels. When set, the remote T-PE/S-PEs need to select tunnel/LSPs with congruent paths (e.g., co-routed bidirectional LSP). If there is no satisfied tunnel, it may trigger the remote T-PE/S-PEs to establish a new tunnel.

S (Strict) bit: This instructs the PEs with respect to the handling of the underlying PSN tunnels. 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 PSN tunnels. When the bit is set, the PSN 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.

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.

In this document two sub-TLVs are defined: the IPv4/IPv6 Tunnel sub-TLVs. The format of the PSN Tunnel sub-TLVs is as follows:

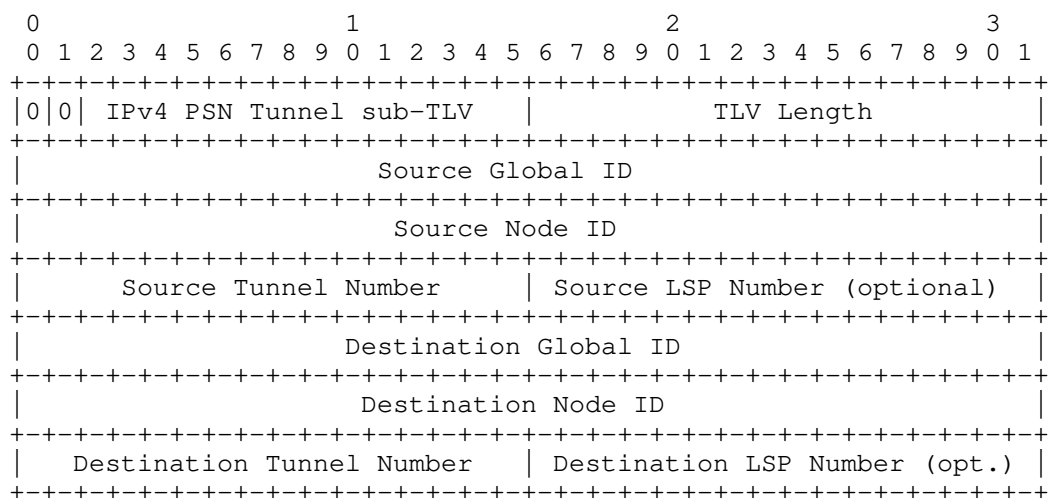


Figure 3: IPv4 PSN Tunnel sub-TLV format

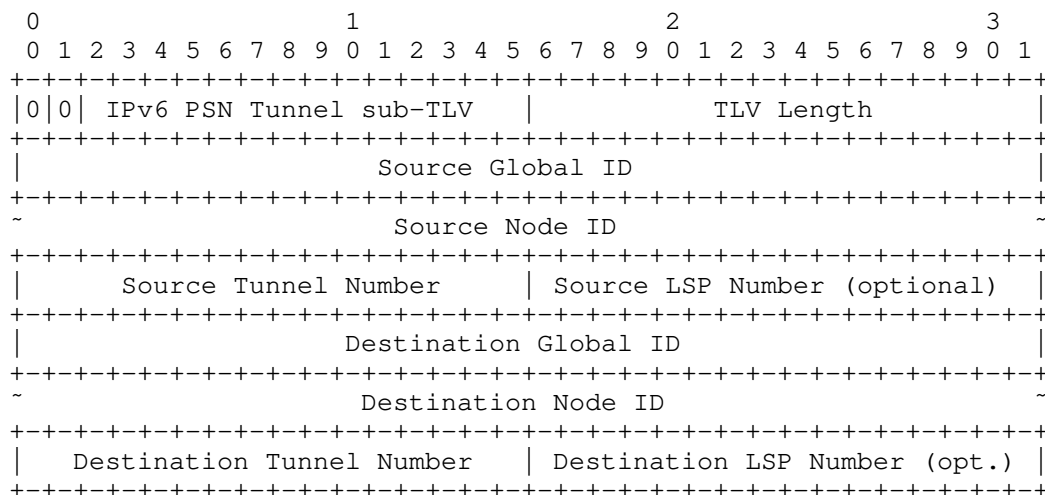


Figure 4: IPv6 PSN Tunnel sub-TLV format

The definition of Source and Destination Global/Node IDs and Tunnel/LSP Numbers are derived from [I-D.ietf-mpls-tp-identifiers]. The notation is designed to describe co-routed or bi-directional LSPs, which is suitable in the context of the work here.

As defined in Section 4.6.1.2 and Section 4.6.2.2 of [RFC3209], the "Tunnel end point 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 determines whether it stands for tunnel or LSP.

When the T-bit is set, it identifies a tunnel, and the Source/Destination LSP Number fields MUST be set to 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 no one available

or satisfies the constraints, the PE may trigger to 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 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 co-routed or bidirectional tunnels/LSPs at the same time. There may cause collisions of tunnel/LSPs selection as both PEs assume the active role.

The PEs can be generally categorized into two types:

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.

Segmented PW has defined the active/passive role election (Section 7.2.1, [RFC6073]). This document will not define any new procedures.

In the remaining of this document, it will elaborate the operation in two situations:

1. SS-PW: In this scenario, both PEs of a PW assume active roles
2. MS-PW: One PE is active, while the other is passive. The PWs are setup using FEC 129

4. PSN Binding Operation for SS-PW

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

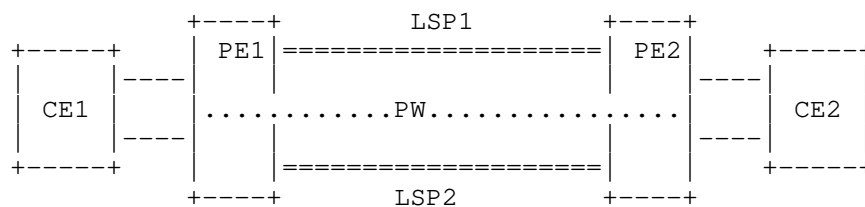


Figure 5: PSN binding operation in SS-PW environment

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), it 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.

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 are congruent.

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, 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 then 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 switch the PW 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. Noting that when T-bit is set, the working LSP broken will not trigger to update the changes if there are protection LSPs.

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

However, if the new Label Binding message does not contain the PSN Tunnel Binding TLV, it declares the removal of any congruent 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.

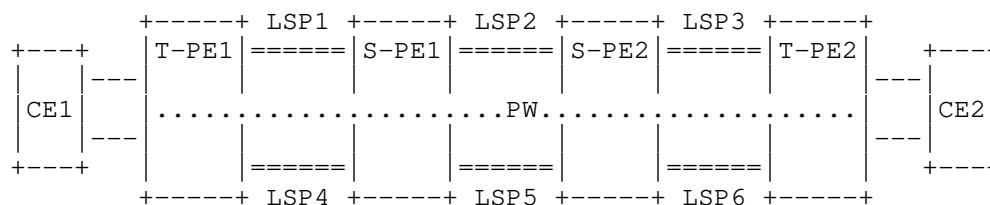


Figure 6: PSN binding operation in MS-PW environment

When the active PE (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 (say 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 (T-PE1) MUST set the S-bit, reset the C-bit and indicates the binding tunnel/LSP to the next-hop S-PE (S-PE1).

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 (S-PE2), with the PSN Tunnel sub-TLV carrying the information of the new PSN tunnel/LSPs selected by S-PE1 for the next PW segment. S-PE2 and subsequent S-PEs will repeat the same operation until the Label Mapping message reaches to the remote T-PE (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, 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 Jourday, Andy Malis, Curtis Villamizar and Luca Martini.

9. References

9.1. Normative References

- [I-D.ietf-mpls-tp-identifiers]
Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-06 (work in progress), June 2011.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

9.2. Informative References

- [I-D.ietf-pwe3-dynamic-ms-pw]
Martini, L., Bocci, M., Balus, F., Bitar, N., Shah, H., Aissaoui, M., Rusmisl, J., Serbest, Y., Malis, A., Metz, C., McDysan, D., Sugimoto, J., Duckett, M., Loomis, M., Doolan, P., Pan, P., Pate, P., Radoaca, V., Wada, Y., and Y. Seo, "Dynamic Placement of Multi Segment Pseudo Wires", draft-ietf-pwe3-dynamic-ms-pw-13 (work in progress),

October 2010.

- [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.

Authors' Addresses

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

Email: mach@huawei.com

Wei Cao
Huawei Technologies Co., Ltd
No. 3 Xinxu Road, Shang-di, Hai-dian District
Beijing 100085
China

Email: wayne.caowei@huawei.com

Attila Takacs
Ericsson
Laborc u. 1.
Budapest 1037
Hungary

Email: attila.takacs@ericsson.com

Ping Pan
Infinera
Sri Mohana Satya Srinivas Singamsetty
US

Email: ppan@infinera.com

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

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

M. Daikoku
KDDI Corporation
October 22, 2012

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

Abstract

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

Status of this Memo

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

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

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

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

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the

document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Conventions used in this document	4
2. Terminology	4
3. ICCP extension requirements	5
3.1. Multi-chassis MSP Protection Model	5
3.2. ICCP aspects	6
4. ICCP TLV extensions for MSP	6
4.1. MSP connect TLV	6
4.2. MSP disconnect TLV	7
4.2.1. MSP disconnect cause TLV	8
4.3. MSP group config TLV	8
4.4. MSP port config TLV	10
4.5. MSP section state TLV	11
4.6. MSP switch command TLV	12
4.7. MSP group state TLV	13
4.8. MSP Synchronization Request TLV	14
4.9. MSP Synchronization Data TLV	15
5. PE Node Failure	16
6. Security Considerations	16
7. IANA Consideration	16
8. References	16
8.1. Normative References	16
8.2. Informative References	16
Authors' Addresses	16

1. Introduction

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

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

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

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

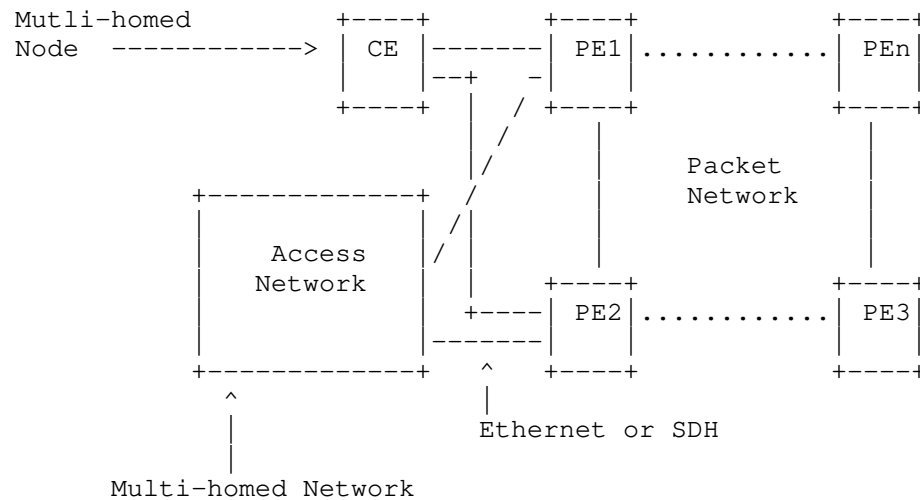


Figure 1: Attachment circuit multi-homed to Packet Network

1.1. Conventions used in this document

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

2. Terminology

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

3. ICCP extension requirements

3.1. Multi-chassis MSP Protection Model

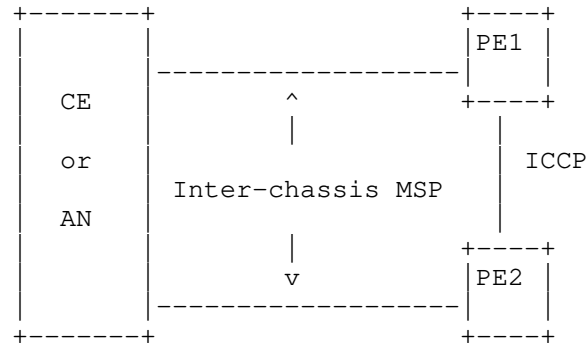


Figure 2: Generic Multi-chassis MSP Protection Model

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

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

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

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

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

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

The ICCP application described in this document is responsible for

the state synchronization between different chassis forming a RG.

3.2. ICCP aspects

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

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

4. ICCP TLV extensions for MSP

The following sections specify the format of MSP application TLVs.

4.1. MSP connect TLV

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

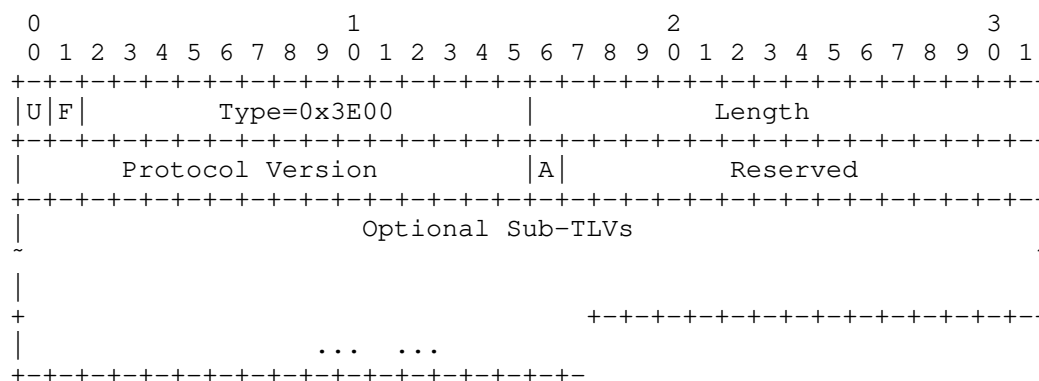


Figure 3: MSP connect TLV

- U and F Bits
Both are set to 0.
- Type
Set temporarily to 0x3E00 for "MSP connect TLV"
- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.
- Protocol Version
The version of this particular protocol for the purposes of ICCP.
This is set to 0x0001.
- A Bit
Acknowledgement Bit. Set to 1 if the sender has received a MSP Connect TLV from the recipient. Otherwise, set to 0.
- Reserved
Reserved for future use.
- Optional Sub-TLVs
There are no optional Sub-TLVs defined for this version of the Protocol.

4.2. MSP disconnect TLV

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

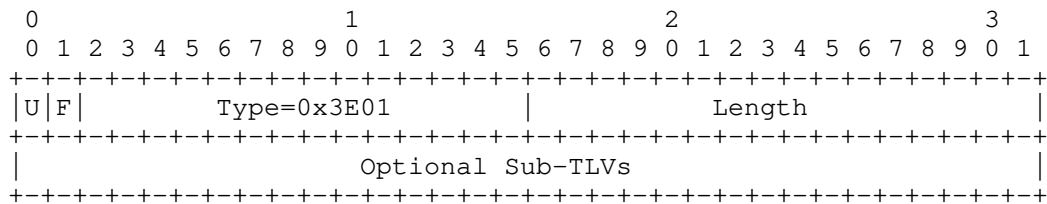


Figure 4: MSP disconnect TLV

- U and F Bits
Both are set to 0.
- Type
Set temporarily to 0x3E01 for "MSP disconnect TLV"

- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.
- Optional Sub-TLVs
There are no optional Sub-TLVs defined for this version of the Protocol.

4.2.1. MSP disconnect cause TLV

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

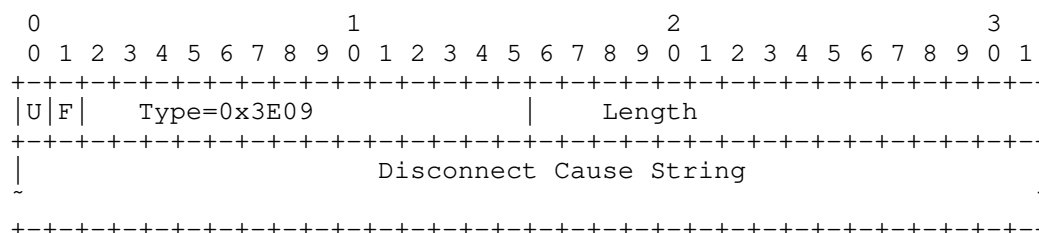


Figure 5: MSP disconnect TLV

- U and F Bits
Both are set to 0.
- Type
Set temporarily to 0x3E09 for "MSP disconnect cause TLV"
- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length fields.
- Disconnect Cause String
Variable length string specifying the reason for the disconnect message. Used for network management.

4.3. MSP group config TLV

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

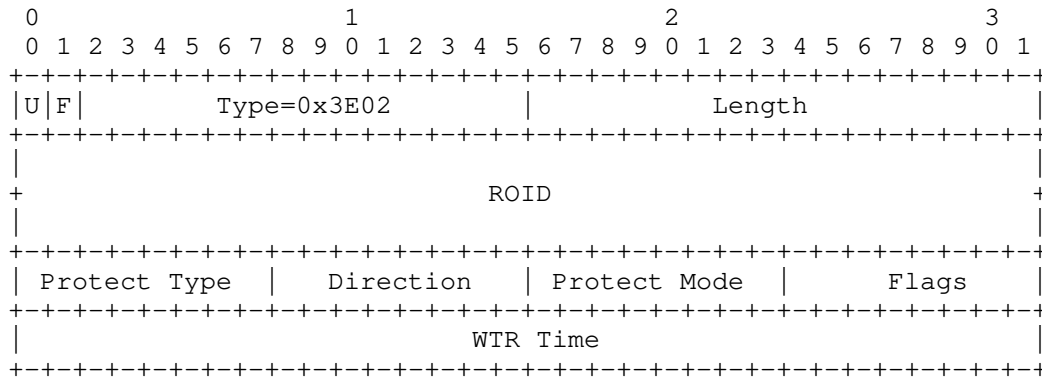


Figure 6: MSP group config TLV

- U and F Bits
Both are set to 0.
- Type
Set temporarily to 0x3E02 for "MSP group config TLV"
- Length
Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.
- ROID
Defined in the [I-D:ietf-pwe3-iccp]. Eight octets, uniquely identifies the Redundant Object.
- Protect Type
One octet encoding the protect type of the MSP protect group as follows:
0x00 1+1
0x01 1:1
0x02-0xFF reserved
- Direction
One octet encoding the architecture of the network as follows:
0x00 unidirectional
0x01 bidirectional
- Reversion Mode
One octet encoding the mode of operation as follows:
0x00 non-revertive operation
0x01 revertive operation

- Flags

One octet. Valid values are:

-i Synchronized (0x01)

Indicates that the sender has concluded transmitting all group configuration information.

-ii Purge Configuration (0x02)

Indicates that the group is no longer configured for MSP operation.

- WTR Time

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

4.4. MSP port config TLV

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

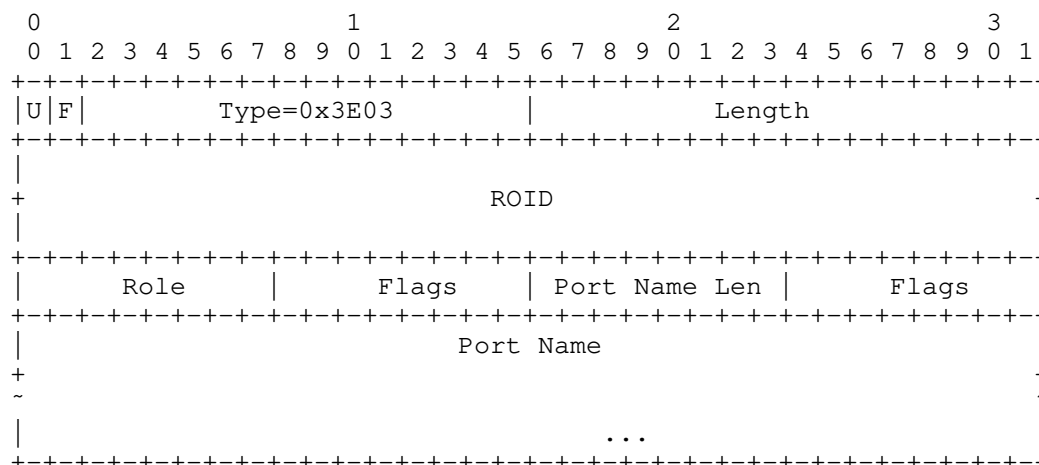


Figure 7: MSP port config TLV

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- ROID
Defined in the [I-D:ietf-pwe3-iccp].Eight octets, uniquely identifies the Redundant Object.
- Role
One octet encoding the role of the section as follows:
0x00 working
0x01 protection
- Flags
One octet. Valid values are:
-i Synchronized (0x01)
Indicates that the sender has concluded transmitting all group configuration information.
-ii Purge Configuration (0x02)
Indicates that the group is no longer configured for MSP operation.
- Port Name Len
One octet, length of the "Port Name" field in octets.
- Port Name
Port name encoded in UTF-8 format, up to a maximum of 32 characters.

4.5. MSP section state TLV

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

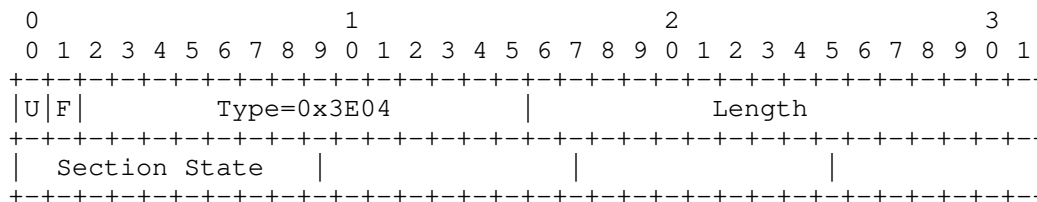


Figure 8: MSP section state TLV

- U and F Bits
Both are set to 0.
- Type
Set temporarily to 0x3E04 for "MSP section state TLV"
- Length
Length of the TLV in octets excluding the U-bit,F-bit,Type,and Length

fields.

- Section State

One octet encoding the section state as follows:

0x00 the signal is ok

0x01 Signal fail high priority

0x02 Signal fail low priority

0x03 Signal degrade high priority

0x04 Signal degrade low priority

4.6. MSP switch command TLV

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

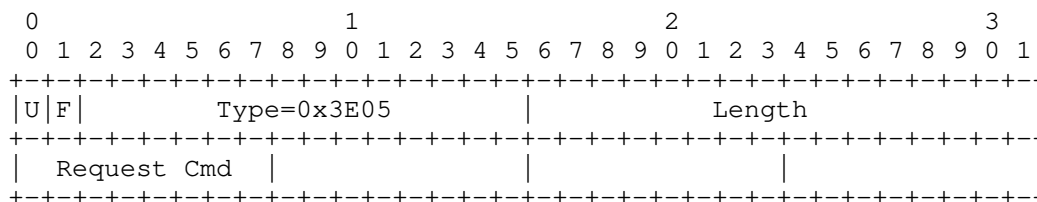


Figure 9: MSP switch command TLV

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Request Cmd

One octet.The switch command issued at the MSP APS controller interface. The following are the possible values, in order of priority from highest to lowest:

(1111) Clear

(1101) Lockout of protection(LP)

(1011) Forced Switch working-to-protection

(1001) Forced Switch protection-to-working

(0111) Manual switch working-to-protection

(0101) Manual switch protection-to-working

(0100) Exercise

4.7. MSP group state TLV

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

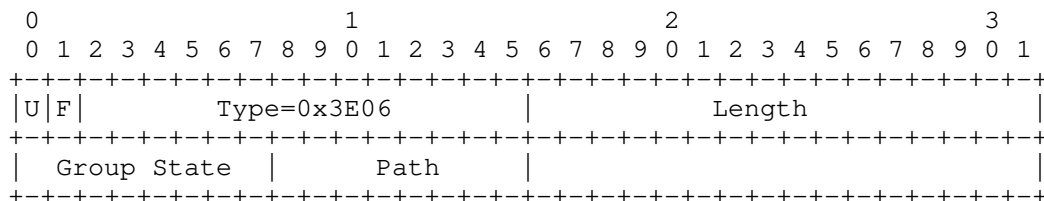


Figure 10: MSP group state TLV

- U and F Bits
Both are set to 0.

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

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

- Group State
One octet encoding the current state of the MSP protect group as follows:

- 0x00 No request
- 0x01 Do not revert
- 0x02 Reverse request
- 0x03 Unused
- 0x04 Exercise
- 0x05 Unused
- 0x06 Wait-to restore
- 0x07 Unused
- 0x08 Manual switch
- 0x09 Unused
- 0x0A Signal degrade low priority
- 0x0B Signal degrade high priority
- 0x0C Signal fail low priority
- 0x0D Signal fail high priority
- 0x0E Forced switch
- 0x0F Lockout of protection

- Path

One octet encoding the active path of the MSP protect group as follows:

0x00 the active path is the working section

0x01 the active path is the protection section

4.8. MSP Synchronization Request TLV

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

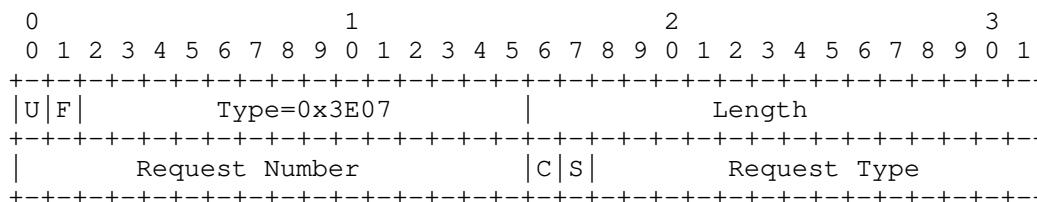


Figure 11: MSP Synchronization Request TLV

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Request Number

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

- C Bit

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

- S Bit

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

- Request Type

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

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

4.9. MSP Synchronization Data TLV

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

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

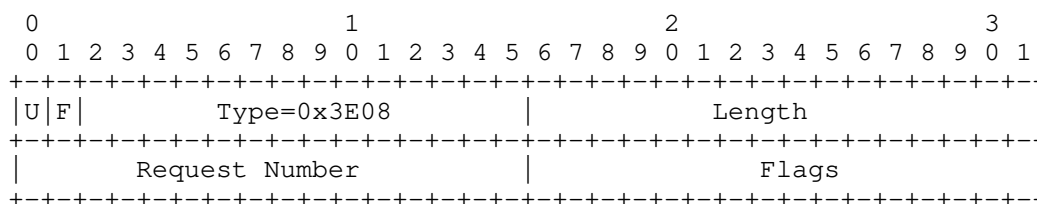


Figure 12: MSP group notify TLV

- U and F Bits
 Both are set to 0.

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

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

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

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

5. PE Node Failure

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

6. Security Considerations

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

7. IANA Consideration

TBD.

8. References

8.1. Normative References

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

8.2. Informative References

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

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

Authors' Addresses

Hongjie Hao
ZTE Corporation

Email: hao.hongjie@zte.com.cn

Yuxia Ma
ZTE Corporation

Email: ma.yuxia@zte.com.cn

Weiqliang Cheng
China Mobile

Email: chengweiqliang@chinamobile.com

Daniel Cohn

Email: daniel.cohn.ietf@gmail.com

Masahiro Daikoku
KDDI Corporation

Email: ms-daikoku@kddi.com

Wanming Cao
ZTE Corporation

Email: cao.wanming@zte.com.cn

Jinghai Yu
ZTE Corporation

Email: yu.jinghai@zte.com.cn

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

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/lid-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 January 7, 2010

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	3
1.1	Requirements Language	3
1.2	Terminology	3
1.3	Notational Conventions in Backus-Naur Form	4
2	PW status refresh reduction protocol	4
2.1	Protocol states	4
2.1.1	INACTIVE	5
2.1.2	STARTUP	5
2.1.3	ACTIVE	5
2.2	Timer value change transition procedure	5
3	PW status refresh reduction procedure	6
4	PW status refresh reduction Message Encoding	6
5	PW status refresh reduction Control Messages	9
5.0.1	Notification message	10
5.0.2	PW Configuration Message	10
5.0.2.1	MPLS-TP Tunnel ID	11
5.0.2.2	PW ID configured List	11
5.0.2.3	PW ID unconfigured List	12
6	PW provisioning verification procedure	13
6.1	PW ID List advertising and processing	13
7	Security Considerations	14
8	IANA Considerations	14
8.1	PW Status Refresh Reduction Message Types	14
8.2	PW Configuration Message Sub-TLVs	14
8.3	PW Status Refresh Reduction Notification Codes	15
9	References	15
9.1	Normative References	15
9.2	Informative References	16
10	Author's Addresses	16

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 than the original one the PE will operate on the new timer value immediately.
- iii. If the new value is smaller than 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

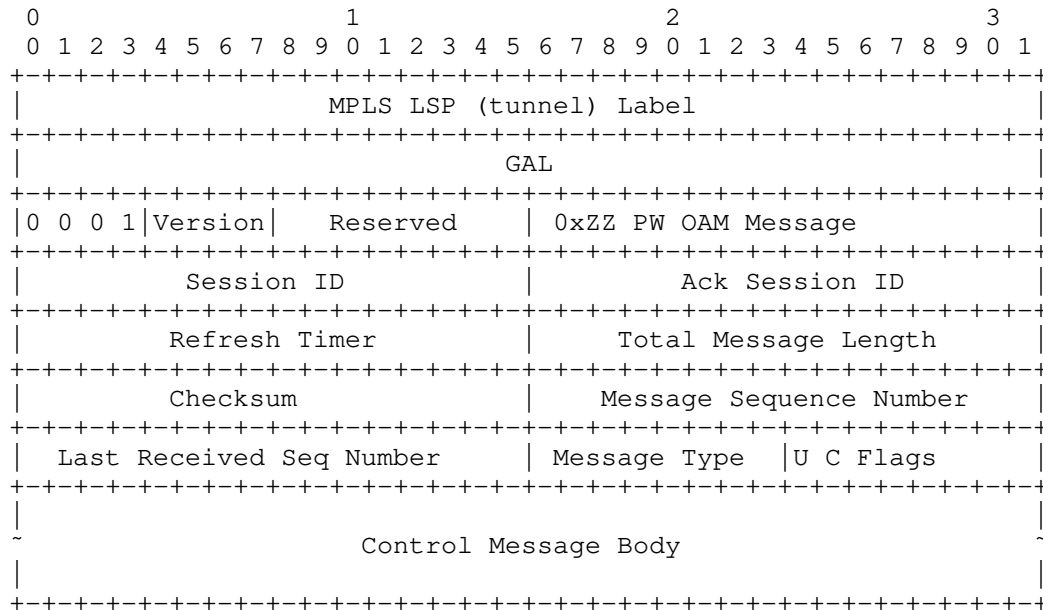
When the refresh reduction protocol, on a particular LSP, is in the ACTIVE state, the PE can send all PW status messages, for PWs on that LSP, with a refresh timer value of zero. This greatly decreases the amount of messages that the PE needs to transmit to the remote PE because once the PW status message 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.

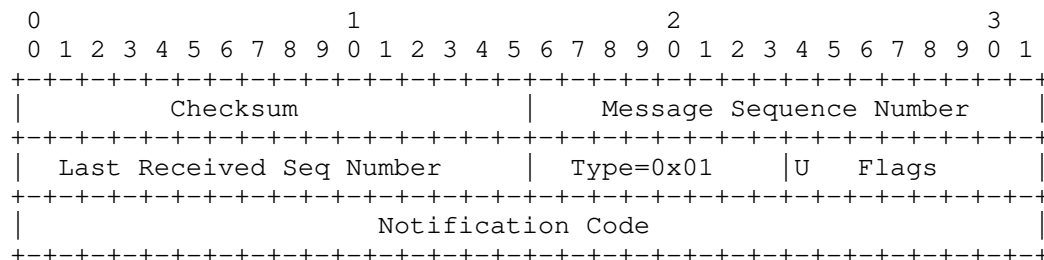
When there is the need to send a PW status refresh reduction Control Messages, the system can attach it to a scheduled PW status refresh reduction or send one ahead of time. In any case PW status refresh reduction Control Messages always piggy back on normal messages.

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

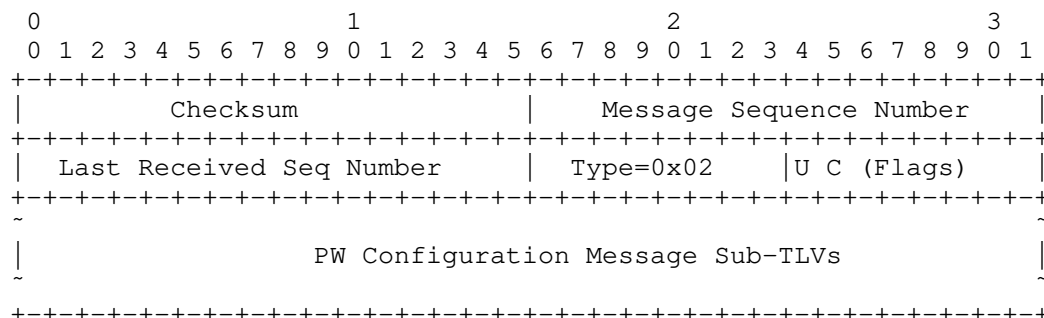
The most common use of the Notification Message is to acknowledge the reception of a message by indicating the received message sequence number in the "Last Received Sequence Number" field. The notification message is encoded as follows:



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).



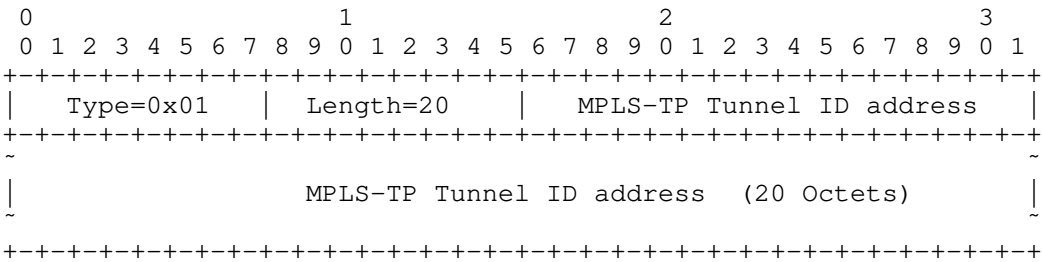
The PW Configuration Message type is set to 0x02. For this message the U-bit is set to 1 as processing of these messages is OPTIONAL.

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:



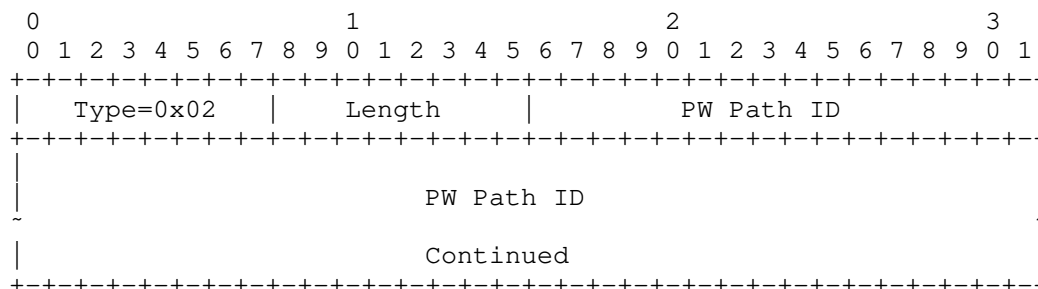
The MPLS-TP point to point tunnel ID is defined in [IDENTIFIER] as follows:

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.

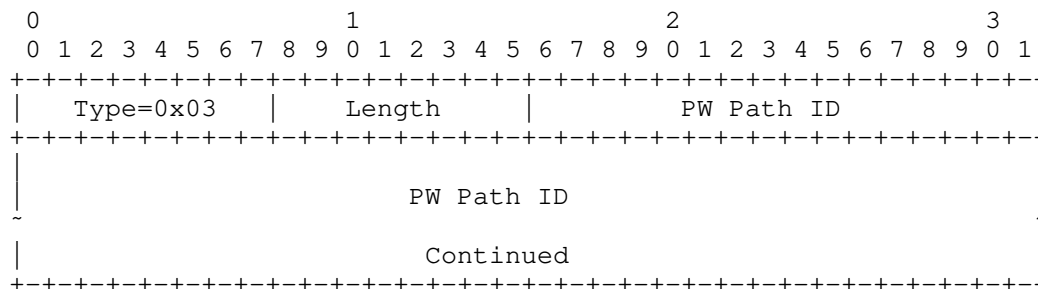


The PW Path ID is a 32 octet pseudowire path identifier specified in [IDENTIFIER] as follows: AGI::Src-Global_ID::Src-Node_ID::Src-AC_ID::Dst-Global_ID::Dst-Node_ID::Dst-AC_ID

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

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.2. Acronyms	5
2. VCCV Control Channel When The Control Word is Used	6
3. VCCV Control Channel When The Control Word is Not Used	6
4. IANA Considerations	19
4.1. VCCV Interface Parameters Sub-TLV	19
4.1.1. MPLS VCCV Control Channel (CC) Type 4	19
5. Security Considerations	24
6. Acknowledgements	25
7. References	26
7.1. Normative References	26
7.2. Informative References	26

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

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.

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.

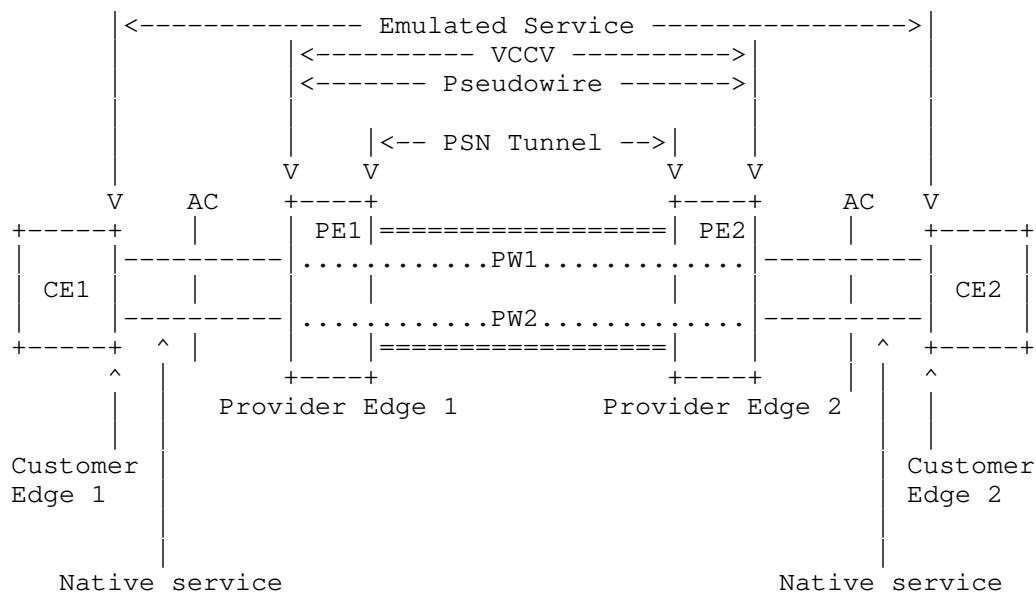


Figure 1: PWE3 VCCV Operation Reference Model

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.

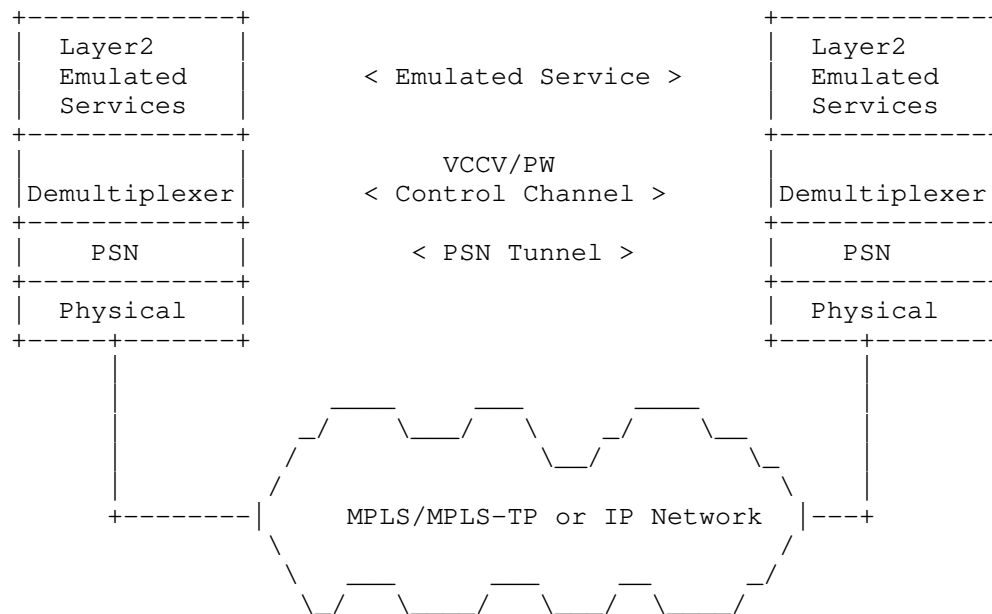


Figure 2: PWE3 Protocol Stack Reference Model including the VCCV Control Channel

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

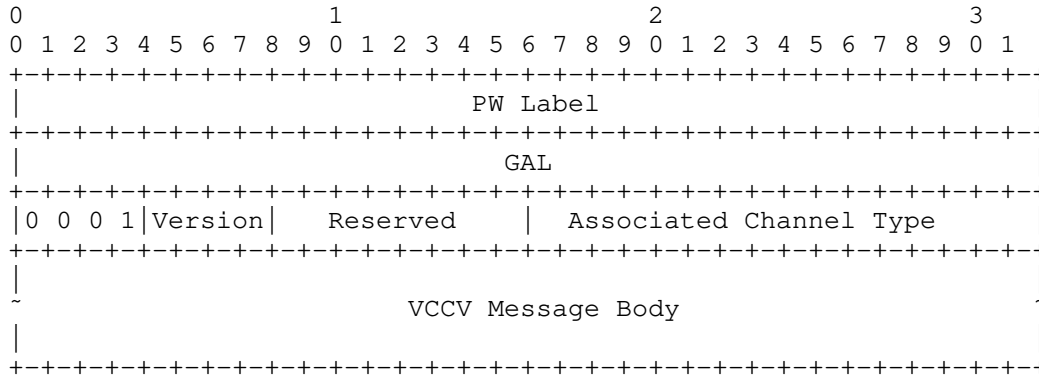
AC	Attachment Circuit [RFC3985].
AVP	Attribute Value Pair [RFC3931].
CC	Control Channel (used as CC Type).
CE	Customer Edge.
CV	Connectivity Verification (used as CV Type).
CW	Control Word [RFC3985].
L2SS	L2-Specific Sublayer [RFC3931].
LCCE	L2TP Control Connection Endpoint [RFC3931].
OAM	Operation and 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
Intended status: Standards Track
Expires: November 21, 2012

P. Kwok
P. Dutta
Alcatel-Lucent
F. Jounay
France Telecom
May 20, 2012

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

Abstract

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

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

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

Requirements Language

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

Status of this Memo

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

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

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

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

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. PW Communities	5
3. Defined PW Community Types	6
3.1. PW Template Community	6
3.1.1. PW Generic Template Community	7
3.2. PW Color Community	7
3.2.1. PW Generic Color Community	7
4. IANA Considerations	7
5. Security Considerations	8
6. Acknowledgements	8
7. References	8
7.1. Normative References	8
7.2. References	8
Authors' Addresses	8

1. Introduction

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

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

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

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

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

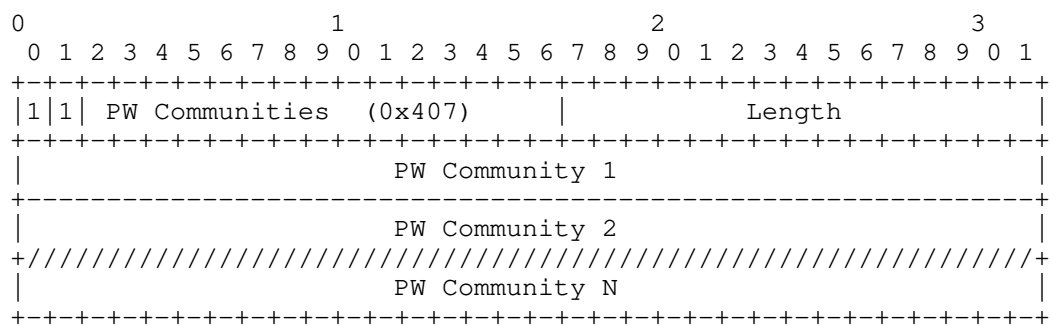
In Dynamic MS-PW, a T-PE or S-PE may need to choose one next-hop from several Equal Cost Multi-Path (ECMP) next-hops provided by best matching PW Route. One way to do ECMP selection is to apply some form of hash function on AGI/SAII/TAII of the PW but that strictly limits the MS-PW addressing schemes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

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

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

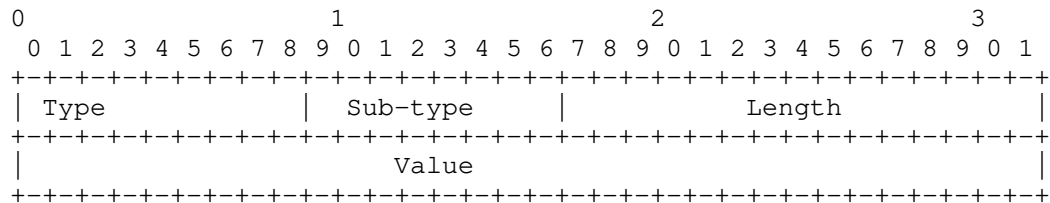
2. PW Communities

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



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

A PW Community is defined as follows:



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

3. Defined PW Community Types

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

3.1. PW Template Community

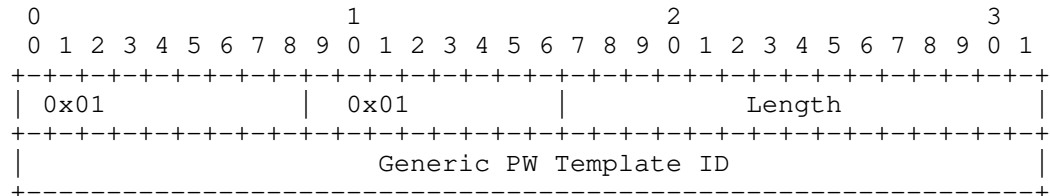
A PW Template community (PW Community Type 0x01) can be considered as a template that has a set of rules defined locally by a T-PE or S-PE. Each T-PE or S-PE can define its own set of rules and its upto the administrative domain to maintain congruities among PW community rules through which PW set-up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

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

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

3.1.1. PW Generic Template Community

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

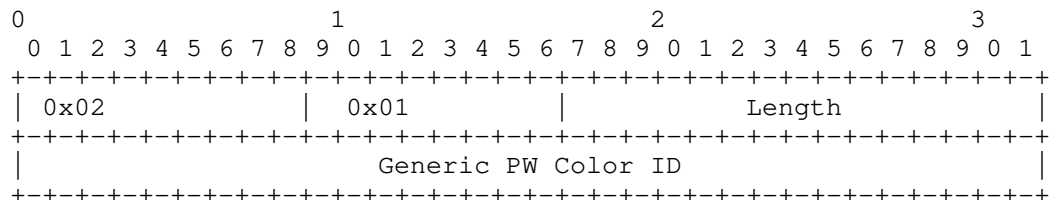


3.2. PW Color Community

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

3.2.1. PW Generic Color Community

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



4. IANA Considerations

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

5. Security Considerations

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

6. Acknowledgements

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

7. References

7.1. Normative References

- [I-D.ietf-pwe3-dynamic-ms-pw]
Martini, L., Bocci, M., and F. Balus, "Dynamic Placement of Multi Segment Pseudowires", draft-ietf-pwe3-dynamic-ms-pw-14 (work in progress), July 2011.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

7.2. References

- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

Authors' Addresses

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

Email: paul.kwok@alcatel-lucent.com

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

Email: pranjal.dutta@alcatel-lucent.com

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

Email: frederic.jounay@orange-ftgroup.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 02, 2014

Yimin Shen, Ed.
Juniper Networks
Rahul Aggarwal
Arktan, Inc
Wim Henderickx
Alcatel-Lucent
July 01, 2013

PW Endpoint Fast Failure Protection
draft-shen-pwe3-endpoint-fast-protection-04

Abstract

This document specifies a fast mechanism for protecting pseudowires (PWs) against egress endpoint failures, including egress attachment circuit failure, egress PE failure, multi-segment PW terminating PE failure, and multi-segment PW switching PE failure. Designed on the basis of multi-homed CE, PW redundancy, upstream label assignment and context specific label switching, the mechanism enables local repair to be performed by a router upstream adjacent to a failure. In particular, the router can restore PW traffic in the order of tens of milliseconds, by transmitting the traffic to a protector through a pre-established bypass tunnel. Therefore, the mechanism can reduce traffic loss before global repair reacts to the failure and the network converges on the topology changes due to the failure.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 02, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Specification of Requirements	4
3. Reference Models and Failure Cases	4
3.1. Single-Segment PW	4
3.2. Multi-Segment PW	6
4. Theory of Operation	7
4.1. Local Repair and Protector	8
4.2. Context Identifier	10
4.2.1. Semantics	10
4.2.2. Advertisement and Path Computation	11
4.3. Protection Models	12
4.3.1. Co-located Protector	12
4.3.2. Centralized Protector	13
4.4. Transport Tunnel	15
4.5. Bypass Tunnel	15
4.6. Forwarding State on Protector	16
4.6.1. Examples of Co-located Protector	16
4.6.2. Examples of Centralized Protector	17
5. LDP Extensions	17
5.1. Egress Protection Capability TLV	18
5.2. PW Label Distribution from Primary PE to Protector	19
5.3. PW Label Distribution from Backup PE to Protector	20
5.4. Protection FEC Element TLV	20
5.4.1. Encoding Format for PWid	21
5.4.2. Encoding Format for Generalized PWid	22
6. Revertive Behavior	24
7. IANA Considerations	25
8. Security Considerations	25
9. Acknowledgements	25
10. References	25
10.1. Normative References	26
10.2. Informative References	27
Authors' Addresses	27

1. Introduction

Per RFC 3985, RFC 4447 and RFC 5659, a pseudowire (PW) or PW segment can be thought of as a connection between a pair of forwarders hosted by two PEs, carrying an emulated layer-2 service over a packet switched network (PSN). In the single-segment PW (SS-PW) case, a forwarder binds a PW to an attachment circuit (AC). In the multi-segment PW (MS-PW) case, a forwarder on a terminating PE (T-PE) binds a PW segment to an AC, while a forwarder on a switching PE (S-PE) binds one PW segment to another PW segment. In each direction between the PEs, PW packets are transported by a PSN tunnel, which is called a transport tunnel.

In order to protect the layer-2 service against network failures, it is necessary to protect every link and node along the entire data path. For the traffic in a given direction, this include ingress AC, ingress (T-)PE, intermediate routers of transport tunnel, S-PEs, egress (T-)PE, and egress AC. To minimize service disruption upon a failure, it is also desirable that each of these components is protected by a fast protection mechanism based on local repair. Such a mechanism generally involves a bypass path that is pre-computed and pre-installed on the router upstream adjacent to a failure. The bypass path has the property that it can guide traffic around the failure, while remaining unaffected by the topology changes resulting from the failure. Thus, when the failure occurs, the router can invoke the bypass path to achieve fast restoration for the service.

Today, fast protection against ingress AC failure and ingress (T-)PE failure is achievable by using a multi-homed CE and redundant PWs. Fast protection against failure of intermediate router is achievable through RSVP fast-reroute (RFC 4090) or IP/LDP fast-reroute (RFC 5714 and RFC 5286). However, there is a lack of equivalent mechanism against egress AC failure, egress (T-)PE failure, and S-PE failure. For these failures, service restoration has to rely on global repair or control plane repair. Global repair is normally driven by ingress CE or ingress (T-)PE, and dependent on status notification or end-to-end OAM. Control plane repair is dependent on protocol convergence. Therefore, both mechanisms are relatively slow in reacting to the failures and restoring traffic.

This document is intended to serve the above need. It specifies a fast protection mechanism based on local repair technique to protect PWs against the following egress endpoint failures.

- a. Egress AC failure.
- b. Egress PE failure: Node failure of an egress PE of an SS-PW, or a T-PE of an MS-PW.
- c. Switching PE failure: Node failure of an S-PE of an MS-PW.

The mechanism is applicable to LDP signaled PWs. It is relevant to networks with redundant PWs and multi-homed CEs. It is designed on the basis of MPLS upstream label assignment and context-specific label switching (RFC 5331). Fast protection refers to the ability to restore traffic upon a failure in the order of tens of milliseconds. This is achieved by establishing local protection at the router upstream adjacent to an anticipated failure. Compared with the existing global repair and control plane repair, this mechanism can provide faster service restoration. However, it is intended to complement those mechanisms, rather than replacing them in any way.

2. Specification of Requirements

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.

3. Reference Models and Failure Cases

This document refers to the following topologies to describe failure scenarios and protection procedures. These topologies involve multi-homed CEs and redundant PWs, which are commonly seen in networks with global repair mechanisms. The mechanism in this document will also use these topologies for local repair purposes. This SHALL enable local repair and global repair to work in tandem to achieve broader coverage of protection for services.

3.1. Single-Segment PW

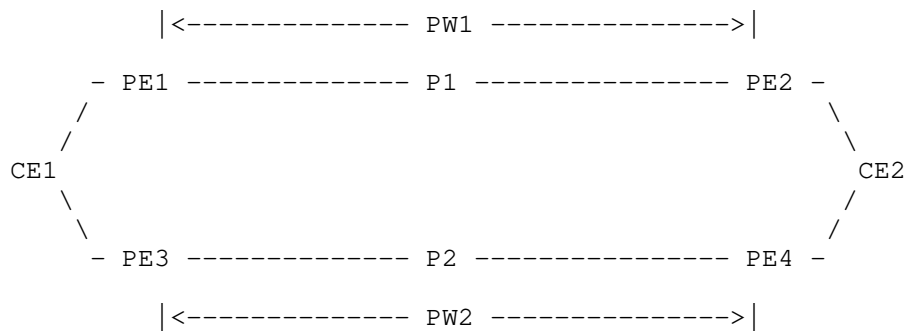


Figure 1

In Figure 1, the IP/MPLS network consists of PE-routers and P-routers. It provides an emulation of a layer-2 service between CE1 and CE2.

Each CE is multi-homed to two PEs. Hence, there are two divergent paths between the CEs. The first path uses PW1 established between PE1 and PE2, connecting the AC CE1-PE1 and the AC CE2-PE2. The second path uses PW2 established between PE3 and PE4, connecting the AC CE1-PE3 and the AC CE2-PE4. The operational states of all the PWs and ACs are up. The transport tunnels of the PWs are not shown in this figure for clarity.

At any given time, each CE sends traffic via only one AC and receives traffic via only one AC. The two ACs MAY or MAY NOT be the same. The AC used to send traffic is determined by the CE, and MAY rely on an end-to-end OAM mechanism between the CEs. The AC used for the CE to receive traffic is determined by the state of the network and the protection mechanism in use, as described later in this document.

From the perspective of traffic flowing towards a given CE, the set of PWs, PEs and ACs involved can be viewed to serve primary and backup (or active and standby) roles. When the network is in a steady state, the PW that is intended to carry the traffic is referred to as a primary PW. The PE at the egress of the primary PW is a primary PE. The AC connecting the CE and the primary PE is a primary AC. The other PW may be used to carry the traffic upon a network failure, and is referred to as a backup PW. The PE at the egress of the backup PW is a backup PE. The AC connecting the CE and the backup PE is a backup AC.

In this document, the following primary and backup roles are assigned for the traffic going from CE1 to CE2:

Primary PW: PW1

Primary PE: PE2

Primary AC: CE2-PE2

Backup PW: PW2

Backup PE: PE4

Backup AC: CE2-PE4

In this case, an egress AC failure refers to the failure of the AC CE2-PE2. An egress node failure refers to the failure of PE2.

The backup PE, backup PW and backup AC may be used to carry traffic after a PW endpoint failure, when CE1 and CE2 switches traffic to PW2 in local repair or global repair, as described later in this document.

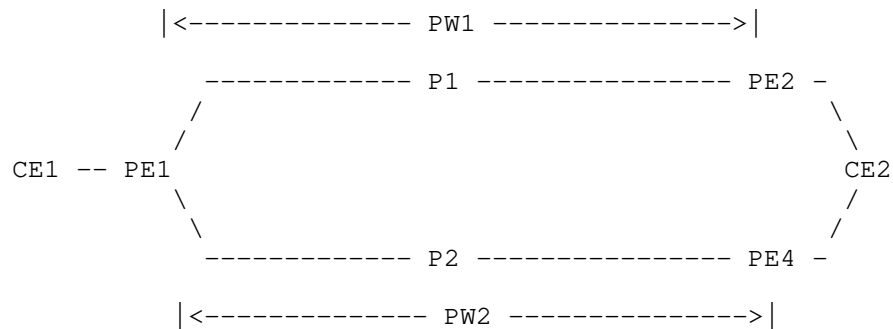


Figure 2

Figure 2 shows another possible scenario, where CE1 is single-homed to PE1, while CE2 remains multi-homed to PE2 and PE4. From the perspective of egress protection for the traffic from CE1 to CE2, this topology is not much different than Figure 1. However, for the traffic in the direction from CE2 to CE1, PE1 must anticipate traffic on both PW1 and PW2, and sends it to CE1 over the AC CE1-PE1.

3.2. Multi-Segment PW

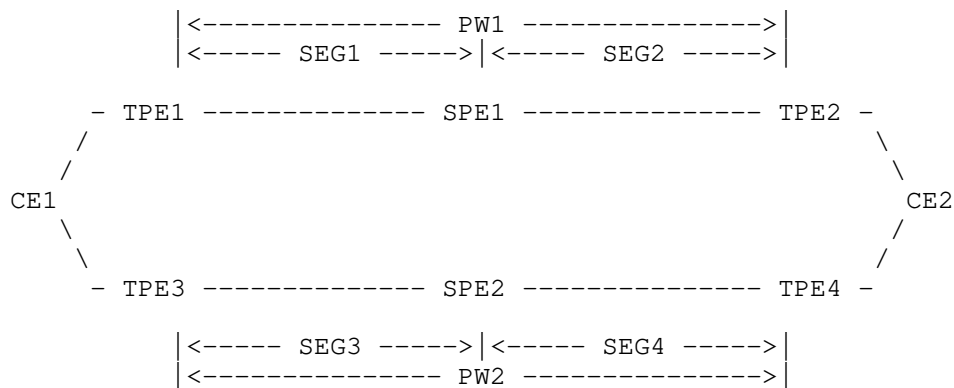


Figure 3

Figure 3 shows a topology that is similar to Figure 1 but in an MS-PW environment. PW1 and PW2 are both MS-PWs. PW1 is established between TPE1 and TPE2, and switched between segments SEG1 and SEG2 at SPE1. PW2 is established between TPE3 and TPE4, and switched between segments SEG3 and SEG4 at SPE2. CE1 is multi-homed to TPE1 and TPE3. CE2 is multi-homed to TPE2 and TPE4. The transport tunnels of the PW segments are not shown in this figure for clarity.

In this document, the following primary and backup roles are assigned for the traffic going from CE1 to CE2:

Primary PW: PW1

Primary T-PE: TPE2

Primary S-PE: SPE1

Primary AC: CE2-TPE2

Backup PW: PW2

Backup T-PE: TPE4

Backup S-PE: SPE2

Backup AC: CE2-TPE4

In this case, an egress AC failure refers to the failure of the AC CE2-TPE2. An egress node failure refers to the failure of TPE2. A switching node failure refers to the failure of SPE1.

The backup T-PE, backup PW and backup AC are used for protecting the primary PW against egress AC failure and egress node failure. The backup S-PE and the backup PW are used for protecting the primary PW against switching node failure, as described later in this document.

For consistency with the SS-PW scenario, primary T-PEs and a primary S-PEs may simply be referred to as primary PEs in this document, where specifics is not required. Similarly, backup T-PEs and backup S-PEs may be referred to as backup PEs.

4. Theory of Operation

The fast protection mechanism in this document provides three types of protection for PWs, corresponding to the three types of failures described in Section 1.

- a. Egress AC protection
- b. Egress (T-)PE node protection
- c. S-PE node protection

The mechanism assumes a multi-homing connectivity from the target CE to a primary PE and a backup PE, and the existence of a backup PW in the network. In S-PE node protection, it also assumes the existence of a backup S-PE on the backup PW.

4.1. Local Repair and Protector

The mechanism relies on local repair to be performed by routers upstream adjacent to failures. Each of these routers is referred to as a "point of local repair" (PLR). A PLR MUST be able to detect a failure by using a rapid mechanism, such as physical layer failure detection, Bidirectional Failure Detection (BFD) (RFC 5880), etc. In anticipation of the failure, the PLR MUST also pre-establish a bypass PSN tunnel to a "protector", and pre-install a bypass route in the FIB (forwarding information base). The bypass tunnel MUST have the property that it is not affected by the topology changes caused by the failure. Upon detecting the failure, the PLR MUST invoke the bypass route in the data plane, and reroute PW traffic to the protector through the bypass tunnel. The protector MUST in turn send the traffic to the target CE. This procedure is referred to as local repair.

Different routers may serve as PLR and protector in different scenarios.

- o In egress AC protection, the PLR is the primary PE that terminates the primary PW and hosts the primary AC. The protector is the backup PE (Figure 4).

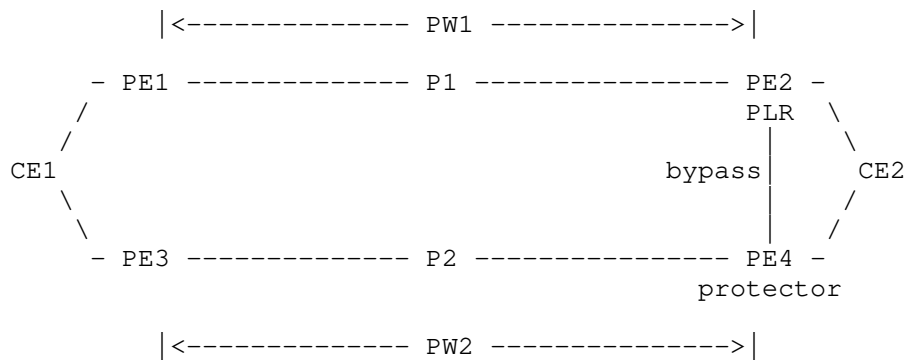


Figure 4

- o In egress PE node protection, the PLR is the penultimate hop router of the transport tunnel of the primary PW, and the protector is the backup PE (Figure 5).

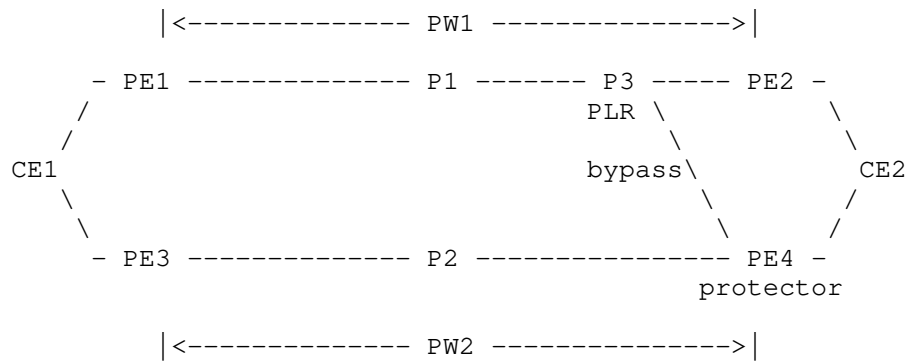


Figure 5

- o In S-PE node protection, the PLR is the penultimate hop router of the transport tunnel of the primary PW segment, and the protector is the backup S-PE (Figure 6).

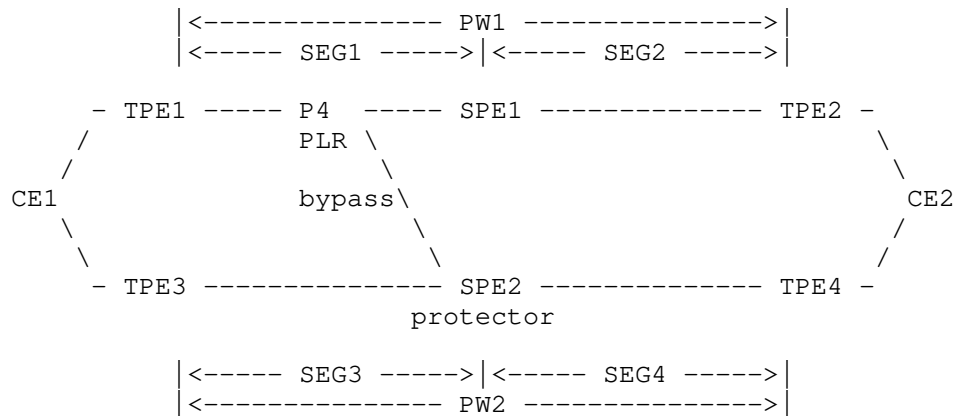


Figure 6

A PLR can realize its role based on configuration or the signaling of transport tunnel. For example, in the case where the transport tunnel is signaled by RSVP, the penultimate hop router could realize that it is the PLR for egress (T-)PE or S-PE failure based on the RRO in Resv message, which should indicate to the router that it is one hop away from the PE. The detail of how this could be achieved on a per-protocol basis is out of the scope of this document.

In all scenarios, when a PLR reroutes traffic through a bypass tunnel to a protector during local repair, it MUST keep the label of the primary PW intact in the packets. This obviates the need for the PLR to maintain forwarding state on a per-PW basis, and allows a single bypass tunnel to protect multiple PWs.

The procedure also requires that the protector SHOULD be able to forward the traffic based on a PW label that is assigned by the primary PE, and ensure the traffic to eventually reach the target CE. From the protector's perspective, this PW label is an upstream assigned label (RFC 5331). To accomplish this, the protector SHOULD learn the PW label from the primary PE prior to the failure, and install proper forwarding state for the PW label in a dedicated label space of the primary PE. During local repair, the protector SHOULD perform PW label lookup in this label space.

The above examples have shown the scenarios where the protectors are backup (S-)PEs. In other scenarios, a protector may be a dedicated router that assumes such role, separate from the backup (S-)PE of a primary PW. During local repair, the PLR MUST still reroute traffic to the protector through a bypass tunnel. The protector MUST then send the traffic to the backup (S-)PE, which MUST in turn send the traffic to the target CE via a backup AC or a backup PW segment. More detail will be described in Section 4.3.

4.2. Context Identifier

A protector MAY serve the protection for multiple primary PEs. The protector MUST maintain a separate label space for each primary PE. Likewise, the PWs terminated on a primary PE MAY be protected by multiple protectors, each for a subset of the PWs. In any case, a given primary PW is associated with one and only one pair of {primary PE, protector}.

An IPv4/v6 address is assigned to each ordered pair of {primary PE, protector} to facilitate protection establishment. This address is referred to as a "context identifier". It MUST be globally unique, or unique in the address space of the network where the primary PE and the protector reside.

4.2.1. Semantics

The semantics of a context identifier is twofold.

- o It identifies a primary PE and an associated protector. In other words, it identifies a primary PE on a per protector basis. A given primary PE may be protected by multiple protectors, each for a subset of the primary PWs terminated on the primary PE. A

distinct context identifier MUST be assigned to the primary PE and each protector.

For each primary PW, its ingress PE MUST set up a transport tunnel with destination as the context identifier of the {primary PE, protector}, rather than a private IP address of the primary PE. This not only allows the transport tunnel to be set up to the primary PE, but also conveys the identity of the protector to the PLR(s) along the transport tunnel. Each PLR can in turn use this information to set up a bypass tunnel to the protector without relying on local configuration.

- o It identifies the primary PE's label space on the protector. The protector may protect PWs for multiple primary PEs. For each primary PE, it MUST maintain a separate label space to store the PW labels assigned by that primary PE. It MUST associate a PW label with a label space via the context identifier of the {primary PE, protector}, as below.

In addition to the normal LDP PW signaling, the primary PE MUST have a targeted LDP session with the protector, and advertise PW labels to the protector via LDP Label Mapping messages (See Section 5 for detail). The primary PE MUST also attach the context identifier to each message. Upon receiving the message, the protector MUST install the advertised PW label in the label space identified by the context identifier.

When a PLR sets up a bypass tunnel to the protector, it MUST set the destination to the context identifier, rather than a private IP address of the protector. Once established, the bypass tunnel, with either its MPLS label or IP tunnel destination address in IP header, is used as the identifier of the label space. On the protector, all PW packets received on the bypass tunnel MUST be forwarded based on a label lookup in that label space.

4.2.2. Advertisement and Path Computation

Using a context identifier as destination for both transport tunnel and bypass tunnel requires both the primary PE and the protector to advertise the context identifier via IGP as an IP address reachable through both routers in routing domain and/or TE domain. This imposes the following requirements on path computation for these tunnels.

- o For the transport tunnel, the ingress PE MUST choose the primary PE as the actual endpoint.

- o For the bypass tunnel, the PLR MUST choose the protector as the actual endpoint. In egress (T-)PE node protection and S-PE node protection, the bypass tunnel MUST avoid the primary (S-)PE.

The detail of how the primary PE and the protector may advertise a context identifier is independent of this mechanism and out of the scope of this document. One approach would be to advertise it as a virtual proxy node connected to both routers, with the link between the proxy node and the primary PE having a more preferable IGP or TE metric than the link between the proxy node and the protector. The ultimate goal is for a path computation algorithm, such as CSPF (constrained shortest path first), LFA (RFC 5286) and MRT ([IP-LDP-FRR-MRT]), to be able to compute the paths that meet the above requirements.

4.3. Protection Models

There are two protection models based on the location of a protector. A network MAY use either model, or a combination of both.

4.3.1. Co-located Protector

In this model, the protector is a backup PE that is directly connected to the target CE via a backup AC, or it is a backup S-PE on a backup PW. That is, the protector is co-located with the backup (S-)PE. Examples of this model have been introduced in Figure 4, Figure 5 and Figure 6 in Section 4.1.

In egress AC protection and egress PE node protection, when a protector receives traffic from the PLR, it forwards the traffic to the CE via the backup AC. This is shown in Figure 7, where PE2 is the PLR for egress AC failure, P3 is the PLR for PE2 failure, and PE4 (the backup PE) is the protector.

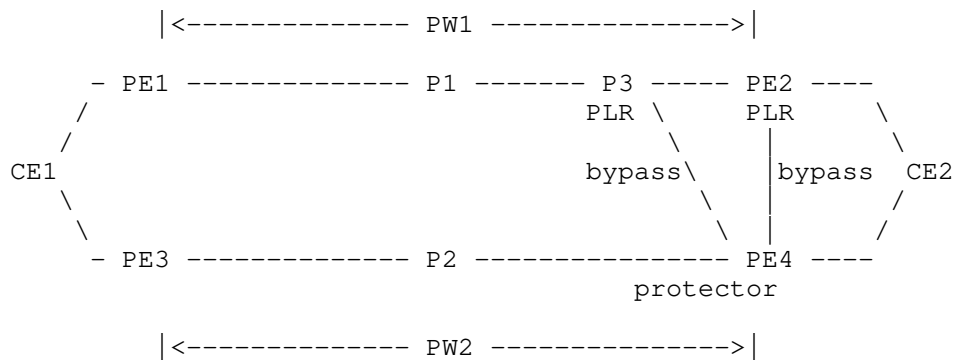


Figure 7

In S-PE node protection, when a protector receives traffic from the PLR, it MUST forward the traffic via the next segment of the backup PW. The T-PE of the backup PW MUST forward the traffic to the CE via a backup AC. This is shown in Figure 8, where P4 is the PLR for SPE1 failure, and SPE2 (the backup S-PE) is the protector for SPE1 (the primary S-PE).

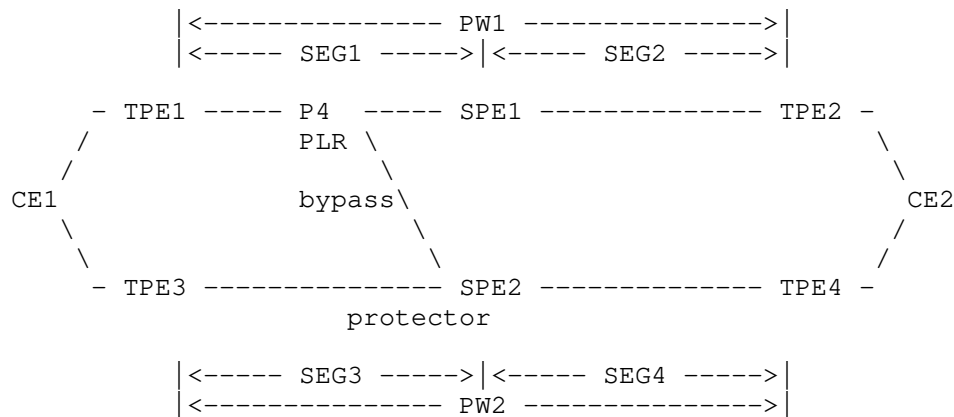


Figure 8

In the co-located protector model, the number of context identifiers needed by a network is the number of distinct {primary PE, backup PE} pairs. From the perspective of scalability, the model is suitable for networks where the number of backup PEs for any given primary PE is relatively small.

4.3.2. Centralized Protector

In this model, the protector is a dedicated P router or PE router that serves the role. In egress AC protection and egress PE node protection, the protector MAY or MAY NOT be a backup PE with a direct connection to the target CE. In S-PE node protection, the protector MAY or MAY NOT be a backup S-PE on the backup PW.

In egress AC protection and egress PE node protection, when the protector receives traffic from the PLR, if the protector has a direct connection (i.e. backup AC) to the CE, it MUST forward the traffic to the CE via the backup AC, which is similar to Figure 7. Otherwise, it MUST forward the traffic to a backup PE, which MUST then forward the traffic to the CE via a backup AC. This is shown in Figure 9, where the protector receives traffic from P3 or PE2 (the PLRs) and forwards the traffic to PE4 (the backup PE). The protector may be protecting other PWs as well, which is not shown in this figure.

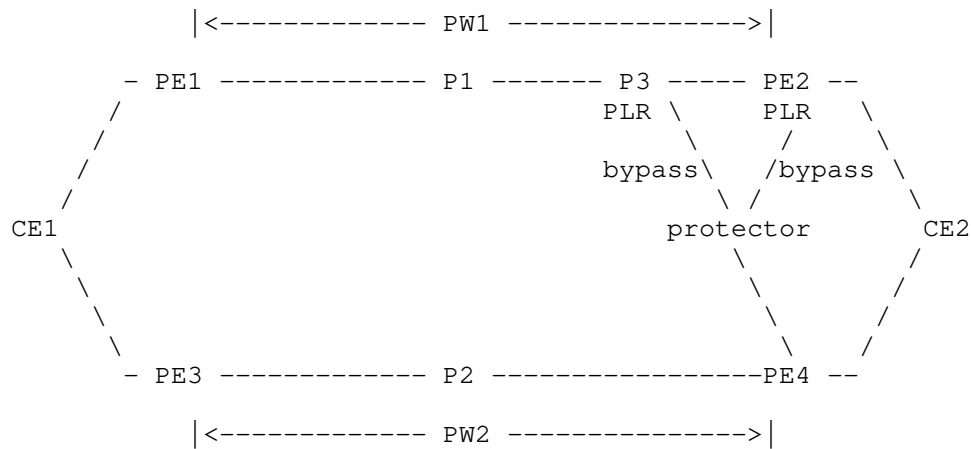


Figure 9

In S-PE node protection, when the protector receives traffic from the PLR, if the protector is a backup S-PE of the backup PW, it MUST forward the traffic via the next segment of the backup PW, and the T-PE of the backup PW MUST forward the traffic to the CE via a backup AC, which is similar to Figure 8. Otherwise, the protector MUST first forward the traffic to the backup S-PE, which MUST then forward the traffic via the next segment of the backup PW. Finally, the T-PE of the backup PW MUST forward the traffic to the CE via a backup AC. This is shown in Figure 10, where the protector forwards traffic to SPE2 (the backup S-PE). The protector may be protecting other PW segments as well, which is not shown in this figure.

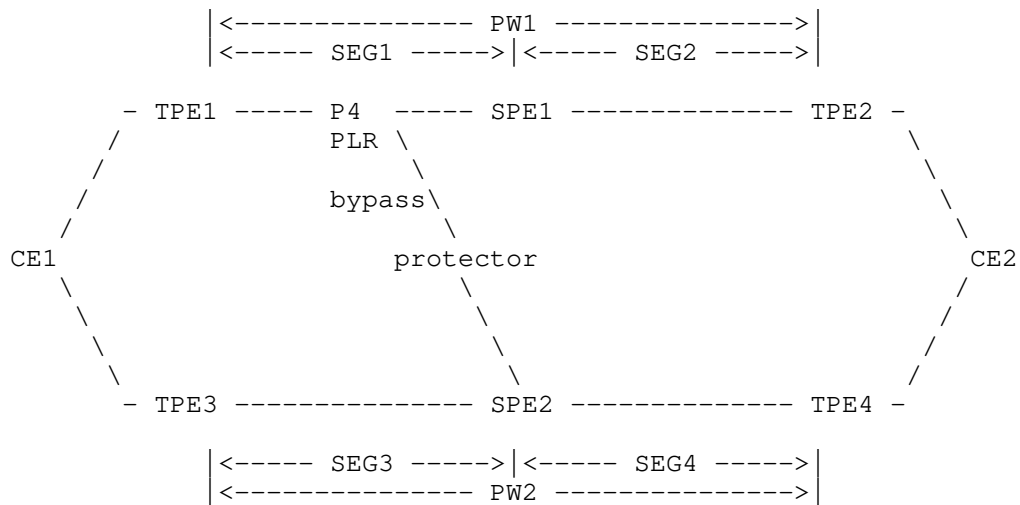


Figure 10

In the centralized protector model, each primary PE MAY only need one protector to protect all of its PWs. From the perspective of scalability, the number of context identifiers needed by a network can be as low as the number of primary PEs.

4.4. Transport Tunnel

The ingress PE of a primary PW (or PW segment) associates the PW with the primary egress PE through LDP signaling. In addition, as mentioned in Section 4.2.1, the ingress PE MUST associate the transport tunnel of the PW with the context identifier of the {primary PE, protector}, and set up the transport tunnel by using the context identifier as destination. This not only ensures that PW traffic be transported to the primary PE, but also facilitates bypass tunnel establishment at PLR(s), as the context identifier implies the identity of the protector as well.

The association between the transport tunnel and the context identifier at the ingress PE MAY be achieved by configuration or an auto-discovery mechanism. In the later case, the ingress PE MAY learn the context identifier from the primary (egress) PE, if the primary PE advertises the context identifier as "third party next hop" in IPv4/v6 Interface_ID TLV (RFC 3471, RFC 3472) in the LDP Label Mapping message of the primary PW.

4.5. Bypass Tunnel

A PLR may protect multiple PWs associated with one or multiple pairs of {primary PE, protector}. The PLR MUST establish a bypass tunnel to each protector for each distinct context identifier associated with that protector. The destination of the bypass tunnel MUST be the context identifier (Section 4.2.1). The PLR may derive the context identifier from the destination of the transport tunnel that traverses it.

For examples, in Figure 7 and Figure 9, a bypass tunnel is established from PE2 (PLR for egress AC failure) to the protector, and another bypass tunnel is established from P3 (PLR for egress node failure) to the protector. In Figure 8 and Figure 10, a bypass tunnel is established from P4 (PLR for switching node failure) to the protector.

During local repair, the PLR reroutes traffic to the protector through the bypass tunnel with PW label intact in the packets. This normally involves pushing a label to the label stack, if the bypass tunnel is an MPLS tunnel, or pushing an IP header to the packets, if the bypass tunnel is an IP tunnel. The protector MUST in turn forward the traffic based on the PW label. To achieve such kind of forwarding, the protector MUST rely on the bypass tunnel as a context to determine the primary PE's label space. If the bypass tunnel is an MPLS tunnel, the protector MUST assign a non-reserved label to the bypass tunnel during the signaling of the bypass tunnel, and treat this label as the context. If the bypass tunnel is an IP tunnel, the protector can know the context directly based on the context identifier carried as destination address in IP header.

A bypass tunnel MUST have the property that it is not affected by the topology changes caused by the failure. Therefore, it can be used to transmit traffic for local repair. It SHOULD remain effective, until the traffic is moved to another fully functional egress AC, PW and/or transport tunnel.

4.6. Forwarding State on Protector

A protector MUST learn PW labels from all the primary PEs that it protects (Section 5.2), and maintain the PW labels in respective label spaces of the primary PEs. In the control plane, a label space is identified by the context identifier of a pair of {primary PE, protector}. In the forwarding plane, it is indicated by the bypass tunnel(s) destined for the context identifier.

4.6.1. Examples of Co-located Protector

In Figure 7, PE4 is a co-located protector that protects PW1 against egress AC failure and egress node failure. It maintains a label

space for PE2, which is identified by the context identifier of {PE2, PE4}. It learns PW1's label from PE2, and installs an forwarding entry for the label in that label space. The nexthop of the forwarding entry indicates a label pop with outgoing interface pointing to the backup AC CE2-PE4.

In Figure 8, SPE2 is a co-located protector that protects PW1 against switching node failure. It maintains a label space for SPE1, which is identified by the context identifier of {SPE1, SPE2}. It learns SEG1's label from SPE1, and installs a forwarding entry in the label space. The nexthop of the forwarding entry indicates a label swap to SEG4's label.

4.6.2. Examples of Centralized Protector

In the centralized protector model, for each primary PW of which the protector is not a backup (S-)PE, the protector MUST also learn the label of the backup PW from the backup (S-)PE (Section 5.3). This is the backup (S-)PE that the protector will forward traffic to. The protector MUST install a forwarding entry with label swap from the primary PW's label to the backup PW's label.

In Figure 9, the protector is a centralized protector that protects PW1 against egress AC failure and egress node failure. It maintains a label space for PE2, which is identified by the context identifier of {PE2, protector}. It learns PW1's label from PE2, and PW2's label from PE4. It installs a forwarding entry for PW1's label in the label space. The nexthop of the forwarding entry indicates a label swap to PW2's label.

In Figure 10, the protector is a centralized protector that protects the PW segment SEG1 of PW1 against switching node failure of SPE1. It maintains a label space for SPE1, which is identified by the context identifier of {SPE1, protector}. It learns SEG1's label from SPE1, and learns SEG3's label from SPE2. It installs a forwarding entry for SEG1's label in the label space. The nexthop of the forwarding entry indicates a label swap to SEG3's label.

5. LDP Extensions

As described in previous sections, a targeted LDP session MUST be established between each pair of primary PE and protector. The primary PE sends Label Mapping message over this session to advertise a primary PW's label to the protector. In the centralized protector model, a targeted LDP session MUST also be established between a backup (S-)PE and a protector. The backup PE sends Label Mapping message over this session to advertise a backup PW's label to the protector.

To facilitate the procedures, this document defines a new "Protection FEC Element" TLV. The Label Mapping messages of both the LDP sessions above MUST carry this TLV to indicate the identity of the primary PW. Specifically, in the centralized protector model, the Protection FEC Element TLV advertised by a backup (S-)PE MUST match the one advertised by the primary PE, so that the protector can associate the primary PW's label with the backup PW's label, and perform a label swap.

This document also defines the encoding of Capability Parameter TLV (RFC 5561) for a new "Egress Protection Capability", to allow a protector to announce its capability of processing the above Protection FEC Element TLV and performing context specific label switching for PW labels.

The procedures in this section are only applicable, if the protector advertises the Egress Protection Capability, the primary PE supports the advertisement of the Protection FEC Element TLV, and in the centralized protector model, the backup PE also supports the advertisement of the Protection FEC Element TLV.

5.1. Egress Protection Capability TLV

A protector MUST advertise the Egress Protection Capability TLV in its Initialization message and Capability message, over the LDP session with a primary PE or a backup PE. The TLV carries the context identifier associated with the {primary PE, protector}. This TLV SHOULD NOT be advertised by the primary PE or the backup PE to the protector.

The processing of the Egress Protection Capability TLV by a receiving router SHOULD follow the procedures defined in RFC 5561. In particular, the router SHOULD advertise PW information to the protector by using the Protection FEC Element TLV, only after it has received the Egress Protection Capability TLV from the protector. It SHOULD validate the context identifier included in the TLV, and advertise the information of only those PWs that are associated with the context identifier. It SHOULD withdraw previously advertised Protection FEC TLVs, when the protector has withdrawn the Egress Protection Capability TLV via Capability message.

The encoding of the Egress Protection Capability TLV is defined as below. It conforms to the format of Capability Parameter TLV specified in RFC 5561.

```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+

```

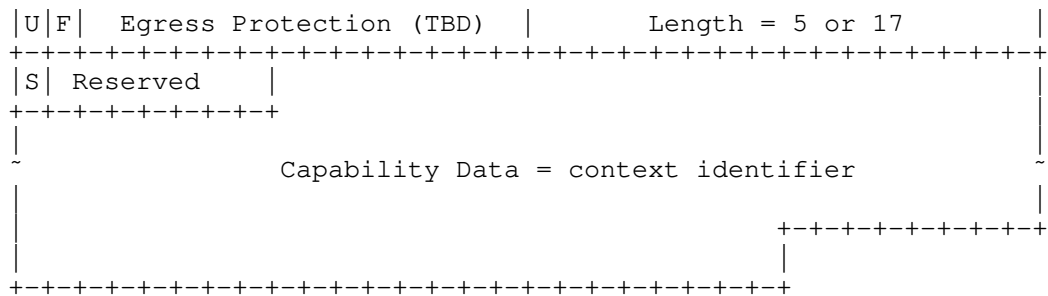


Figure 11

The U-bit MUST be set to 1 so that a receiver MUST silently ignore this TLV if unknown to it, and continue processing the rest of the message.

The F-bit MUST be set to 0 since this TLV is sent only in Initialization and Capability messages, which are not forwarded.

The TLV Code Point is TBD. It needs to be assigned by IANA.

The S-bit indicates whether the sender is advertising (S=1) or withdrawing (S=0) the capability.

The "Capability Data" is encoded with the context identifier of the {primary PE, protector}. Hence, the Length of the TLV MUST be set to 5 if the context identifier is an IPv4 address, or 17 if it is an IPv6 address.

5.2. PW Label Distribution from Primary PE to Protector

A primary PE SHOULD advertise a primary PW's label to a protector by sending a Label Mapping message. The message includes a Protection FEC Element TLV (see Section 5.4 for encoding), and an Upstream-Assigned Label TLV (RFC 6389) encoded with the PW's label. The combination of the Protection FEC Element TLV and the PW label represents the primary PE's forwarding state for the PW. The Label Mapping message SHOULD also carry an IPv4/v6 Interface_ID TLV (RFC 6389, RFC 3471) encoded with the context identifier of the {primary PE, protector}.

The protector that receives this Label Mapping message SHOULD install a forwarding entry for the PW label in the label space identified by the context identifier. The nexthop of the forwarding entry SHOULD ensure packets to be sent towards the target CE via a backup AC or a backup (S-)PE, depending on the protection scenario. The protector SHOULD silently drop a Label Mapping message if the included context identifier is unknown to it.

5.3. PW Label Distribution from Backup PE to Protector

In the centralized protector model, a backup PE SHOULD advertise a backup PW's label to a protector by sending a Label Mapping message. The message includes a Protection FEC Element TLV and a Generic Label TLV encoded with the backup PW's label. This Protection FEC Element MUST be identical to the Protection FEC Element TLV that the primary PE advertises to the protector (Section 5.2). The context identifier SHOULD NOT be encoded in Interface_ID TLV in this message.

The protector that receives this Label Mapping message SHOULD associate the backup PW with the primary PW, based on the common Protection FEC Element TLV. It SHOULD distinguish between the Label Mapping message from the primary PE and the Label Mapping message from the backup PE based on the respective presence and absence of context identifier in Interface_ID TLV. It SHOULD install a forwarding entry for the primary PW's label in the label space identified by the context identifier. The nexthop of the forwarding entry SHOULD indicate a label swap to the backup PW's label, followed by a label push or IP header push for a transport tunnel to the backup PE.

5.4. Protection FEC Element TLV

The Protection FEC Element TLV has type 0x83. Its format is defined as below:

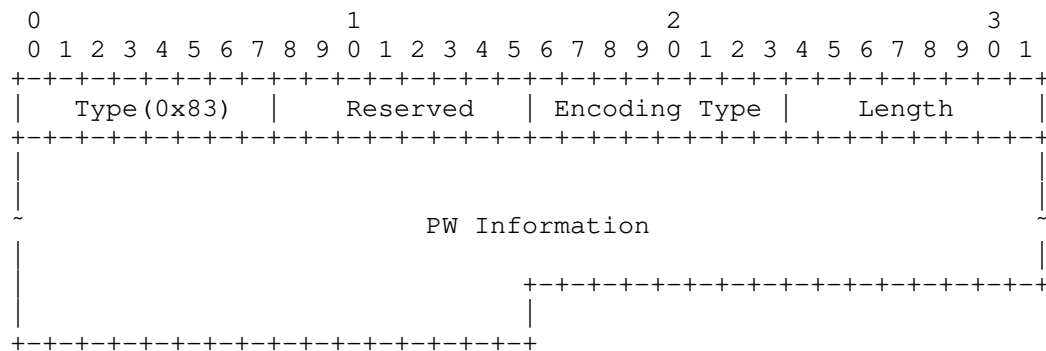


Figure 12

- Encoding Type

Type of format that PW Information field is encoded.

- Length

Length of PW Information field in octets.

- PW Information

Field of variable length that specifies a PW

For Encoding Type, 1 is defined for the PWidth FEC Element format, and 2 is defined for the Generalized PWidth FEC Element format (RFC 4447).

5.4.1. Encoding Format for PWin

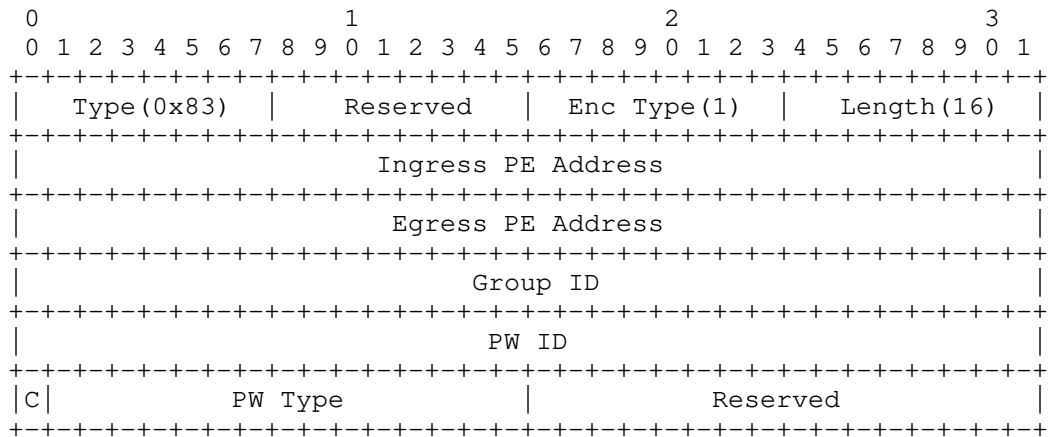


Figure 13

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Group ID

An arbitrary 32-bit value that represents a group of PWs and that is used to create groups in the PW space.

- PW ID

A non-zero 32-bit connection ID that, together with the PW Type field, identifies a particular PW.

- Control word bit (C)

A bit that flags the presence of a control word on this PW. If C = 1, control word is present; If C = 0, control word is not present.

- PW Type

A 15-bit quantity that represents the type of PW.

5.4.2. Encoding Format for Generalized PWid

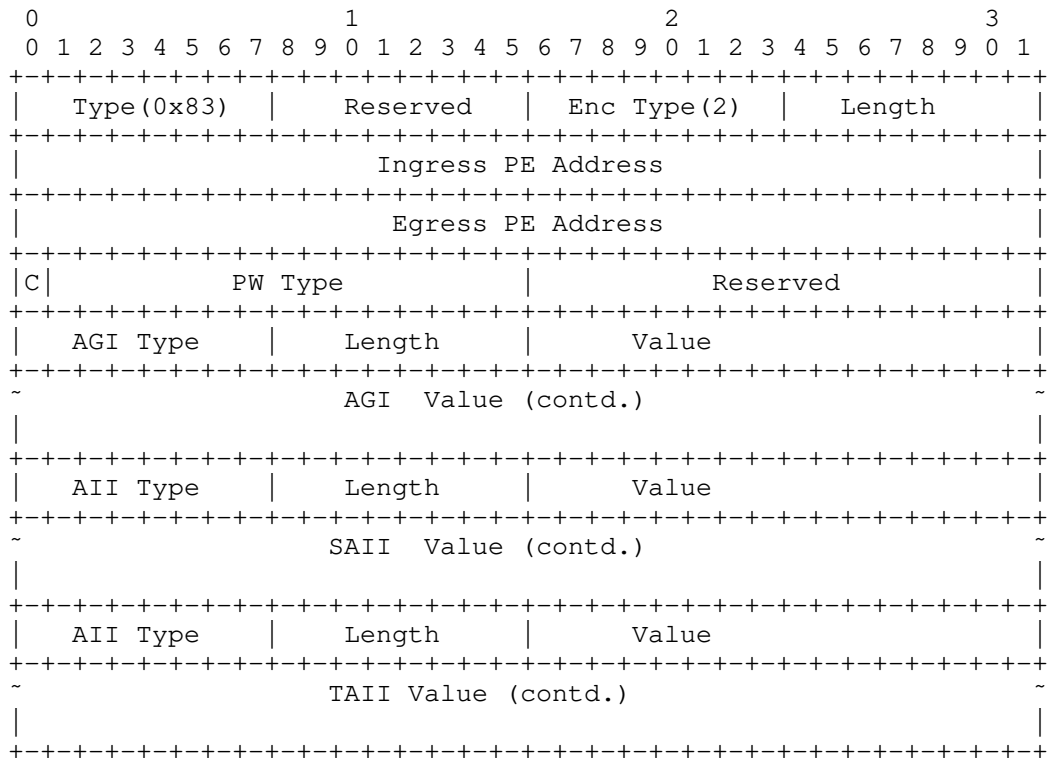


Figure 14

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Control word bit (C)

A bit that flags the presence of a control word on this PW. If C = 1, control word is present; If C = 0, control word is not present.

- PW Type

A 15-bit quantity that represents the type of PW.

- AGI Type, Length, Value, AGI Value

Attachment Group Identifier of PW.

- SAII Type, Length, Value, SAII Value

Source Attachment Individual Identifier of PW.

- TAII Type, Length, Value, TAII Value

Target Attachment Individual Identifier of PW.

6. Revertive Behavior

Subsequent to local repair, there are three strategies for the network to restore traffic to a fully functional PW.

- o Global revertive mode

If the ingress CE is multi-homed (Figure 1), it MAY switch the traffic to a backup AC which is bound to a backup PW. Alternatively, if the CE is single-homed to the ingress PE whereas the ingress PE hosts a backup PW (Figure 2), the ingress PE MAY switch the traffic to the backup PW. These procedures are referred to as global repair. Possible triggers of a global repair include PW status, OAM, and BFD.

- o Control plane revertive mode

In egress PE node protection and S-PE node protection, it is possible that the failure is limited to the link between the PLR and the primary (S-)PE, whereas the primary (S-)PE is still up. In this case, the PLR or an upstream router along the transport tunnel MAY reroute the tunnel around the failed link via an alternative path. Thus, the transport tunnel can continue to be used to carry the PW traffic to the primary (S-)PE. This procedure is driven by control plane convergence, and is referred to as control plane repair.

- o Local revertive mode

The PLR MAY move traffic back to the primary PW, after the failure is resolved. In egress AC protection, upon detecting that the primary AC is restored, the PLR MAY start forwarding traffic over the AC again. Likewise, in egress PE node protection and S-PE node protection, upon detecting that the primary PE is restored, the PLR MAY re-establish the primary transport tunnel through the primary PE, and move the traffic from the bypass tunnel back to the transport tunnel. These procedures are referred to as local reversion.

The fast protection mechanism in this document SHOULD be used in tandem with the global revertive mode. Particularly in the case of egress (S-)PE failure, if the ingress PE or the protector loses communication with the (S-)PE for an extensive period of time, the LDP session between them may go down. Consequently, the ingress PE may bring down the primary PW, or the protector may remove the forwarding entry of the primary PW label. In either case, the service will be disrupted. In other words, although the fast protection can temporarily repair traffic, control plane state may eventually time out if the failure persists. Therefore, it is recommended that the global revertive mode SHOULD be set up in advance, so that traffic can be moved to a fully functional backup PW shortly after the local repair.

The control plane revertive mode may happen as part of the convergence of control plane protocols. It is only applicable to some specific topologies.

The local revertive mode is optional. In the circumstances where the failure is caused by resource flapping, local reversion MAY be dampened to limit potential disruptions. Local revertive mode MAY be disabled completely by configuration.

7. IANA Considerations

This document defines the encoding of the Capability Parameter TLV for the new "Egress Protection Capability" in Section 5. This would require IANA to assign a TLV Code Point to it.

This document defines a new LDP Protection FEC Element TLV in Section 5. IANA has assigned the type value 0x83 to it.

8. Security Considerations

The security considerations discussed in RFC 5036, RFC 5331, RFC 3209, and RFC 4090 apply to this document.

9. Acknowledgements

This document leverages work done by Hannes Gredler, Yakov Rekhter, Minto Jeyanthan and several others on MPLS edge protection. Thanks to Nischal Sheth, Bhupesh Kothari, and Kevin Wang for their contribution. Thanks to Yakov Rekhter and John E Drake for reviewing the document. Thanks to Andrew G Malis for valuable comments.

10. References

10.1. Normative References

- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC5659] Bocci, M. and S. Bryant, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.
- [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.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [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.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

- [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC6389] Aggarwal, R. and JL. Le Roux, "MPLS Upstream Label Assignment for LDP", RFC 6389, November 2011.
- [IP-LDP-FRR-MRT]
Atlas, A. and R. Kebler, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees ", draft-ietf-rtgwg-mrt-frr-architecture (work in progress), 2011.

10.2. Informative References

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

Authors' Addresses

Yimin Shen (editor)
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Phone: +1 9785890722
Email: yshen@juniper.net

Rahul Aggarwal
Arktan, Inc

Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.be