Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 12, 2011

Yuanlin Bao
Lizhong Jin
ZTE Corporation
Ruiquan Jing
Xiaoli Huo
China Telecom
Oct 9, 2010

LDP Extensions for Pseudo Wire (PW) Transfer in an MPLS-TP Network

draft-bao-pwe3-pw-transfer-00.txt

Abstract

As defined in [RFC5654] MPLS-TP transport path includes LSP and PW. And the possibility of transferring the ownership and control of an existing and in-use path between the management plane and the control plane, without actually affecting data plane traffic being carried over it, is a valuable option for carrier. [RFC5493] and [RFC5852] describe the LSP transfer. This memo gives the requirement and LDP extensions for PW transfer in an MPLS-TP network.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

As defined in [RFC5654], MPLS-TP transport path corresponds to an LSP or a PW which is beared in an LSP. And LSP includes unidirectional LSP, co-routed bidirectional LSP and associated bidirectional LSP, while PW includes Single-Segment Pseudowire (SS-PW) and Multi-Segment Pseudowire (MS-PW).

For MPLS-TP LSP, it can be created/deleted via GMPLS signaling, see [RFC3945]. However, the creation/deletion of PW can be completed by LDP, and [RFC4447] gives these procedures of SS-PW while [SEG-PW] and [DYNAMIC-MS-PW] decribes the ones of MS-PW.

Nowdays, some service providers have deployed MPLS-TP network for mobile backhaul. But, most of the MPLS-TP paths are statically configured by management plane in the first stage. So, it is desirable for provider to transfer the control of paths from the management plane (MP) to control plane (CP) in future. In addition, the control transfer in the opposite direction, from CP to MP should be possible as well.

Both the requirement 55 in [RFC5654] and requirement 47 in [MPLS-TP-CP-FWK] state that an MPLS-TP control plane MUST provide a mechanism for dynamic ownership transfer of the control of MPLS-TP transport paths from the management plane to the control plane and vice versa. Furthermore, section 5.3.3 of [MPLS-TP-CP-FWK] describes the requirement for PW transfer. So, this memo considers the detailed requirements for PW transfer, and the corresponding LDP extensions is also described.

1.1. Comparison with Make-before-Break

The Make-Before-Break (MBB) technology is an alternative method for PW transfer which has three steps. Firstly, a new PW (has the same parameters with the one to be transferred) will be created; then the PW will be switched from old PW to the new one; and after the PW switching completed successfully the old PW will be deleted. From this process, we can find there’re many drawbacks with MBB.

The creation and swithing steps of MBB will lead to instant interruption which is acceptable if it can be controlled within 50ms. Furthermore, extra resource is need, in the circumstance that the network is almost saturate, there maybe not enough resource for the new PWs, so MBB will be unavailable. Otherwise, MBB will lead to label modification which will make the bundling relationship between PW and LSP must modified at the same time. This will triggre many problems, and a new detection mechanism needs to be defined which may be very complex. In addition, since control plane is used to create the new PW while management plane is responsible for the deletion of the old PW. Thus batch operation cann't be used for this process. If there're a large number PWs needed to be transfered, the operator's time will be engaged by this tedious operation which is inefficiency. However, the PW transfer method described in this document will not affect the data plane, the traffic and it's configuration. So it's preference for PW transfer. However, the PW transfer method described in this document will not affect the data plane, the traffic and it's configuration. So it's preference for PW transfer.

1.2. Conventions used in this document

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

2. Terminology

- o Transport Path: A network connection as defined in G.805 [ITU.G805.2000]. In an MPLS-TP environment, a transport path corresponds to an LSP or a PW (see RFC5654).

- o Single-Segment Pseudowire (SS-PW): A PW setup directly between two T-PE devices. Each PW in one direction of a SS-PW traverses one PSN tunnel that connects the two T-PEs.

- o Multi-Segment Pseudowire (MS-PW): A static or dynamically configured set of two or more contiguous PW segments that behave and function as a single point-to-point PW. Each end of a MS-PW by definition MUST terminate on a T-PE.

- o PW Segment: A part of a single-segment or multi-segment PW, which traverses one PSN tunnel in each direction between two PE devices, T-PEs and/or S-PEs.

- o Resource Ownership: A resource used by an MPLS-TP path is said to be 'owned' by the plane that was used to set up the MPLS-TP path through that part of the network. So, a resource owned by the management/control plane means the resource was used to set up the MPLS-TP path through the management/control plane. See RFC5493 for detailed description.

3. Overview of the PW Transfer

The PW transfer includes two reverse procedures. One is the MP to CP (MP2CP) transfer procedure, another is the CP to MP (CP2MP) transfer procedure.

For MP2CP transfer procedure, a PW set up and owned by MP needs to be transferred to CP control. To conduct this transfer, the T-LDP session will be created in CP for PW. After this transfer procedure, the resource ownership must be transferred, that is the resource owned by MP will be transferred to CP.

The CP2MP transfer procedure is the reverse one compared to MP2CP procedure. However, since a LDP session may be shared by multi PWs, the T-LDP session may be retained after one PW transferring from CP to MP, if there’re still another PWs remain untransferred. So, the CP2MP procedure needs to check whether this signaling session should be retained or not.

As an requirement listed in [RFC5493], during both MP2P and CP2MP transfer procedures, if PW is carrying traffic, its control transfer has to be done without any disruption to the data plane traffic.

Furthermore, both MP2CP and CP2MP transfer procedures can be conducted in a batch manner, that is, multiple LSPs or PWs can be transferred all at one time. For example, all PWs on a node can be transferred at one time. However, this transfer manner is out of this document.

4. Requirements for PW Transfer

[RFC5493] describes the requirements for the conversion between permanent connection (PC) and switched connection (SC) in a GMPLS network. The terminologies "PC" and "SPC" come from ITU-T standard [G.8081], Because associated bidirectional LSP isn’t defined in ITU-T standard. So, both PC and SPC can only be considered as unidirectional LSP and co-routed bidirectional LSP. Therefore, these requirements fully apply to unidirectional LSP and co-routed bidirectional LSP in a MPLS-TP network. Although, some requirements defined in [RFC5493] apply to PW, but other new requirements also need to be explored.

This section lists the special requirements for PW transfer.

1) PW attributes MUST not be changed

The PW attributes, such as bandwith, PWid , PW type, Control Word, VCCV, Interface Parameter, MUST not be changed during and after the PW transfer.

2) PW transfer MUST be independent of LSP

The PW transfer SHOULD not depend on whether the LSP (bearing this PW) is controlled by MP or CP. Since PW transfer procedure will not impact the data plane path, so PW transfer MUST leave LSP alone. The relationship between PW and LSP MUST NOT be changed.

3) Support partial MS-PW segments transfer

Since a MS-PW transit multi domains and these domains may belong to different providers. In this scenario, if some providers have deployed control plane while others not, the PW segments in these domains that control plane are deployed SHOULD be allowed to transfer between MP and CP while other PW segments keep their original states.

5. LDP Extension for PW Transfer

5.1. LDP Extension

5.1.1. Support PW Transfer with LDP

A new Capability Parameter TLV is defined, the PW Transfer Capability. Following is the format of the PW Transfer Capability Parameter.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|0|PW Transfer Capability(TBD)|     Length (= 1)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1| Reserved    |
+-+-+-+-+-+-+-+-+
```

Figure 1: PW Transfer Capability

The PW Transfer Capability TLV MUST be supported in the LDP Initialization Message([RFC5561]). Advertisement of the PW Transfer Capability indicates support of the procedures for PW transfer between MP and CP detailed in this document. If the peer has not advertised the corresponding capability, then no PW transfer label messages should be sent to the peer.

5.1.2. PW Ownership Transfer TLV

To ensure the PW ownership transfer between MP and CP automatically, T-PE/S-PE SHOULD has the knowledge of the PW transfer signaling message. So, the PW path and PW transfer indication MUST be carried in the LDP Label Mapping message.

Since [SEG-PW] has defined PW switching point TLV (S-PE TLV) and Sub-TLV to the switching points that the PW traverses, so these TLV and Sub-TLV can be used to carry the PW path. Therefore, this section only defines a new LDP TLV - Transfer TLV - which can be used to indicate a PW transfer signaling procedure.

The PW Ownership Transfer TLV (PW-OH TLV), is defined as follows (TLV type needs to be assigned by IANA):

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0|   PW Transfer  (0x0105)   |         Length                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|POT|                       Reserved                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: PW Ownership Transfer TLV

POT (2 bits): PW Ownership Transfer. PE MUST carry this TLV in LDP Label Mapping and Notification message defined in [RFC5036] when transferring from MP to CP, or CP to MP. The value of POT is following:

1 - PW ownership transfer from management plane to control plane

2 - PW ownership transfer from control plane to management plane

Reserved(30 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

5.2. Procedures

5.2.1. PW Ownership Transfer from MP to CP

Before transferring from MP to CP, there MUST be a T-LDP session between two T-PE for SS-PW, or T-PE and S-PE for MS-PW. During the LDP initialization stage, the LDP speaker MUST announce it’s PW transfer capability according to [RFC5561] by sending the peer a Capability message carrying the PW transfer capability TLV.

To conduct the MP2CP PW transfer, operator sends the MP2CP PW transfer command to the source and destination T-PEs which will inform MP and CP to initiate the MP2CP PW transfer process. When CP gets all the information of the PW to be transferred , the CP of source and destination nodes will build the LDP mapping message based on the procedures described in [RFC 4447], and send the mapping message to its peer T-PE or S-PE.

The differences between the normal and the MP2CP PW transfer Label Mapping message are:

1. PW-OH TLV with POT value equals 1 will be encoded into the "Optional Parameters" of the Mapping message for both SS-PW and MS-PW MP2CP transfer.

2. For MS-PW, the PW path will be encoded into S-PE TLVs and Sub-TLVs with local S-PE address according to [SEG-PW].

When the Label Mapping message is build up, it will be send to source/destination T-PE for SS-PW and to S-PE for MS-PW.

For SS-PW, when the source/destination T-PE receives the MP2CP PW transfer Label Mapping message, and also send MP2CP PW transfer Label Mapping message to its peer, it will transfer the PW control from MP to CP.

For MS-PW, when the S-PE receives the MP2CP PW transfer Label Mapping message, it will decode the next hop S-PE from local IP address Sub-TLVs in S-PE TLVs then forward this Label Mapping message to the next hop S-PE. Only when S-PE receive the MP2CP PW transfer label mapping message from the reverse direction of PW, it will transfer the PW control from MP to CP. When the source/destination T-PE receives the MP2CP PW transfer Label Mapping message, it will deal with it in the same way as SS-PW described above.

5.2.1.1. MP2CP PW Transfer Failure

If T-PEs or S-PE fail to PW transfer capability negotiation, the procedures in [RFC5561] SHOULD be performed.

Since T-LDP runs over TCP, and there is only one hop between T-PEs in SS-PW, if the T-LDP sesseion is created successfully, the PW transfer Label Mapping can be sent and received reliably.

For MS-PW, if one of the PW segment fails to transfer from MP to CP, a Notification message SHOULD be sent to source/destionation T-PE to report the failure. And the PW segments successfully transferred SHOULD be remained.

5.2.2. PW Ownership Transfer from CP to MP

Since multiple PWs can share a single T-LDP session, when a PW transferred from CP to MP, the LDP session may be retained for other PWs. So when a PW transfers from CP to MP, a Notification message carring the corresponding PW FEC and PW-OH TLV with the POT value equals 2 SHOULD be send out. All the other S-PEs along the PW received this Notification message, SHOULD send the notification message to next hop S-PE. Only when S-PE receives notification message from reverse direction of PW, it will transfer the PW control from CP to MP and remain the corresponding LDP session. When there is no PW, the session MAY be still remained for the future use. Thus, whether to delete the LDP session depends on the provider's policy. If the provider want to delete the LDP session in which there is no PW, the procedures in [RFC5036] can be conducted.

5.2.2.1. CP2MP PW Transfer Failure

Since the PW transfer capability is negotiated before T-LDP session set up, and the T-LDP runs over TCP, CP2MP PW transfer can be performed reliably.

For MS-PW, if one PW segment fails to transfer from CP to MP, a Notification message SHOULD be sent to source/destionation T-PE to report the failure.

6. Security Considerations

[RFC5036] and [RFC4447] describe the security considerations that apply to the T-LDP specification. The same security framework and considerations apply to the capability mechanism described in this document.

7. IANA considerations

TBD.

8. Acknowledgements

The authors would like to thank Weilian Jiang, and Kan Hu for their useful comments.

9. References

9.1. Normative References

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

[RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.

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

[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.

[RFC5493] Caviglia, D., Bramanti, D., Li, D., and D. McDysan, "Requirements for the Conversion between Permanent Connections and Switched Connections in a Generalized Multiprotocol Label Switching (GMPLS) Network", RFC 5493, April 2009.

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

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

9.2. Informative References

[DYNAMIC-MS-PW]
Luca Martini, Matthew Bocci, and Florin Balus, "Dynamic Placement of Multi Segment Pseudo Wires", draft-ietf-pwe3-dynamic-ms-pw-10.txt .

[G.8081] International Telecommunications Union, "Terms and definitions for Automatically Switched Optical Networks (ASON)", Recommendation G.8081/Y.1353, June 2004 .

[MPLS-TP-CP-FWK]
Loa Andersson, Lou Berger, and Luyuan Fang, "MPLS-TP Control Plane Framework", draft-ietf-ccamp-mpls-tp-cp-framework-02.txt .

[MPLS-TP-FWK]
M. Bocci and S. Bryant etc., "A Framework for MPLS in Transport Networks", draft-ietf-mpls-tp-framework-11.txt .

[SEG-PW] Luca Martini and Chris Metz, "Segmented Pseudowire", draft-ietf-pwe3-segmented-pw-13.txt .

Authors' Addresses

Yuanlin Bao
ZTE Corporation
5F, R&D Building 3, ZTE Industrial Park, XiLi LiuXian Road,
Nanshan District, Shenzhen 518055
P.R.China

Phone: +86 755 26773731
Email: bao.yuanlin@zte.com.cn
URI: http://www.zte.com.cn/

Lizhong Jin
ZTE Corporation
889, Bibo Road, Pudong District
Shanghai 201203
P.R.China

Email: lizhong.jin@zte.com.cn
URI: http://www.zte.com.cn/

Ruiquan Jing
China Telecom

Email: jingrq@ctbri.com.cn

Xiaoli Huo
China Telecom

Email: huoxl@ctbri.com.cn

Network Working Group
Internet Draft
Intended status: Informational

Expires: April 2011

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

Cisco Systems, Inc.

October 15, 2010

Stitching Procedures for Static PW in MPLS-TP Environment
draft-boutros-pwe3-mpls-tp-ms-pw-00.txt

Status of this Memo

This Informational Internet-Draft is aimed at achieving IETF Consensus before publication as an RFC and will be subject to an IETF Last Call.

[RFC Editor, please remove this note before publication as an RFC and insert the correct Streams Boilerplate to indicate that the published RFC has IETF Consensus.]

Abstract

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

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

Table of Contents

1. Introduction

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

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

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

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

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

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

2. Terminology

LDP: Label Distribution Protocol.

MEP: Maintenance End Point.

MIP: Maintenance Intermediate Point.

MPLS: Multi Protocol Label Switching.

MPLS-TP: MPLS Transport Profile.

MS-PW: Multi-Segment PseudoWire.

LB: Loopback.

LI: Lock Instruct.

LSP: Label Switched Path.

OAM: MPLS Operations, Administration and Maintenance.

PE: Provide Edge Node.

PW: PseudoWire.

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

SS-PW: Single-Segment PseudoWire.

TLV: Type, Length, and Value.

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

3. Operation

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

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

```
+-------+    (PW1)     +-------+     (PW2)     +-------+
|       |------------->|       |-------------->|       |
| T-PE1 |              | S-PE  |               | T-PE2 |
|       |<-------------|       |<--------------|       |
+-------+              +-------+               +-------+
```

Figure 1. Reference Model for LI/LB Mechanism

3.1. Lock Operation

3.1.1. Locking MPLS-TP LSP

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

3.1.1.1. LI originated at T-PE

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

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

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

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

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

3.1.1.2. LI originated at S-PE

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

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

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

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

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

3.1.2. Locking PW

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

3.1.2.1. LI originated at T-PE

In our example, let’s assume that, T-PE1 sends an LI request message to lock PW1. S-PE is the intended recipient (based on the TTL value of the PW label).If S-PE is able to lock PW1, it sends a PW status message with the status code 0x00000018 (Local PSN-facing PW Receive/Transmit Fault) over PW2 to T-PE2, and locks PW1. S-PE then sends a positive LI reply to T-PE1. Upon receiving the positive LI reply, T-PE locks PW1. If S-PE is unable to lock PW1, it sends a negative LI reply to T-PE1. PW status code is sent over PW OAM message or LDP message depending on whether the segment PW2 is static or dynamic. When T-PE2 receives the PW status codes, it processes them as described in [3] or [4] depending on whether PW2 is dynamic or static.

3.2. Loopback Operation

3.2.1. Loopback at MPLS-TP LSP Level

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

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

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

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

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

3.2.2. Loopback at PW Level

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

In our example, let’s assume that PW1 is to be operated in a loopback mode such that S-PE loops all incoming packets over PW1 back to T-PE1. To setup this mode of operation, T-PE1 sends an LB request message to S-PE. TTL value of the PW label is chosen so as to expire on the intended recipient (in our example TTL value should be 1 so that LB request can be processed at S-PE). If S-PE can successfully setup PW1 in loopback mode, it sends a positive LB response to T-PE1.

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

3.3. Switching Point PE TLV

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

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

3.4. LSP-Ping/Trace

TBD

4. Security Considerations

This document does not introduce any additional security constraints.

5. IANA Considerations

Not applicable.

6. References

6.1. Normative References

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

6.2. Informative References

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

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

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

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

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

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

Author’s Addresses

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

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

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

Full Copyright Statement

Intellectual Property Statement

Acknowledgment

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

Network working group W. Cao
Internet Draft M. Chen
Category: Standards Track Huawei Technologies Co.,Ltd
Created: October 25, 2010 A. Takacs
Expires: April 2011 Ericsson

LDP extensions for Explicit Pseudowire to transport LSP mapping

draft-cao-pwe3-mpls-tp-pw-over-bidir-lsp-01.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 some extensions to LDP that allow both ends of a PW (or segment of a MS-PW) to select and bind to the same bidirectional LSP or use unidirectional LSPs with congruent paths.

Status of this Memo

Conventions used in this document

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

Table of Contents

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 some transport services may require bidirectional traffic to follow congruent paths. When bidirectional LSPs 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. When there are multiple unidirectional/bidirectional LSPs that may be used to provide different levels of Quality of Service (QOS) or protection between the PEs a selection must be made and some form of control is required to ensure that the correct LSPs are used.

```
           +----+   +--+ LSP1 +--+   +----+
+-----+    | PE1|===|P1|======|P2|===| PE2|    +-----+
|     |----|    |   +--+      +--+   |    |----|     |
| CE1 |    |............PW................|    | CE2 |
|     |----|    |      +--+          |    |----|     |
+-----+    |    |======|P3|==========|    |    +-----+
           +----+      +--+ LSP2     +----+
```

Figure 1 SS-PW scenario

Figure 1 shows an example of inconsistent binding in a Single-Segment PW (SS-PW) scenario. There are two bidirectional LSPs (LSP1 and LSP2, along diverse paths) and a bidirectional PW service between PE1 and PE2. With the current 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, and PE2 may select LSP2 (PE1-P3-PE2) as the PSN tunnel for the PE2->PE1 direction of the PW, so the bidirectional PW service is bound to two separate bidirectional LSPs. If the service requirement is that the two directions of the PW service are routed in the same way through the network, this outcome will be unacceptable. The problem also exists in Multi-Segment PW (MS-PW) scenarios.

One possible way to resolve this issue is to bind the PSN tunnel manually at each PE, but this is prone to configuration errors and it is difficult to maintain a large number of PWs in such a manner. To allow for minimal manual intervention and configuration, this draft discusses an automatic solution by extending FEC 128/129 PW based on [RFC4447].

2. PW to LSP Binding TLV

In this document two new OPTIONAL TLVs are defined: the IPv4/IPv6 PW to LSP Binding TLVs. They are used to communicate the selected LSPs between the two PEs of a PW or segment of MS-PW.

When using LDP to signal the PW, the identifiers of the LSP are carried in the Label Mapping message utilizing the new TLVs defined in this document.

The format of the PW to LSP Binding LSP TLVs is as follows, the value fields are derived from the definition of [I-D.ietf-mpls-tp-identifiers].

(Editor notes: In I-D.ietf-mpls-tp-identifiers, an LSP is identified by the combination of Src-Global_ID, Src-Node_ID, Src-Tunnel_Num, Dst-Global_ID, Dst-Node_ID, Dst-Tunnel_Num, LSP_Num, this is fine for unidirectional and co-routed bidirectional LSP, but it is not enough for associated bidirectional LSP that is combined with two reverse unidirectional LSPs and hence two LSP_Nums are required.)

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0| IPv4 PW to LSP binding TLV|           TLV Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Source Global ID                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Source Node ID                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Source Tunnel Number      |      Source LSP Number        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Destination Global ID                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Destination Node ID                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Destination Tunnel Number   |    Destination LSP Number     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           Figure2 IPv4 PW to LSP Binding TLV format
```

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0| IPv6 PW to LSP Binding TLV|         TLV Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Source Global ID                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Source Node ID                         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Source Tunnel Number      |      Source LSP Number        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Destination Global ID                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     Destination Node ID                       ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Destination Tunnel Number   |    Destination LSP Number     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           Figure3 IPv6 PW to LSP Binding TLV format
```

As defined in [RFC3209] and [RFC3473], an RSVP-TE LSP is identified by the combination of LSP ID, Tunnel ID, Tunnel Extended ID, Tunnel end point address, Tunnel sender address, and a mapping between these fields to the fields of IPv4/v6 PW to LSP Binding TLV is needed. The mapping defined in Section 5.3 of [I-D.ietf-mpls-tp-identifiers] applies here.

In addition, for co-routed bidirectional LSP, since the Source and Destination Tunnel/LSP ID is the same, Destination Tunnel Number and Destination LSP Number MUST be set to the same as the Source Tunnel Number and Source LSP Number, respectively.

For associated bidirectional LSP, Destination Tunnel Number and Destination LSP Number MUST be set to the Tunnel ID and LSP ID of the reverse direction component LSP of the associated bidirectional LSP, respectively.

For unidirectional LSPs, when the reverse direction tunnel LSP is determined in advance (e.g., in an active/passive mode, the active end may explicitly specify the reverse tunnel LSP for a PW), Destination Tunnel Number and Destination LSP Number SHOULD be set to the Tunnel ID and LSP ID of the reverse LSP, respectively. If the reverse direction tunnel LSP can not be determined in advance, Destination Tunnel Number and Destination LSP Number MUST be set to zero.

(Editor notes: In I-D.ietf-mpls-tp-identifiers, the Source/Destination Node ID is defined as a 32-bit ID, but for a MPLS/GMPLS TE based LSP, the Extended Tunnel ID, Tunnel end point address, and Tunnel sender address may be IPv6 addresses, so the current Source/Destination Node ID does not cover this and can not map to IPv6 based Tunnel Extended ID, Tunnel end point address, and Tunnel sender address.)

3. LDP Extensions

Before sending a Label Mapping message to set up a PW or PW Segment, a PE has to select candidate LSPs to act as PSN tunnels. The selected LSPs are carried by the PW to LSP binding TLV and sent with the Label Mapping message to the target/switching PE. Therefore, there may be some collisions of tunnel LSP selection when both PEs assume the active role and independently signal the PW or PW Segment. In order to reduce and resolve the collision of tunnel selection, two types of PEs are identified here:

a) Active PE: the PE which initiates the selection of the tunnel LSPs and informs the remote PE;

b) Passive PE: the PE which obeys the active PE’s suggestion.

The role of a PE is based on the role that it takes in the signaling of a specific PW. The active/passive role election is defined in the Section 7.2.1 of [SEG-PW] and applies here, this document does not define any new role election procedures. There exist two situations:

Active/Active - Both PEs of a PW or PW Segment assume active roles (e.g., SS-PW, LDP using FEC 128 MS-PW).

Active/Passive - One PE is Active and the other is passive (e.g., LDP using FEC 129 MS-PW).

3.1.1. Active/Active Signaling Procedures

In a bidirectional LSP scenario, both PEs (say PE1 and PE2) send a Label Mapping message carrying their own selected bidirectional LSP to each other. If the bidirectional LSP in the received message from other PE is as same as it was in the Label Mapping message sent by itself, then the PW signaling has converged on an mutually agreed tunnel LSP and selection is completed. Otherwise, when the bidirectional LSP selected by one PE (say PE1) differs from the bidirectional LSP selected by the other PE (say PE2), PE1 and PE2 have to make a choice between two tunnel LSPs. In this case PE1 and PE2 can compare the Node ID, and the LSP selected by the node with numerically higher ID will be determined to carry the PW.

In case of unidirectional LSPs, each PE may select a unidirectional tunnel LSP that is used for its own forward direction of the PW and send it with the Label Mapping message to the other PE. It is possible that the two LSPs are not congruent. The mechanisms to determine which LSPs are congruent are out of scope, but it is assumed that each PE is able to look at the paths of LSPs (from information supplied to or by the control plane, or from information supplied by the management plane). In addition, each PE may explicitly specify both the forward and reverse direction tunnel LSPs of the PW and send them with the Label Mapping message to each other. If the two PEs of the PW have the same tunnel selection (e.g., for a specific PW, the forward direction tunnel LSP selected by one PE is the same as the reverse direction tunnel LSP selected by the other PE, and vice versa), then the PW signaling is completed and has converged on an mutually agreed tunnel LSPs. Otherwise, when the tunnel LSPs selected by one PE differ from the tunnel LSPs selected by the other PE, the LSPs selected by the node with numerically higher Node ID will be determined as the tunnel.

In case of one PE selects a pair of unidirectional LSPs and the other PE selects a bidirectional LSP, the LSPs selected by the node with numerically higher Node ID will be determined as the tunnel.

3.1.2. Active/Passive Signaling Procedures

3.1.2.1. Active PE Signaling Procedure

Before sending the Label Mapping message, the active PE, say PE1, MUST select the tunnel LSPs for the PW or Segment PW. Then PE1 generates a PW to LSP Binding TLV that identifies the selected LSP and sends the Label Mapping message containing it to the passive PE, in this case PE2.

In case of bidirectional LSPs, if PE1 receives a Label Mapping message in which the bidirectional LSP is the same as the bidirectional LSP it selected then both directions of the PW or Segment PW are setup.

In case of unidirectional LSPs, if PE1 specifies both the forward and reverse direction tunnel LSPs in a previous Label Mapping message sent by itself, when PE1 receives a Label Mapping message in which the reverse tunnel LSP is the same as the forward tunnel LSP and the forward tunnel LSP is the same as the reverse tunnel LSP it selected, then both directions of the PW or segment PW are setup. According to the passive PE procedures described in Section 3.1.2.2, the identified LSPs SHOULD match. If they do not, the active PE MUST assume that the peer PE is also in active role, and MUST apply the procedures described in Section 3.1.1.

3.1.2.2. Passive PE Signaling Procedure

When a Label Mapping message carrying a PW to LSP Binding TLV is received by the passive PE (say PE2) it may decide, based on local policy and/or success or failure in matching the LSP to accept or reject it.

If the suggested tunnel LSPs cannot be matched successfully or if local policy prohibits its acceptance, a Label Release message MUST be sent, with a "No matched tunnel LSPs" code, and the processing of the Label Mapping message is complete.

If the tunnel LSPs proposed by PE1 are accepted by PE2 then PE2 attempts setup of the PW in the opposite (PE2->PE1) direction, it sends a Label Mapping message to PE1, with a PW to LSP Binding TLV that identifies the tunnel LSPs, proposed by PE1, that it has accepted for this PW. That is, for bidirectional LSPs, the PW to LSP Binding TLV SHOULD identify the same bidirectional LSP proposed by PE1. In case of unidirectional LSPs, if the received PW to LSP Binding TLV including both forward and reverse direction tunnel LSPs, the Source Tunnel Number and LSP Number of the PW to LSP Binding LSP SHOULD be exchanged for each other. Accordingly, the Source/Destination Node ID/Global ID of the PW to LSP Binding TLV SHOULD be exchanged as well.

4. Security Considerations

The ability to control which LSPs are used to carry a PW is a potential security risk both for denial of service and for interception of traffic. It is RECOMMENDED that PEs do not accept the use of LSPs identified in the LSP 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 [RFC5306] and [RFC5920].

5. IANA Considerations

5.1. LDP TLV Types

This document defines two new TLVs [Section 2 of this document] for inclusion in LDP Label Mapping message. IANA is required to assigned TLV type values to the new defined TLVs from LDP "TLV Type Name Space" registry.

IPv4 PW to LSP Binding TLV - 0x0971 (to be confirmed by IANA)

IPv6 PW to LSP Binding TLV - 0x0972 (to be confirmed by IANA)

5.2. 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.

"No matched tunnel LSPs" - 0x0000003B (to be confirmed by IANA)

6. Acknowledgments

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 has benefited from discussions with Nabil Bitar, Paul Doolan, Frederic Journay and Andy Malis.

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.

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

[I-D.ietf-mpls-tp-identifiers] Bocci, M. and G. Swallow, "MPLS-TP Identifiers", "draft-ietf-mpls-tp-identifiers-01", work in progress.

7.2. Informative References

[SEG-PW] Luca Martini, et al., "Segmented Pseudowire", "draft-ietf-pwe3-segmented-pw-15.txt", work in progress.

[TP-CP-FWK] Loa Andersson, Lou Berger, Luyuan Fang, Nabil Bitar, "MPLS-TP Control Plane Framework", "draft-ietf-ccamp-mpls-tp-cp-framework", work in progess.

[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.

[RFC3473] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling", RFC 3473, January 2003.

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

Authors' Addresses

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd.
No. 3 Xinxi Road
Shangdi Information Industry Base
Hai-Dian District, Beijing 100085
China

EMail: mach@huawei.com

Wei Cao
Huawei Technologies Co., Ltd.
No. 3 Xinxi Road
Shangdi Information Industry Base
Hai-Dian District, Beijing 100085
China

EMail: caoweigne@huawei.com

Attila Takacs
Ericsson
Laborc u. 1.
Budapest, 1037
Hungary

EMail: attila.takacs@ericsson.com

Network Working Group N. Del Regno, Ed.
Internet-Draft Verizon Communications
Intended status: Standards Track T. Nadeau
Expires: April 18, 2011 Huawei
V. Manral
IP Infusion
D. Ward
Juniper Networks
October 15, 2010

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

Abstract

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

Requirements Language

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

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

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

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

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

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

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

Unfortunately, some early implementations of PWE3 standard (and/or prestandard) encapsulations are limited in their support for Control Word for the above encapsulations due to A) hardware deficiencies, B) software deficiencies or C) a combination of the two. In other cases, deployed implementations support control word, but the service provider has had no impetus to suffer the minor loss of overhead efficiency. However, this document asserts based on operational feedback of the PWE3 protocols in actual deployments, that the benefits of requiring a mandatory control word in the PWE3 standards outweigh the minor efficiencies gained when not using it.

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

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

2. Mandatory Control Word

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

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

3. Backward Compatibility

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

PREFERRED and the subsequent c=1 is negotiated.

4. IANA Considerations

This document makes no request of IANA.

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

5. Security Considerations

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

6. Acknowledgements

7. Normative References

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

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

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

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

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

Authors' Addresses

Nick Del Regno (editor)
Verizon Communications

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

Thomas Nadeau
Huawei

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

Vishwas Manral
IP Infusion

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

David Ward
Juniper Networks

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

Network Working Group N. Del Regno, Ed.
Internet-Draft Verizon
Intended status: BCP V. Manral, Ed.
Expires: April 18, 2011 IPInfusion Inc.
R. Kunze
M. Paul
Deutsche Telekom
T. Nadeau
Huawei
October 15, 2010

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

Abstract

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

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

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

Requirements Language

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

Table of Contents

1. Introduction

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

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

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

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

2. Comparison of Alternative Control Channel Types

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

2.1. Control Channel Type 1: Control Word

As noted in [RFC5085], an in-band control channel is ideal, since this ensures that the connectivity verification messages follow the same path as the PWE3 traffic. VCCV Control Channel Type 1, also known as "PWE3 Control Word with 0001b as first nibble," is the only "in-band" control channel specified. It uses the control word as opposed to using the label to indicate the presence of the Connectivity Verification message (CV). This ensures that the control channel follows the forwarding path of the associated traffic in all cases, including in the case of ECMP hashing.

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

2.2. Control Channel Type 2: MPLS Router Alert Label

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

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

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

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

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

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

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

Implementations MUST support the use of this option.

3. Mandatory Control Channels

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

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

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

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

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

4. Mandatory CV Types

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

5. IANA Considerations

This document makes no request of IANA.

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

6. Security Considerations

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

7. Acknowledgements

8. References

8.1. Normative References

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

8.2. Informative References

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

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

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

Authors' Addresses

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

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

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

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

Ruediger Kunze
Deutsche Telekom

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

Manuel Paul
Deutsche Telekom

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

Thomas Nadeau
Huawei

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

Network Working Group
Internet Draft
Expiration Date: April 2011
Intended status: Standards Track

Luca Martini (Ed.)
Cisco Systems Inc.

Matthew Bocci (Ed.)
Florin Balus (Ed.)
Alcatel-Lucent

October 13, 2010

Dynamic Placement of Multi Segment Pseudo Wires

draft-ietf-pwe3-dynamic-ms-pw-13.txt

Status of this Memo

Abstract

There is a requirement for service providers to be able to extend the reach of pseudo wires (PW) across multiple Packet Switched Network domains. A Multi-Segment PW is defined as a set of two or more contiguous PW segments that behave and function as a single point-to-point PW. This document describes extensions to the PW control protocol to dynamically place the segments of the multi segment pseudo wire among a set of Provider Edge (PE) routers.

Table of Contents

1. Major Co-authors

The editors gratefully acknowledge the following additional co-authors: Mustapha Aissaoui, Nabil Bitar, Mike Loomis, David McDysan, Chris Metz, Andy Malis, Jason Rusmeisel, Himanshu Shah, Jeff Sugimoto.

2. Acknowledgements

The editors also gratefully acknowledge the input of the following people: Mike Ducket, Paul Doolan, Prayson Pate, Ping Pan, Vasile Radoaca, Yeongil Seo, Yetik Serbest, Yuichiro Wada.

3. Introduction

3.1. Scope

[MS-REQ] describes the service provider requirements for extending the reach of pseudo-wires across multiple PSN domains. This is achieved using a Multi-segment Pseudo-Wire (MS-PW). A MS-PW is defined as a set of two or more contiguous PW segments that behave and function as a single point-to-point PW. This architecture is described in [MS-ARCH].

The procedures for establishing PWs that extend across a single PWE3 domain are described in [RFC4447], while procedures for setting up PWs across multiple domains, or control planes are described in [PW-SEG].

The purpose of this draft is to specify extensions to the PWE3 control protocol [RFC4447], and [PW-SEG] procedures, to enable multi-segment PWs to be automatically placed. The proposed procedures follow the guidelines defined in [RFC5036] and enable the reuse of existing TLVs, and procedures defined for SS-PWs in [RFC4447].

3.2. 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.

3.3. Terminology

[MS-ARCH] provides terminology for multi-segment pseudo wires.

This document defines the following additional terms:

- Source Terminating PE (ST-PE). A Terminating PE (T-PE), which assumes the active signaling role and initiates the signaling for multi-segment PW.
- Target Terminating PE (TT-PE). A Terminating PE (T-PE) that assumes the passive signaling role. It waits and responds to the multi-segment PW signaling message in the reverse direction.
- Forward Direction: ST-PE to TT-PE.
- Reverse Direction: TT-PE to ST-PE
- Forwarding Direction: Direction of control plane, signaling flow
- Pseudo wire Routing (PW routing). The dynamic placement of SS-PWs that compose an MS-PW, as well as the automatic selection of S-PEs.

3.4. Architecture Overview

The following figure describes the reference models which are derived from [MS-ARCH] to support PW emulated services across multi-segment PWs.

```
       Native   |<-------------Pseudo Wire----------->|  Native
       Service  |                                     |  Service
        (AC)    |     |<-PSN1-->|     |<-PSN2-->|     |   (AC)
          |     V     V         V     V         V     V     |
          |     +-----+         +-----+         +-----+
    +----+ |     |T-PE1|=========|S-PE1|=========|T-PE2|     |    +----+
    |    |-------|.....PW.Seg't1........PW Seg't3......|----------|    |
    | CE1| |     |     |         |     |         |     |     |    |CE2 |
    |    |-------|.....PW.Seg't2.......|PW Seg't4......|----------|    |
    +----+ |     |     |=========|     |=========|     |     |    +----+
         ^       +-----+         +-----+         +-----+          ^
         |   Provider Edge 1        ^       Provider Edge 2       |
         |                          |                             |
         |                          |                             |
         |               PW switching point                       |
         |                                                        |
         |<---------------- Emulated Service -------------------->|
```

Figure 1: PW switching Reference Model

Figure 1 shows the architecture for a simple multi-segment case. T-PE1 and T-PE2 provide PWE3 to CE1 and CE2. These PEs reside in different PSNs. A PSN tunnel extends from T-PE1 to S-PE1 across PSN1, and a second PSN tunnel extends from S-PE1 to T-PE2 across PSN2. PWs are used to connect the attachment circuits (ACs) attached to T-PE1 to the corresponding AC attached to T-PE2. A PW on the tunnel across PSN1 is connected to a PW in the tunnel across PSN2 at S-PE1 to complete the multi-segment PW (MS-PW) between T-PE1 and T-PE2. S-PE1 is therefore the PW switching point and will be referred to as the switching provider edge (S-PE). PW Segment 1 and PW Segment 3 are segments of the same MS-PW while PW Segment 2 and PW Segment 4 are segments of another MS-PW. PW segments of the same MS-PW (e.g., PW segment 1 and PW segment 3) MUST be of the same PW type, and PSN tunnels (e.g., PSN1 and PSN2) can be the same or different technology. An S-PE switches an MS-PW from one segment to another based on the PW identifiers. (PWid , or AII) How the Pw PDUs are switched at the S-PE depends on the PSN tunnel technology: in case of an MPLS PSN to another MPLS PSN PW switching the operation is a standard MPLS label switch operation.

Note that although Figure 1 only shows a single S-PE, a PW may transit more one S-PE along its path. For instance, in the multi-provider case, there can be an S-PE at the border of one provider domain and another S-PE at the border of the other provider domain.

4. Applicability

In this document we describe the case where the PSNs carrying the SS-PW are only MPLS PSNs using the generalized FEC 129. Interactions with an IP PSN using L2TPv3 as described in [PW-SEG] section 7.4 are left for further study.

4.1. Requirements Addressed

Specifically the following requirements are addressed [MS-REQ]:

- Dynamic End-to-end Signaling
- Scalability and Inter-domain Signaling and Routing
- Minimal number of provisioning touches (provisioning only at the T-PEs)
- Same set of T-PEs/S-PEs for both directions of a MS-PWs
- QoS Signaling, Call Admission Control
- Resiliency
- End-to-end negotiation of OAM Capability

4.2. Changes to Existing PW Signaling

The procedures described in this document make use of existing LDP TLVs and related PW signaling procedures described in [RFC4447] and [PW-SEG]. Only an optional Bandwidth TLV is added to address the QoS Signaling requirements (see "MS-PW Next Hop Bandwidth Signaling" section for details).

5. PW layer 2 addressing

Single segment pseudo wires on an MPLS PSN use Attachment circuit identifiers for a PW using FEC 129. In the case of an automatically placed MS-PW, there is a requirement to have individual global addresses assigned to PW attachment circuits, for reachability , and manageability of the PW. Referencing figure 1 above, individual globally unique addresses MUST be allocated to all the ACs , and S-PEs composing an MS-PW.

5.1. Attachment Circuit Addressing

The attachment circuit addressing is derived from [RFC5003] AII type 2 shown here:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  AII Type=02  |    Length     |        Global ID              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Global ID (contd.)      |        Prefix                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Prefix (contd.)         |        AC ID                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      AC ID                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Implementations of the following procedure MUST interpret the AII type to determine the meaning of the address format of the AII, irrespective of the number of segments in the MS-PW.

A unique combination Global ID, Prefix, and AC ID parts of the AII type 2 will be assigned to each AC. In general the same global ID and prefix will be assigned for all ACs belonging to the same T-PE, however this is not a strict requirement. A particular T-PE might have more than one prefix assigned to it, and likewise a fully qualified AII with the same Global ID/Prefix but different AC IDs might belong to different T-PEs.

For the purpose of MS-PW the AII MUST be globally unique across all interconnected PW domains.

5.2. S-PE addressing

The T-PE may elect to select a known specific path along a set of S-PEs for a specific PW. This requires that each S-PE be uniquely addressable in terms of pseudo wires. For this purpose at least one AI address of the format similar to AII type 2 [RFC5003] composed of the Global ID, and Prefix part only MUST be assigned to each S-PE.

6. Dynamic placement of MS-PWs

[PW-SEG] describes a procedure for connecting multiple pseudo wires together. This procedure requires each S-PE to be manually configured with the information required to terminate and initiate the SS-PW part of the MS-PW. The procedures in the following sections describe an method to extend [PW-SEG] by allowing the automatic selection of pre-defined S-PEs, and automatically setting up a MS-PW between two T-PEs.

6.1. Pseudo wire routing procedures

The AII type 2 described above contains a Global ID, Prefix, and AC ID. The TAII is used by S-PEs to determine the next SS-PW destination for LDP signaling.

Once an S-PE receives a MS-PW label mapping message containing a TAII with an AII that is not locally present, the S-PE performs a lookup in a local Layer 2 AII PW routing table. If this lookup results in an IP address of the next PE that advertised reachability information for the AII in question, then the S-PE will initiate the necessary LDP messaging procedure for setting up the next PW segment. If the AII PW routing table lookup does not result in a IP address of the next PE, the destination AII has become unreachable, and the PW MUST fail to setup. In this case the next PW segment is considered unprovisioned, and a label release MUST be returned to the T-PE with a status message of "AII Unreachable".

If the TAI of a MS-PW label mapping message, received by a PE, contains the prefix of a locally provisioned prefix on that PE, but an AC ID that is not provisioned, then the LDP liberal label retention procedures apply, and the label mapping message is retained.

To allow for dynamic end-to-end signaling of MS-PWs, information must be present in S-PEs to support the determination of the next PW signaling hop. Such information can be provisioned (static route equivalent) on each S-PE system or disseminated via regular routing protocols (e.g. BGP).

6.1.1. AII PW routing table Lookup aggregation rules

All PEs capable of dynamic multi segment pseudowire path selection, must build a PW routing table to be used for PW next hop selection.

The PW addressing scheme (AII type 2 in [RFC5003]) consists of a

Global Id, a 32 bit prefix and a 32 bit Attachment Circuit ID.

An aggregation scheme similar with the one used for classless IPv4 addresses can be employed. An (8 bits) length mask is specified as a number ranging from 0 to 96 that indicates which Most Significant Bits (MSB) are relevant in the address field when performing the PW address matching algorithm.

```
 0        31 32    63 64    95 (bits)
+-----------+--------+--------+
| Global ID | Prefix | AC ID  |
+-----------+--------+--------+
```

During the signaling phase, the content of the (fully qualified) TAII type 2 field from the FEC129 TLV is compared against routes from the PW Routing table. Similar with the IPv4 case, the route with the longest match is selected, determining the next signaling hop and implicitly the next PW Segment to be signaled.

6.1.2. PW Static Route

For the purpose of determining the next signaling hop for a segment of the pseudo wire, the PEs MAY be provisioned with fixed route entries in the PW next hop routing table. The static PW entries will follow all the addressing rules and aggregation rules described in the previous sections. The most common use of PW static provisioned routes is this example of the "default" route entry as follows:

Global ID = 0 Prefix = 0 AC ID = 0 , Prefix Length = 0 Next Signaling Hop = S-PE1

6.1.3. Dynamic advertisement with BGP

Any suitable routing protocol capable of carrying external routing information may be used to propagate MS-PW path information among S-PE, and T-PE. However, T-PE, and S-PEs, MAY choose to use Boundary Gateway Protocol (BGP) [RFC4760] to propagate PW address information throughout the PSN.

Contrary to other l2vpn signaling methods that use BGP [L2-SIGNALING], in the case of the dynamically placed MS-PW if the source T-PE knows a priori (by provisioning) the address of the terminating T-PE. Hence there is no need to advertise a "fully qualified" 96 bit address on a per PW Attachment Circuit basis. Only the T-PE Global ID, Prefix, and prefix length needs to be advertised as part of well known BGP procedures - see [RFC4760].

As PW Endpoints are provisioned in the T-PEs. The ST-PE will use this information to obtain the first S-PE hop (i.e., first BGP next hop) to where the first PW segment will be established. Any subsequent S-PEs will use the same information (i.e. the next BGP next-hop(s)) to obtain the next-signaling-hop(s) on the path to the TT-PE.

The PW dynamic path NLRI is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [RFC4760]. The [AFI, SAFI] value pair used to identify this NLRI is (AFI=25, SAFI=6 (pending IANA allocation)).

The Next Hop field of MP_REACH_NLRI attribute shall be interpreted as an IPv4 address, whenever the length of the NextHop address is 4 octets, and as a IPv6 address, whenever the length of the NextHop address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix comprising an 8-octet Route Distinguisher, the Global ID, the Prefix, and the AC-ID, and encoded as defined in section 4 of [RFC4760].

This NLRI is structured as follows:

```
 Bit
 0     7 8             71 72      103 104  135 136   167
 +------+----------------+-----------+--------+--------+
 |Length|  Route Dist    | Global ID | Prefix | AC ID  |
 +------+----------------+-----------+--------+--------+
```

The Length field is the prefix length of the Route Distinguisher + Global ID + Prefix + AC-ID in bits.

Except for the default PW route, which is encoded as a 0 length prefix, the minimum value of the length field is 96 bits. Lengths of 128 bits to 159 bits are invalid as the AC ID field cannot be aggregated. The maximum value of the Length field is 160 bits. BGP advertisements received with invalid prefix lengths MUST be rejected as having a bad packet format.

6.2. LDP Signaling

The LDP signaling procedures are described in [RFC4447] and expanded in [PW-SEG]. No new LDP Signaling components are required for setting up a dynamically placed MS-PW. However some optional signaling extensions are described below.

6.2.1. MS-PW Bandwidth Signaling

In the SS-PW case the PW QoS requirements may easily be met by selecting a MPLS PSN tunnel at the S-PE that meets the PW QoS requirements. However in the case of an automatically placed MS-PW the QoS requirements for a SS-PW not initiating on a T-PE MAY need to be indicated along with the MS-PW addressing. This is accomplished by including an OPTIONAL PW Bandwidth TLV. The PW Bandwidth TLV is specified as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|0|     PW BW  TLV  (0x096E)   |         TLV  Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Forward SENDER_TSPEC                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Reverse SENDER_TSPEC                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The complete definitions of the content of the SENDER_TSPEC objects are found in [TSPEC] section 3.1. The forward SENDER_TSPEC refers to the data path in the direction of ST-PE to TT-PE. The reverse SENDER_TSPEC refers to the data path in the direction TT-PE to ST-PE.

In the forward direction, after a next hop selection is determined, a T/S-PE SHOULD reference the forward SENDER_TSPEC object to determine an appropriate PSN tunnel towards the next signaling hop. If such a tunnel exists, the MS-PW signaling procedures are invoked with the inclusion of the PW Bandwidth TLV. When the PE searches for a PSN tunnel, any tunnel which points to a next hop equivalent to the next hop selected will be included in the search.(The LDP address TLV is used to determine the next hop equivalence)

When an S/T-PE receives a PW Bandwidth TLV, once the PW next hop is selected, the S/T-PE MUST request the appropriate resources from the PSN. The resources described in the reverse SENDER_TSPEC are allocated from the PSN toward the originator of the message or previous hop. When resources are allocated from the PSN for a specific PW, then the PSN SHOULD account for the PW usage of the resources.

In the case where PSN resources towards the previous hop are not available the following procedure MUST be followed:

- -i. The PSN MAY allocate more QoS resources, e.g. Bandwidth, to the PSN tunnel.
- -ii. The S-PE MAY attempt to setup another PSN tunnel to accommodate the new PW QoS requirements.
- -iii. If the S-PE cannot get enough resources to setup the segment in the MS-PW a label release MUST be returned to the previous hop with a status message of "Bandwidth resources unavailable"

In the latter case, the T-PE receiving the status message MUST also withdraw the corresponding PW label mapping for the opposite direction if it has already been successfully setup.

If an ST-PE receives a label mapping message the following procedure MUST be followed:

If the ST-PE has already sent a label mapping message for this PW then the ST-PE must check that this label mapping message originated from the same LDP peer to which the corresponding label mapping message for this particular PW was sent. If it is the same peer, the PW is established. If it is a different peer, then ST-PE MUST send a label release message, with a status code of "Duplicate AII" to the PE that originate the LDP label mapping message.

If the PE has not yet sent a label mapping message for this particular PW , then it MUST send the label mapping message to this same LDP peer, regardless of what the PW TAII routing lookup result is.

6.2.2. Active/Passive T-PE Election Procedure

When a MS-PW is signaled, Each T-PE might independently start signaling the MS-PW, this could result in a different path selected for each T-PE PW. To avoid this situation one of the T-PE MUST start the PW signaling (active role), while the other waits to receive the LDP label mapping before sending the respective PW LDP label mapping message. (passive role). The Active T-PE (the ST-PE) and the passive T-PE (the TT-PE) MUST be identified before signaling is initiated for a given MS-PW.

The determination of which T-PE assume the active role SHOULD be done as follows: the SAII and TAII are compared as unsigned integers, if the SAII is bigger then the T-PE assumes the active role.

The selection process to determine which T-PE assumes the active role MAY be superseded by manual provisioning.

6.2.3. Detailed Signaling Procedures

On receiving a label mapping message, the S-PE MUST inspect the FEC TLV. If the receiving node has no local AII matching the TAII for that label mapping then the S-PE will check if the FEC is already installed for the forward direction:

- If it is already installed, and the received mapping was received from the same LDP peer where the forward LDP label mapping was sent, then this label mapping represents signaling in the reverse direction for this MS-PW segment.
- Otherwise this represents signaling in the forward direction.

For the forward direction:

-i. Determine the next hop S-PE or T-PE according to the procedures above.

-ii. Check that a PSN tunnel exists to the next hop S-PE or T-PE. If no tunnel exists to the next hop S-PE or T-PE the S-PE MAY attempt to setup a PSN tunnel.

-iii. Check that a PSN tunnel exists to the previous hop. If no tunnel exists to the previous hop S-PE or T-PE the S-PE MAY attempt to setup a PSN tunnel.

-iv. If the S-PE cannot get enough PSN resources to setup the segment to the next or previous S-PE or T-PE, a label release MUST be returned to the T-PE with a status message of "Resources Unavailable".

-v. If the label mapping message contains a Bandwidth TLV, allocate the required resources on the PSN tunnels in the forward and reverse directions according to the procedures above.

-vi. Allocate a new PW label for the forward direction.

-vii. Install the FEC for the forward direction.

-viii. Send the label mapping message with the new forward label and the FEC to the next hop S-PE/T-PE.

For the reverse direction:

-i. Install the received FEC for the reverse direction.

-ii. Determine the next signaling hop by referencing the LDP sessions used to setup the LSP in the Forward direction.

-iii. Allocate a new PW label for the reverse direction.

-iv. Install the FEC for the reverse direction.
-v. Send the label mapping message with a new label and the FEC to the next hop S-PE/ST-PE.

6.2.4. Support for Explicit PW Path

The Explicit Route TLV format defined in [RFC3212] section 4.1 MAY be used to signal an explicit path for a MS-PW. An Explicit PW path may be required to provide a simple solution for 1:1 protection with diverse primary and backup path or to enable controlled signaling (strict or loose) for special PWs. Details of its usage to be provided in a future study.

7. Failure Handling Procedures

7.1. PSN Failures

Failures of the PSN tunnel MUST be handled by PSN mechanisms. If the PSN is unable to re-establish the PSN tunnel, then the S-PE SHOULD follow the procedures defined in Section 8 of [PW-SEG].

7.2. S-PE Reachability Failures

For defects in an S-PE, the procedures defined in [PW-SEG] SHOULD be followed. However in general an established MS-PW will not be affected by changes in L2 PW reachability information.

T-PEs that receive a label release message with a status of "AII Unreachable" MUST re-attempt to establish the PW immediately. However the T-PE MUST throttle its PW setup message retry attempts with an exponential backoff in situations where PW setup messages are being constantly released. It is also recommended that a T-PE detecting such a situation take action to notify an operator.

If there is a change in the L2 PW reachability information in the forward direction only, the T-PE MAY elect to tear down the MS-PW by sending a label withdraw message and re-establish the MS-PW. In the same case, an S-PE MAY do the same by sending a label withdraw message in the forward direction, and a label release message in the opposite direction along the MS-PW.

A change in L2 reachability information in the reverse direction has no effect on an MS-PW.

8. Operations and Maintenance (OAM)

The OAM procedures defined in [PW-SEG] may be used also for MS-PWs. A PW switching point TLV is used [PW-SEG] to record the switching points that the PW traverses.

In the case of a MS-PW where the PW Endpoints are identified though using a globally unique, FEC 129-based AII addresses, there is no PWID defined on a per segment basis. Each individual PW segment is identified by the address of adjacent S-PE(s) in conjunction with the SAI and TAI. In this case, the following type MUST be used in place of type 0x01 in the PW switching point TLV:

Type	Length	Description
0x06	12	L2 PW address of PW Switching Point

The above field MUST be included together with type 0x02 in the TLV once per individual PW Switching Point following the same rules and procedures as described in [PW-SEG].

9. Security Considerations

This document specifies only extensions to the protocols already defined in [RFC4447], and [PW-SEG]. Each such protocol may have its own set of security issues, but those issues are not affected by the extensions specified herein. Note that the protocols for dynamically distributing PW Layer 2 reachability information may have their own security issues, however those protocols specifications are outside the scope of this document.

10. IANA Considerations

This document uses several new LDP TLV types, IANA already maintains a registry of name "TLV TYPE NAME SPACE" defined by RFC3036. The following value is suggested for assignment:

TLV type	Description
0x096E	Bandwidth TLV

10.1. LDP Status Codes

This document uses several new LDP status codes, IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by RFC3036. The following values have been pre-allocated:

Range/Value	E	Description	Reference
0x00000037	0	Bandwidth resources unavailable	RFCxxxx
0x00000038	0	Resources Unavailable	RFCxxxx
0x00000039	0	AII Unreachable	RFCxxxx
0x0000003A	0	PW Loop Detected	RFCxxxx

10.2. BGP SAFI

IANA needs to allocate a new BGP SAFI for "Network Layer Reachability Information used for Dynamic Placement of Multi-Segment Pseudiwires" from the IANA "Subsequence Address Family Identifiers (SAFI)" registry. The following value has been pre-allocated:

Value	Description	Reference
6	Network Layer Reachability Information used for Dynamic Placement of Multi-Segment Pseudowires	[RFCxxxx]

11. Normative References

[PW-SEG] Martini et.al. "Segmented Pseudo Wire", draft-ietf-pwe3-segmented-pw-1.txt, IETF Work in Progress, September 2010

[TSPEC] Wroclawski, J. "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997

[RFC5036] Andersson, Minei, Thomas. "LDP Specification" RFC5036, October 2007

[RFC4447] "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", Martini L.,et al, RFC 4447, June 2005.

[RFC5003] "Attachment Individual Identifier (AII) Types for Aggregation", Metz, et al, RFC5003, September 2007

[RFC3212] B. Jamoussi, et al. "Constraint-Based LSP Setup using LDP", RFC3212, January 2002.

12. Informative References

[MS-REQ] Martini et al, "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3)", RFC5023, Bitar, Martini, Bocci, October 2008

[MS-ARCH] Bocci at al, "An Architecture for Multi-Segment Pseudo Wire Emulation Edge-to-Edge", RFC5659,October 2009.

[RFC4760] Bates, T., Rekhter, Y., Chandra, R. and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[L2-SIGNALING] E. Rosen, W. Luo, B. Davie, V. Radoaca, "Provisioning, Autodiscovery, and Signaling in L2VPNs", draft-ietf-l2vpn-signaling-08.txt May 3, 2006.(work in progress)

13. Author’s Addresses

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

Matthew Bocci
Alcatel-Lucent,
Voyager Place
Shoppenhangers Road
Maidenhead
Berks, UK
e-mail: matthew.bocci@alcatel-lucent.co.uk

Florin Balus
Alcatel-Lucent
701 E. Middlefield Rd.
Mountain View, CA 94043
e-mail: florin.balus@alcatel-lucent.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
e-mail: nabil.bitar@verizon.com

Himanshu Shah
Ciena Corp
35 Nagog Park,
Acton, MA 01720
e-mail: hshah@ciena.com

Mustapha Aissaoui
Alcatel-Lucent
600 March Road
Kanata
ON, Canada
e-mail: mustapha.aissaoui@alcatel-lucent.com

Jason Rusmisel
Alcatel-Lucent
600 March Road
Kanata
ON, Canada
e-mail: Jason.rusmisel@alcatel-lucent.com

Yetik Serbest
SBC Labs
9505 Arboretum Blvd.
Austin, TX 78759
e-mail: Yetik_serbest@labs.sbc.com

Andrew G. Malis
Verizon
117 West St.
Waltham, MA 02451
e-mail: andrew.g.malis@verizon.com

Chris Metz
Cisco Systems, Inc.
3700 Cisco Way
San Jose, Ca. 95134
e-mail: chmetz@cisco.com

David McDysan
Verizon
22001 Loudoun County Pkwy
Ashburn, VA, USA 20147
e-mail: dave.mcdysan@verizon.com

Jeff Sugimoto
Nortel
3500 Carling Ave.
Ottawa, Ontario, CANADA
e-mail: sugimoto@nortel.com

Mike Duckett
Bellsouth
Lindbergh Center D481
575 Morosgo Dr
Atlanta, GA 30324
e-mail: mduckett@bellsouth.net

Mike Loomis
Nortel
600, Technology Park Dr
Billerica, MA, USA
e-mail: mloomis@nortel.com

Paul Doolan
Mangrove Systems
IO Fairfield Blvd
Wallingford, CT, USA 06492
e-mail: pdoolan@mangrovesystems.com

Ping Pan
Hammerhead Systems
640 Clyde Court
Mountain View, CA, USA 94043
e-mail: ppan@hammerheadsystems.com

Prayson Pate
Overture Networks, Inc.
507 Airport Blvd, Suite 111
Morrisville, NC, USA 27560
e-mail: prayson.pate@overturenetworks.com

Vasile Radoaca
Alcatel-Lucent
Optics Divison, Westford, MA, USA
email: vasile.radoaca@alcatel-lucent.com

Yuichiro Wada
NTT Communications
3-20-2 Nishi-Shinjuku, Shinjuke-ku
Tokyo 163-1421, Japan
e-mail: yuichiro.wada@ntt.com

Yeongil Seo
Korea Telecom Corp.
463-1 Jeonmin-dong, Yusung-gu
Daejeon, Korea
e-mail: syi1@kt.co.kr

Copyright Notice

Expiration Date: April 2011

Internet Engineering Task Force Luca Martini
Internet Draft Samer Salam
Intended status: Standards Track Ali Sajassi
Expires: April 13, 2011*(L1

Matthew Bocci Satoru Matsushima
Alcatel-Lucent Softbank

Thomas D. Nadeau
Huawei Technologies

October 13, 2010

Inter-Chassis Communication Protocol for L2VPN PE Redundancy

draft-ietf-pwe3-iccp-04.txt

Status of this Memo

Abstract

This document specifies an inter-chassis communication protocol (ICCP) 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, for the purpose of synchronizing data

amongst the systems. It accommodates multi-chassis attachment circuit as well as pseudowire redundancy mechanisms.

Table of Contents

1. 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.

2. Introduction

Network availability is a critical metric for service providers as it has a direct bearing on their profitability. Outages translate not only to lost revenue but also to potential penalties mandated by contractual agreements with customers running mission-critical applications that require tight SLAs. This is true for any carrier network, and networks employing Layer 2 Virtual Private Network (L2VPN) technology are no exception. Network high-availability can be achieved by employing intra and inter-chassis redundancy mechanisms. The focus of this document is on the latter. The document defines an Inter-Chassis Communication Protocol (ICCP) that allows synchronization of state and configuration data between a set of two or more PEs forming a Redundancy Group (RG). The protocol supports multi-chassis redundancy mechanisms that can be employed on either the attachment circuit or pseudowire front.

3. ICCP Overview

3.1. Redundancy Model & Topology

The focus of this document is on PE node redundancy. It is assumed that a set of two or more PE nodes are designated by the operator to form a Redundancy Group (RG). Members of a Redundancy Group fall under a single administration (e.g. service provider) and employ a common redundancy mechanism towards the access (attachment circuits or access pseudowires) and/or towards the core (pseudowires) for any given service instance. It is possible, however, for members of an RG to make use of disparate redundancy mechanisms for disjoint services. The PE devices may be offering any type of L2VPN service, i.e. VPWS or VPLS. As a matter of fact, the use of ICCP may even be applicable for Layer 3 service redundancy, but this is considered to be outside the scope of this document.

The PEs in an RG offer multi-homed connectivity to either individual devices (e.g. CE, DSLAM, etc...) or entire networks (e.g. access network). Figure 1 below depicts the model.

```
                                    +=================+
                                    |                 |
   Mutli-homed         +----+       |  +-----+        |
   Node  ------------> | CE |-------|--| PE1 ||<------|---Pseudowire-->|
                       |    |--+   -|--|     ||<------|---Pseudowire-->|
                       +----+  |  / |  +-----+        |
                               | /  |     ||          |
                               |/   |     || ICCP     |--> Towards Core
              +-------------+  /    |     ||          |
              |             | /|    |  +-----+        |
              |    Access   |/ +----|--| PE2 ||<------|---Pseudowire-->|
              |   Network   |-------|--|     ||<------|---Pseudowire-->|
              |             |       |  +-----+        |
              |             |       |                 |
              +-------------+       |   Redundancy    |
                ^                   |     Group       |
                |                   +=================+
                |
         Multi-homed Network
```

Figure 1: Generic Multi-chassis Redundancy Model

In the topology of Figure 1, the redundancy mechanism employed towards the access node/network can be one of a multitude of technologies, e.g. it could be IEEE 802.3ad Link Aggregation Groups with Link Aggregation Control Protocol (LACP), or SONET APS. The specifics of the mechanism are out of the scope of this document. However, it is assumed that the PEs in the RG are required to communicate amongst each other in order for the access redundancy mechanism to operate correctly. As such, it is required to run an inter-chassis communication protocol among the PEs in the RG in order to synchronize configuration and/or running state data.

Furthermore, the presence of the inter-chassis communication channel allows simplification of the pseudowire redundancy mechanism. This is primarily because it allows the PEs within an RG to run some arbitration algorithm to elect which pseudowire(s) should be in active or standby mode for a given service instance. The PEs can then advertise the outcome of the arbitration to the remote-end PE(s), as opposed to having to embed a hand-shake procedure into the pseudowire redundancy status communication mechanism, and every other possible Layer 2 status communication mechanism.

3.2. ICCP Interconnect Scenarios

When referring to 'interconnect' in this section, we are concerned with the links or networks over which Inter-Chassis Communication Protocol messages are transported, and not normal data traffic between PEs. The PEs which are members of an RG may be either physically co-located or geo-redundant. Furthermore, the physical interconnect between the PEs over which ICCP is to run may comprise of either dedicated back-to-back links or a shared connection through the packet switched network (PSN); for e.g., MPLS core network. This gives rise to a matrix of four interconnect scenarios, described next.

3.2.1. Co-located Dedicated Interconnect

In this scenario, the PEs within an RG are co-located in the same physical location (POP, CO). Furthermore, dedicated links provide the interconnect for ICCP among the PEs.

```
+==================+     +------------------+
|CO                |     |                  |
|  +-----+         |     |                  |
|  | PE1 |_________|_____|                  |
|  |     |         |     |                  |
|  +-----+         |     |                  |
|     ||           |     |                  |
|     || ICCP      |     |      Core        |
|     ||           |     |     Network      |
|  +-----+         |     |                  |
|  | PE2 |_________|_____|                  |
|  |     |         |     |                  |
|  +-----+         |     |                  |
|                  |     |                  |
+==================+     +------------------+
```

Figure 2: ICCP Co-located PEs Dedicated Interconnect Scenario

Given that the PEs are connected back-to-back in this case, it is possible to rely on Layer 2 redundancy mechanisms to guarantee the robustness of the ICCP interconnect. For example, if the interconnect comprises of IEEE 802.3 Ethernet links, it is possible to provide link redundancy by means of IEEE 802.3ad Link Aggregation Groups.

3.2.2. Co-located Shared Interconnect

In this scenario, the PEs within an RG are co-located in the same physical location (POP, CO). However, unlike the previous scenario, there are no dedicated links between the PEs. The interconnect for ICCP is provided through the core network to which the PEs are connected. Figure 3 depicts this model.

```
        +=================+     +-----------------+
        |CO               |     |                 |
        |  +-----+        |     |                 |
        |  | PE1 |________|_____|                 |
        |  |     |<=================+             |
        |  +-----+   ICCP |     |  ||             |
        |                 |     |  ||             |
        |                 |     |  ||   Core      |
        |                 |     |  ||  Network    |
        |  +-----+        |     |  ||             |
        |  | PE2 |________|_____|  ||             |
        |  |     |<=================+             |
        |  +-----+        |     |                 |
        |                 |     |                 |
        +=================+     +-----------------+
```

Figure 3: ICCP Co-located PEs Shared Interconnect Scenario

Given that the PEs in the RG are connected over the packet switched network (PSN), then PSN Layer mechanisms can be leveraged to ensure the resiliency of the interconnect against connectivity failures. For example, it is possible to employ RSVP LSPs with Fast Re-Route (FRR) and/or end-to-end backup LSPs.

3.2.3. Geo-redundant Dedicated Interconnect

In this variation, the PEs within a Redundancy Group are located in different physical locations to provide geographic redundancy. This may be desirable, for example, to protect against natural disasters or the like. A dedicated interconnect is provided to link the PEs, which is a costly option, especially when considering the possibility of providing multiple such links for interconnect robustness. The resiliency mechanisms for the interconnect are similar to those highlighted in the co-located interconnect counterpart.

```
+=================+     +-----------------+
|CO 1             |     |                 |
|  +-----+        |     |                 |
|  | PE1 |________|_____|                 |
|  |     |        |     |                 |
|  +-----+        |     |                 |
+=====||==========+     |                 |
      || ICCP           |      Core       |
+=====||==========+     |     Network     |
|  +-----+        |     |                 |
|  | PE2 |________|_____|                 |
|  |     |        |     |                 |
|  +-----+        |     |                 |
|CO 2             |     |                 |
+=================+     +-----------------+
```

Figure 4: ICCP Geo-redundant PEs Dedicated Interconnect Scenario

3.2.4. Geo-redundant Shared Interconnect

In this scenario, the PEs of an RG are located in different physical locations and the interconnect for ICCP is provided over the PSN network to which the PEs are connected. This interconnect option is more likely to be the one used for geo-redundancy as it is more economically appealing compared to the geo-redundant dedicated interconnect. The resiliency mechanisms that can be employed to guarantee the robustness of the ICCP transport are PSN Layer mechanisms as has been described in the "Co-located Shared Interconnect" section above.

```
           +=================+     +-----------------+
           |CO 1             |     |                 |
           |  +-----+        |     |                 |
           |  | PE1 |________|_____|                 |
           |  |     |<==================+            |
           |  +-----+   ICCP |     |  ||             |
           +=================+     |  ||             |
                                   |  ||   Core      |
           +=================+     |  ||  Network    |
           |  +-----+        |     |  ||             |
           |  | PE2 |________|_____|  ||             |
           |  |     |<==================+            |
           |  +-----+        |     |                 |
           |CO 2             |     |                 |
           +=================+     +-----------------+
```

Figure 5: ICCP Geo-redundant PEs Shared Interconnect Scenario

3.3. ICCP Requirements

The Inter-chassis Communication Protocol SHOULD satisfy the following requirements:

-i. Provide a control channel for communication between PEs in a Redundancy Group (RG). Nodes may be co-located or remote (refer to "Interconnect Scenarios" section above). It is expected that client applications which make use of ICCP services will only use this channel to communicate control information and not data-traffic. As such the protocol should cater for relatively low bandwidth, low-delay and highly reliable message transfer.

-ii. Accommodate multiple client applications (e.g. multi-chassis LACP, PW redundancy, SONET APS, etc...). This implies that the messages should be extensible (e.g. TLV-based) and the protocol should provide a robust application registration and versioning scheme.

-iii. Provide reliable message transport and in-order delivery between nodes in a RG with secure authentication mechanisms built into the protocol. The redundancy applications that are clients of ICCP expect reliable message transfer, and as such will assume that the protocol takes care of flow-control and retransmissions. Furthermore, given that the applications will rely on ICCP to communicate data used to synchronize state-machines on disparate nodes, it is critical that ICCP guarantees in-order message delivery. Loss of messages or out-of-sequence messages would have adverse side-effects to the operation of the client applications.

-iv. Provide a common mechanism to actively monitor the health of PEs in an RG. This mechanism will be used to detect PE node failure and inform the client applications. The applications require this to trigger failover according to the procedures of the employed redundancy protocol on the AC and PW. It is desired to achieve sub-second detection of loss of remote node (˜ 50 - 150 msec) in order to give the client applications (redundancy mechanisms) enough reaction time to achieve sub-second service restoration time.

-v. Provide asynchronous event-driven state update, independent of periodic messages, for immediate notification of client applications' state changes. In other words, the transmission of messages carrying application data should be on-demand rather than timer-based to minimize inter-chassis state synchronization delay.

-vi. Accommodate multi-link and multi-hop interconnect between nodes. When the devices within an RG are located in different physical locations, the physical interconnect between them will comprise of a network rather than a link. As such, ICCP should accommodate the case where the interconnect involves multiple hops. Furthermore, it is possible to have multiple (redundant) paths or interconnects between a given pair of devices. This is true for both the co-located and geo-redundant scenarios. ICCP should handle this as well.

-vii. Ensure transport security between devices in an RG. This is especially important in the scenario where the members of an RG are located in different physical locations and connected over a shared network (e.g. PSN).

-viii. Must allow operator to statically configure members of RG. Auto-discovery may be considered in the future.

-ix. Allow for flexible RG membership. It is expected that only two nodes per an RG will cover most of the redundancy applications for common deployments. ICCP should not preclude supporting more than two nodes in an RG by virtue of design. Furthermore, it is required to allow a single node to be member of multiple RGs simultaneously.

4. ICC LDP Protocol Extension Specification

To address the requirements identified in the previous section, ICCP is modeled to comprise of three layers:

-i. Application Layer: This provides the interface to the various redundancy applications that make use of the services of ICCP. ICCP is concerned with defining common connection management procedures and the formats of the messages exchanged at this layer; however, beyond that, it does not impose any restrictions on the procedures or state-machines of the clients, as these are deemed application-specific and lie outside the scope of ICCP. This guarantees implementation inter-operability without placing any unnecessary constraints on internal design specifics.

-ii. Inter Chassis Communication (ICC) Layer: This layer implements the common set of services which ICCP offers to the client applications. It handles protocol versioning, RG membership, Redundant Object identification, PE node identification and ICCP connection management.

-iii. Transport Layer: This layer provides the actual ICCP message transport. It is responsible for addressing, route resolution, flow-control, reliable and in-order message delivery, connectivity resiliency/redundancy and finally PE node failure detection. The Transport layer may differ depending on the Physical Layer of the inter-connect.

4.1. LDP ICCP Capability Advertisement

When an RG is enabled on a particular PE, the capability of supporting ICCP must be advertised to all LDP peers in that RG. This is achieved by using the methods in [RFC5561] and advertising the ICCP LDP capability TLV. If an LDP peer supports the dynamic capability advertisement, this can be done by sending a new capability message with the S bit set for the ICCP capability TLV when the first RG is enabled on the PE. If the peer does not support dynamic capability advertisement, then the ICCP TLV MUST be included in the LDP initialization procedures in the capability parameter [RFC5561].

4.2. RG Membership Management

ICCP defines a mechanism that enables PE nodes to manage their RG membership. When a PE is configured to be a member of an RG, it will first advertise the ICCP capability to its peers. Subsequently, the PE sends an RG Connect message to the peers that have also advertised ICCP capability. The PE then waits for the peers to send their own RG Connect messages, if they haven't done so already. For a given RG, the ICCP connection between two devices is considered to be operational only when both have sent and received ICCP RG Connect messages for that RG.

If a PE that has sent a particular RG Connect message doesn't receive a corresponding RG Connect (or a Notification message with NAK) from a destination, it will remain in a state expecting the corresponding RG Connect message (or Notification message). The RG will not become operational until the corresponding RG Connect Message has been received. If a PE that has sent an RG Connect message receives a Notification message with a NAK, it will stop attempting to bring up the ICCP connection immediately. The PE MUST resume bringing up the connection after it receives an RG Connect message from the peer PE for the RG in question. This is achieved by responding to the incoming RG Connect message with an appropriate RG Connect.

A device MUST send a NAK for an RG Connect message if at least one of the following conditions is satisfied:

-i. the PE is not a member of the RG;

-ii. the maximum number of simultaneous ICCP connections that the PE can handle is exceeded.

A PE sends an RG Disconnect message to tear down the ICCP connection for a given RG. This is a unilateral operation and doesn't require any acknowledgement from the other PEs. Note that the ICCP connection for an RG MUST be operational before any client application can make use of ICCP services in that RG.

4.2.1. ICCP Connection State Machine

The ICCP Connection state machine is defined to have six states as depicted in the state transition table and state transition diagram that follow.

ICCP Connection State Transition Table

STATE	EVENT	NEW STATE
NON EXISTENT	LDP session established	INITIALIZED
INITIALIZED	Transmit LDP ICCP Capability	CAPSENT
	Receive LDP ICCP Capability Action: Transmit LDP ICCP Capability	CAPREC
	LDP session torn down	NON EXISTENT
CAPSENT	Receive LDP ICCP Capability	CAPREC
	LDP session torn down	NON EXISTENT
CAPREC	Transmit RG Connect Message	CONNECTING
	Receive acceptable RG Connect Message Action: Transmit RG Connect Message	OPERATIONAL
	Receive any other ICCP Message Action: Transmit NAK in RG Notification Message	CAPREC
	LDP session torn down	NON EXISTENT
CONNECTING	Receive acceptable RG Connect Message	OPERATIONAL
	Receive any other ICCP Message Action: Transmit NAK in RG Notification Message	CAPREC
	LDP session torn down	NON EXISTENT
OPERATIONAL	Receive acceptable RG Disconnect Message	CAPREC
	Transmit RG Disconnect Message	CAPREC
	LDP session torn down	NON EXISTENT

ICCP Connection State Transition Diagram

```
                                +------------+
                                |            |
          +------------------->|NON EXISTENT|    LDP session torn down
          |                   |            |<--------------------------+
          |                   +------------+                           |
          |         LDP session  |    ^ LDP session                    |
          |         established  |    | torn down                      |
          |                      V    |                                |
          |                  +-----------+                             |
   LDP    |                  |           |  Tx LDP ICCP                |
   session|                  |INITIALIZED|    capability               |
   torn   |              +---|           |---------------+             |
   down   |  Rx other    |   +-----------+               |             |
          |  ICCP msg/   |Rx LDP ICCP                    |             |
          |   Tx NAK     |  capability/                  |             |
          |      +---+   |Tx LDP ICCP capability         |             |
          |      |   |   |                               |             |
          |      V   |   V                               V             |
          |   +-----------+   Rx LDP ICCP         +--------+           |
          +---|           |     capability        |        |           |
              |CAPREC     |<----------------------|CAPSENT |---------->+
          +---|           |-------------------+   |        |           |
          |   +-----------+                   |   +--------+           |
          |       ^    ^                      |                        |
   Tx     |       |    |                      |                        |
   RG     |       |    |Rx RG Disconnect msg  |                        |
   Connect|       |    | or                   |Rx RG Connect msg /     |
   Msg    |       |    |Tx RG Disconnect msg  | Tx RG Connect msg       |
          |       |    |                      V                        |
          |       |    |                    +------------+             |
          |       |    +--------------------|            |             |
          |       |                         |OPERATIONAL |------------>+
          |       |                         |            |             |
          |       |Rx other ICCP msg/       +------------+             |
          |       | Tx NAK                        ^                    |
          |       |                               |                    |
          |      +----------+  Rx RG Connect msg  |                    |
          |      |          |---------------------+                    |
          +----->|CONNECTING|                                          |
                 |          |----------------------------------------->+
                 +----------+
```

4.3. Redundant Object Identification

ICCP offers its client applications a uniform mechanism for identifying links, ports, forwarding constructs and more generally objects (e.g. interfaces, pseudowires, VLANs, etc...) that are being protected in a redundant setup. These are referred to as Redundant Objects (RO). An example of an RO is a multi-chassis link-aggregation group that spans two PEs. ICCP introduces a 64-bit opaque identifier to uniquely identify ROs in an RG. This identifier, referred to as Redundant Object ID (ROID), MUST match between RG members for the protected object in question. That allows separate systems in an RG to use a common handle to reference the protected entity irrespective of its nature (e.g. physical or virtual) and in a manner that is agnostic to implementation specifics. Client applications that need to synchronize state pertaining to a particular RO SHOULD embed the corresponding ROID in their TLVs.

4.4. Application Connection Management

ICCP 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. It is assumed that the association of applications with RGs is known a priori, e.g. by means of device configuration. ICCP then sends an Application-specific Connect TLV (carried in RG Connect message), on behalf of each client application, to each remote PE within the RG. The client may piggyback application-specific information in that Connect TLV, which for example can be used to negotiate parameters or attributes prior to bringing up the actual application connection. The procedures for bringing up the application connection are similar to those of the ICCP connection: An application connection between two nodes is up only when both nodes have sent and received RG Connect Messages with the proper Application-specific Connect TLVs. A PE MUST send a Notification Message to NAK an application connection request if one of the following conditions is encountered:

-i. the application doesn’t exist or is not configured for that RG;

-ii. the application connection count exceeds the PE’s capabilities.

When a PE receives such a NAK notification, it MUST stop attempting to bring up the application connection until it receives a new application connection request from the remote PE. This is done by responding to the incoming RG Connect message (carrying an Application-specific Connect TLV) with an appropriate RG Connect message (carrying a corresponding Application-specific Connect TLV).

When an application is stopped on a device or it is no longer associated with an RG, it MUST signal ICCP to trigger sending an Application-specific Disconnect TLV (in the RG Disconnect message). This is a unilateral notification to the other PEs within an RG, and as such doesn’t trigger any response.

4.4.1. Application Versioning

During application connection setup time, a given application on one PE can negotiate with its counterpart on a peer PE the proper application version to use for communication. If no common version is agreed upon, then the application connection is not brought up. This is achieved through the following set of rules:

- If an application receives an Application-specific Connect TLV with a version number that is higher than its own, it MUST send a Notification message with a NAK TLV indicating status code "Incompatible Protocol Version" and supplying the version that is locally supported by the PE.

- If an application receives an Application-specific Connect TLV with a version number that is lower than its own, it MAY respond with an RG Connect that has an Application-specific Connect TLV using the same version that was received. Alternatively, the application MAY respond with a Notification message to NAK the request using the "Incompatible Protocol Version" code, and supplying the version that is supported. The above allows an application to operate in either backwards compatible or incompatible mode.

- If an application receives an Application-specific Connect TLV with a version that is equal to its own, then the application MUST honor or reject the request based on whether the application is configured for the RG in question, and whether or not the application connection count has been exceeded.

4.4.2. Application Connection State Machine

The Application Connection state machine has six states as described in the state transition table and diagram that follow.

ICCP Application Connection State Transition Table

STATE	EVENT	NEW STATE
NON EXISTENT	ICCP connection established	RESET
RESET	ICCP connection torn down	NON EXISTENT
	Transmit Application Connect TLV	CONNSENT
	Receive Application Connect TLV	CONNREC
	Receive any other Application TLV Action: Transmit NAK TLV	RESET
CONNSENT	Receive NAK TLV	RESET
	Receive Application Connect TLV with A-bit=1 Action: Transmit Application Connect TLV with A-bit=1	OPERATIONAL
	Receive any other Application TLV Action: Transmit NAK TLV	RESET
	ICCP connection torn down	NON EXISTENT
CONNREC	Transmit NAK TLV	RESET
	Transmit Application Connect TLV with A-bit=1	CONNECTING
	Receive any Application TLV except Connect Action: Transmit NAK TLV	RESET
	ICCP connection torn down	NON EXISTENT
CONNECTING	Receive Application Connect TLV with A-bit=1	OPERATIONAL
	Receive any other Application TLV Action: Transmit NAK TLV	RESET
	ICCP connection torn down	NON EXISTENT
OPERATIONAL	Receive Application Disconnect TLV	RESET
	Transmit Applicaton Disconnect TLV	RESET

```
                  ICCP connection torn down              NON EXISTENT


   ICCP Application Connection State Transition Diagram
                            +------------+
                            |            |
          +---------------->|NON EXISTENT|  ICCP connection torn down
          |                 |            |<--------------------------+
          |                 +------------+                           |
          |     ICCP connection|    ^ ICCP connection                |
          |       established  |    | torn down                      |
          |                    |    |                                |
          |                    V    |          Rx other App TLV/     |
          |                +-----------+<-----+  Tx NAK              |
   ICCP   |    Rx App      |           |      |                      |
   connect|    Connect TLV |   RESET   |------+                      |
   torn   |  +-------------|           |---------------+             |
   down   |  |             +-----------+    Tx App     |             |
          |  |              ^  ^   ^  ^     Connect TLV|             |
          |  |      Tx NAK  |  |   |  |                |             |
          |  |      or      |  |   |  |                |             |
          |  |      Rx non  |  |   |  |                |             |
          |  |      Connect |  |   |  |                |             |
          |  V      TLV/Tx NAK |   |  |Rx NAK TLV      V             |
          | +-----------+   |  |   |  |or       +--------+           |
          +-|           |---+  |   |  +---------|        |           |
            |CONNREC    |      |   |   Rx other |CONNSENT|---------->+
          +-|           |      |   |   App TLV/ |        |           |
          | +-----------+      |   |     Tx NAK +--------+           |
          |                    |   |                 |Rx App Connect |
          |  Tx App Connect    |   |                 |TLV (A=1) /    |
          |  TLV (A=1)         |   |Rx App Disconn   | Tx App        |
          |                    |   |or               | Connect TLV   |
          |                    |   |Tx App Disconn   V (A=1)         |
          |                    |   |      +------------+             |
          |                    |   +------|            |             |
          |       Rx other App |          |OPERATIONAL |------------>+
          |       TLV / Tx NAK |          |            |             |
          |             +------+          +------------+             |
          |             |                       ^ Rx App Connect     |
          |    +----------+                     | TLV (A=1)          |
          |    |          |---------------------+                    |
          +--->|CONNECTING|                                          |
               |          |----------------------------------------->+
               +----------+
```

4.5. Application Data Transfer

When an application has information to transfer over ICCP it triggers the transmission of an Application Data message. ICCP guarantees in-order and loss-less delivery of data. An application may NAK a message or a set of one or more TLVs within a message by using the Notification Message with NAK TLV. Furthermore, an application may implement its own ACK mechanism, if deemed required, by defining an application-specific TLV to be transported in an Application Data message.

It is left up to the application to define the procedures to handle the situation where a PE receives a NAK in response to a transmitted Application Data message. Depending on the specifics of the application, it may be favorable to have the PE, which sent the NAK, explicitly request retransmission of data. On the other hand, for certain applications it may be more suitable to have the original sender of the Application Data message handle retransmissions in response to a NAK. ICCP supports both models.

4.6. Dedicated Redundancy Group LDP session

For certain ICCP applications, it is required to exchange a fairly large amount of RG information in a very short period of time. In order to better distribute the load in a multiple processor system, and to avoid head of line blocking to other LDP applications, it may be required to initiate a separate TCP/IP session between the two LDP speakers.

This procedure is OPTIONAL, and does not change the operation of LDP or ICCP.

A PE that requires a separate LDP session will advertise a separate LDP adjacency with a non-zero label space identifier. This will cause the remote peer to open a separate LDP session for this label space. No labels need to be advertised in this label space, as it is only used for one or a set of ICCP RGs. All relevant LDP and ICCP procedures still apply as described in the relevant documents.

5. ICCP PE Node Failure Detection Mechanism

ICCP provides its client applications a notification when a remote PE that is member of the RG fails. This is used by the client applications to trigger failover according to the procedures of the employed redundancy protocol on the AC and PW. To that end, ICCP does not define its own KeepAlive mechanism for purpose of monitoring the health of remote PE nodes, but rather reuses existing fault detection mechanisms. The following mechanisms may be used by ICCP to detect PE node failure:

- BFD

  Run a BFD session [RFC5880] between the PEs that are members of a given RG, and use that to detect PE node failure. This assumes that resiliency mechanisms are in place to protect connectivity to the remote PE nodes, and hence loss of BFD periodic messages from a given PE node can only mean that the node itself has failed.

- IP Reachability Monitoring

  It is possible for a PE to monitor IP layer connectivity to other members of an RG that are participating in IGP/BGP. When connectivity to a given PE is lost, the local PE interprets that to mean loss of the remote PE node. This assumes that resiliency mechanisms are in place to protect the route to the remote PE nodes, and hence loss of IP reachability to a given node can only mean that the node itself has failed.

It is worth noting here that loss of the LDP session with a PE in an RG is not a reliable indicator that the remote PE itself is down. It is possible, for e.g. that the remote PE encounters a local event that leads to resetting the LDP session, while the PE node remains operational for purpose of traffic forwarding.

6. ICCP Message Formats

This section defines the messages exchanged at the Application and ICC layers.

6.1. Encoding ICC into LDP Messages

ICCP requires reliable, in-order, state-full message delivery, as well as capability negotiation between PEs. The LDP protocol offers all these features, and is already in wide use in the applications that would also require the ICCP protocol extensions. For these reasons, ICCP takes advantage of the already defined LDP protocol infrastructure.

[RFC5036] Section 3.5 defines a generic LDP message structure. A new set of LDP message types is defined to communicate the ICCP information. LDP message types in the range of 0x700 to 0x7ff will be used for ICCP.

Message types are allocated by IANA, and requested in the IANA section below.

6.1.1. ICC Header

Every ICCP message comprises of an ICC specific LDP Header followed by message data. The format of the ICC Header is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|   Message Type              |      Message Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Message ID                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type=0x0005 (ICC RG ID)     |           Length=4            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          ICC RG ID                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                     Mandatory Parameters                      |
~                                                               ~

+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                     Optional Parameters                       |
~                                                               ~

+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U-bit

  Unknown message bit. Upon receipt of an unknown message, if U is clear (=0), a notification is returned to the message originator; if U is set (=1), the unknown message is silently ignored. The following sections which define messages specify a value for the U-bit.

- Message Type

  Identifies the type of the ICCP message, must be in the range of 0x0700 to 0x07ff.

- Message Length

  Two octet integer specifying the total length of this message in octets, excluding the U-bit, Message Type and Length fields.

- Message ID

  Four octet value used to identify this message. Used by the sending PE to facilitate identifying RG Notification messages that may apply to this message. A PE sending an RG Notification message in response to this message SHOULD include this Message ID in the "NAK TLV" of the RG Notification message; see Section "RG Notification Message".

- ICC RG ID TLV

  A TLV of type 0x0005, length 4, containing 4 octets unsigned integer designating the Redundancy Group which the sending device is member of. RG ID value 0x00000000 is reserved by the protocol.

- Mandatory Parameters

  Variable length set of required message parameters. Some messages have no required parameters.

  For messages that have required parameters, the required parameters MUST appear in the order specified by the individual message specifications in the sections that follow.

- Optional Parameters

  Variable length set of optional message parameters. Many messages have no optional parameters.

For messages that have optional parameters, the optional parameters may appear in any order.

6.1.2. Message Encoding

The generic format of an ICC parameter is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|       Type                |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   TLV(s)                                                      |
~                                                               ~

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U-bit

  Unknown TLV bit. Upon receipt of an unknown TLV, if U is clear (=0), a notification MUST be returned to the message originator and the entire message MUST be ignored; if U is set (=1), the unknown TLV MUST be silently ignored and the rest of the message processed as if the unknown TLV did not exist. The sections following that define TLVs specify a value for the U-bit.

- F-bit

  Forward unknown TLV bit. This bit applies only when the U-bit is set and the LDP message containing the unknown TLV is to be forwarded. If F is clear (=0), the unknown TLV is not forwarded with the containing message; if F is set (=1), the unknown TLV is forwarded with the containing message. The sections following that define TLVs specify a value for the F-bit. By setting both the U- and F-bits, a TLV can be propagated as opaque data through nodes that do not recognize the TLV.

- Type

  Fourteen bits indicating the parameter type.

- Length

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

- TLV(s): A set of 0 or more TLVs, that vary according to the message type.

6.1.3. ROID Encoding

The Redundant Object Identifier (ROID) is a generic opaque handle that uniquely identifies a Redundant Object (e.g. link, bundle, VLAN, etc...) which is being protected in an RG. It is encoded as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              ROID                             |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where: ROID is an 8 octets field encoded as an unsigned integer. The ROID value of 0 is reserved.

The ROID is carried within application specific TLVs.

6.2. RG Connect Message

The RG Connect Message is used to establish the ICCP RG connection in addition to individual Application connections between PEs in an RG. An RG Connect message with no "Application-specific connect TLV" signals establishment of the ICCP RG connection. Whereas, an RG Connect message with a valid "Application-specific connect TLV" signals the establishment of an Application connection, in addition to the ICCP RG connection if the latter is not already established.

An implementation MAY send a dedicated RG Connect message to set up the ICCP RG connection and a separate RG Connect message per client application. However, all implementations MUST support the receipt of an RG Connect message that triggers the setup of the ICCP RG connection as well as a single Application connection simultaneously.

A PE sends an RG Connect Message to declare its membership in a Redundancy Group. One such message should be sent to each PE that is member of the same RG. The set of PEs to which RG Connect Messages should be transmitted is known via configuration or an auto-discovery mechanism that is outside the scope of this specification. If a device is member of multiple RGs, it MUST send separate RG Connect Messages for each RG even if the receiving device(s) happen to be the same.

The format of the RG Connect Message is as follows:

-i. ICC header with Message type = "RG Connect Message" (0x0700)
-ii. ICC Sender Name TLV
-iii. Zero or one Application-specific connect TLV

The currently defined Application-specific connect TLVs are:

- PW Redundancy Connect TLV

- mLACP Connect TLV

The details of these TLVs are discussed in the "Application TLVs" section.

The RG Connect message can contain zero or one Application-specific connect TLV, but no application connect TLV can be sent more than once.

6.2.1. ICC Sender Name TLV

A TLV that carries the hostname of the sender encoded in UTF-8. This is used primarily for purpose of management of the RG and easing network operations. The specific format is shown below:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|       Type = 0x0001       |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Sender Name                                                  |
+                                             +-+-+-+-+-+-+-+-+-+
~                                             ~
|      ...                                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U=F=0

- Type set to 0x0001 (from ICC parameter name space).

- Length

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

- Sender Name

  Hostname of sending device encoded in UTF-8, and SHOULD NOT exceed 80 characters.

6.3. RG Disconnect Message

The RG Disconnect Message serves dual-purpose: to signal that a particular Application connection is being closed within an RG, or that the ICCP RG connection itself is being disconnected because the PE wishes to leave the RG. The format of this message is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|   Message Type=0x0701       |      Message Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Message ID                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type=0x0005 (ICC RG ID)     |           Length=4            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     ICC RG ID                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Disconnect Code TLV                          |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Optional Application-specific Disconnect TLV        |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Optional Parameter TLVs                          |
+                                                               +
|                                                               |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U-bit

  U=0

- Message Type

  The message type for RG Disconnect Message is set to (0x0701)

- Length

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

- Message ID

  Defined in the "ICC Header" section above.

- ICC RG ID

  Defined in the "ICC Header" section above.

- Disconnect Code TLV

  The format of this TLV is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|         Type=0x0004       |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      ICCP Status Code                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U,F Bits

  both U and F are set to 0.

- Type

  set to "Disconnect Code TLV" (0x0004)

- Length

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

- ICCP Status Code

  A status code that reflects the reason for the disconnect message. Allowed values are "ICCP RG Removed" and "ICCP Application Removed from RG".

- Optional Application-specific Disconnect TLV

  Zero or one Application-specific Disconnect TLVs which are defined later in the document. If the RG Disconnect message has a status code of "RG Removed", then it MUST NOT contain any Application-specific Disconnect TLVs, as the sending PE is signaling that it has left the RG and, thus, is disconnecting the ICCP RG connection, with all associated client application connections. If the message has a status code of "Application Removed from RG", then it MUST contain exactly one Application-specific Disconnect TLV, as the sending PE is only tearing down the connection for the specified application. Other applications, and the ICCP RG connection are not to be affected.

- Optional Parameter TLVs

  None are defined for this message in this document. This is specified to allow for future extensions.

6.4. RG Notification Message

A PE sends an RG Notification Message to indicate one of the following: to reject an ICCP connection, to reject an application connection, to NAK an entire message or to NAK one or more TLV(s) within a message. The Notification message can only be sent to a PE that is already part of an RG.

The RG Notification Message MUST NOT be used to NAK messages or TLVs corresponding to multiple ICCP applications simultaneously. In other words, there is a limit of at most a single ICCP application per RG Notification Message.

The format of the RG Notification Message is:

 -i. ICC header with Message type = "RG Notification Message" (0x0702)
-ii. Notification Message TLVs.

The currently defined Notification message TLVs are:

 -i. ICC Sender Name TLV
-ii. NAK TLV.

6.4.1. Notification Message TLVs

The ICC Sender Name TLV uses the same format as in the RG Connect message, and was described above.

The NAK TLV is defined as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|         Type=0x0002       |     Length                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        ICCP Status Code                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Rejected Message ID                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Optional TLV(s)                             |
+                                                               +
|                                                               |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U,F Bits

  both U and F are set to 0.

- Type

  set to "NAK TLV" (0x0002)

- Length

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

- ICCP Status Code

  A status code that reflects the reason for the NAK TLV. Allowed values are:

  -i. Unknown RG (0x00010001)

  This code is used to reject a new incoming ICCP connection for an RG that is not configured on the local PE. When this code is used, the Rejected Message ID field MUST contain the message ID of the rejected "RG Connect" message.

-ii. ICCP Connection Count Exceeded (0x00010002)

This is used to reject a new incoming ICCP connection that would cause the local PE's ICCP connection count to exceed its capabilities. When this code is used, the Rejected Message ID field MUST contain the message ID of the rejected "RG Connect" message.

-iii. Application Connection Count Exceeded (0x00010003)

This is used to reject a new incoming application connection that would cause the local PE's ICCP connection count to exceed its capabilities. When this code is used, the Rejected Message ID field MUST contain the message ID of the rejected "RG Connect" message and the corresponding Application Connect TLV MUST be included in the "Optional TLV".

-iv. Application not in RG (0x00010004)

This is used to reject a new incoming application connection when the local PE doesn't support the application, or the application is not configured in the RG. When this code is used, the Rejected Message ID field MUST contain the message ID of the rejected "RG Connect" message and the corresponding Application Connect TLV MUST be included in the "Optional TLV".

-v. Incompatible Protocol Version (0x00010005)

This is used to reject a new incoming application connection when the local PE has an incompatible version of the application. When this code is used, the Rejected Message ID field MUST contain the message ID of the rejected "RG Connect" message and the corresponding Application Connect TLV MUST be included in the "Optional TLV".

-vi. Rejected Message (0x00010006)

This is used to reject an RG Application Data message, or one or more TLV(s) within the message. When this code is used, the Rejected Message ID field MUST contain the message ID of the rejected "RG Application Data" message.

-vii. ICCP Administratively Disabled (0x00010007)

This is used to reject any ICCP messages from a peer from which the PE is not allowed to exchange ICCP messages due to local administrative policy.

- Rejected Message ID

If non-zero, four octets value that identifies the peer message to which the NAK TLV refers. If zero, no specific peer message is being identified.

- Optional TLV(s)

A set of one or more optional TLVs. If the status code is "Rejected Message" then this field contains the TLV(s) that were rejected. If the entire message is rejected, all its TLVs MUST be present in this field; otherwise, the subset of TLVs that were rejected MUST be echoed in this field.

If the status code is "Incompatible Protocol Version" then this field contains the original "Application Connect TLV" sent by the peer, in addition to the "Requested Protocol Version TLV" defined below:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0003             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Connection Reference        |   Requested Version           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

Both are set to 0.

- Type

set to 0x0003 for "Requested Protocol Version TLV"

- Length

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

- Connection Reference

  This field is set to the Type field of the Application specific Connect TLV that was rejected because of incompatible version.

- Requested Version

  The version of the application supported by the transmitting device. For this version of the protocol it is set to 0x0001.

6.5. RG Application Data Message

The RG Application Data Message is used to transport application data between PEs within an RG. A single message can be used to carry data from only one application. Multiple application TLVs are allowed in a single message, as long as all of these TLVs belong to the same application. The format of the Application Data Message is:

-i. ICC header with Message type = "RG Application Data Message" (0x703)
-ii. "Application specific TLVs"

The details of these TLVs are discussed in the "Application TLVs" section. All application specific TLVs in one RG Application Data Message MUST belong to a single application but MAY reference different ROs.

7. Application TLVs

7.1. Pseudowire Redundancy (PW-RED) Application TLVs

This section discusses the ICCP TLVs for the Pseudowire Redundancy application.

7.1.1. PW-RED Connect TLV

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

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type=0x0010             |    Length                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |      Protocol Version         |A|         Reserved            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      Optional Sub-TLVs                        |
   ~                                                               ~
   |                                                               |
   +                                 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |             ...                 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0010 for "PW-RED 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 PW-RED 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.

7.1.2. PW-RED Disconnect TLV

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

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type=0x0011             |    Length                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                        Optional Sub-TLVs                      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0011 for "PW-RED Disconnect TLV"

- Length

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

- Optional Sub-TLVs

  The only optional Sub-TLV defined for this version of the protocol is the "PW-RED Disconnect Cause" TLV defined next:

7.1.2.1. PW-RED Disconnect Cause TLV

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type=0x0019             |    Length                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      Disconnect Cause String                  |
   ~                                                               ~
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0019 for "PW-RED 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. Used for network management.

7.1.3. PW-RED Config TLV

The PW-RED Config TLV is used in the RG Application Data message and has the following format:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type = 0x0012           |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             ROID                              |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      PW Priority              |             Flags             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Service Name TLV                             |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            PW ID TLV or Generalized PW ID TLV                 |
~                                                               ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0012 for "PW-RED Config TLV"

- Length

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

- ROID

  As defined in the ROID section above.

- PW Priority

  Two octets Pseudowire Priority. Used to indicate which PW has better priority to go into Active state. Numerically lower numbers are better priority. In case of a tie, the PE with the numerically lower identifier (i.e. IP Address) has better priority.

- Flags

  Valid values are:

  -i. Synchronized (0x01)

      Indicates that the sender has concluded transmitting all pseudowire configuration for a given service.

  -ii. Purge Configuration (0x02)

      Indicates that the pseudowire is no longer configured for PW-RED operation.

- Sub-TLVs

  The "PW-RED Config TLV" includes the following two sub-TLVs:

-i. Service Name TLV

-ii. PW ID TLV or Generalized PW ID TLV

The format of the sub-TLVs is as follows:

7.1.3.1. Service Name TLV

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type                    |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Service Name                       |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0013 for "Service Name TLV"

- Length

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

- Service Name

  The name of the L2VPN service instance encoded in UTF-8 format and up to 80 character in length.

7.1.3.2. PW ID TLV

This TLV is used to communicate the configuration of PWs for VPWS.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type                    |     Length                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          Peer ID                              |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          Group ID                             |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          PW ID                                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0014 for "PW ID TLV"

- Length

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

- Peer ID

  Four octet LDP Router ID of the peer at the far end of the PW.

- Group ID

  Same as Group ID in [RFC4447] section 5.2.

- PW ID

  Same as PW ID in [RFC4447] section 5.2.

7.1.3.3. Generalized PW ID TLV

This TLV is used to communicate the configuration of PWs for VPLS.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type = 0x0015           |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   AGI Type    |    Length     |      Value                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                    AGI  Value (contd.)                        ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   AII Type    |    Length     |      Value                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                   SAII  Value (contd.)                        ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   AII Type    |    Length     |      Value                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                   TAII Value (contd.)                         ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F bits

  both set to 0.

- Type

  set to 0x0015

- Length

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

- AGI, AII, SAII and TAII

  defined in [RFC4447] section 5.3.2.

7.1.4. PW-RED State TLV

The PW-RED State TLV is used in the RG Application Data Message. This TLV is used by a device to report its PW status to other members in the RG.

The format of this TLV is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0016            |    Length                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              ROID                             |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Local PW State                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Remote PW State                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0016 for PW-RED State TLV.

- Length

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

- ROID

  As defined in the ROID section above.

- Local PW State

  The status of the PW as determined by the sending PE, encoded in the same format as the "Status Code" field of the "PW Status TLV" defined in [RFC4447].

- Remote PW State

  The status of the PW as determined by the remote peer of the sending PE. Encoded in the same format as the "Status Code" field of the "PW Status TLV" defined in [RFC4447]. The same code points listed above are used here as well.

7.1.5. PW-RED Synchronization Request TLV

The PW-RED Synchronization Request TLV is used in the RG Application Data message. This TLV is used by a device to request from its peer to retransmit configuration or operational state. The following information can be requested:

- configuration and/or state for one or more pseudowires

- configuration and/or state for all pseudowires

- configuration and/or state for all pseudowires in a given service

The format of the TLV is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0017             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Request Number           |C|S|    Request Type           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Optional Sub-TLVs                          |
~                                                               ~
|                                                               |
+                                +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             ...                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0017 for "PW-RED Synchronization Request TLV"

- Length

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

- Request Number

  2 octets. Unsigned integer uniquely identifying the request. Used to match the request with a response. The value of 0 is reserved for unsolicited synchronization, and MUST NOT be used in the PW-

RED 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 pseudowire(s)
0x01   Request Data for all pseudowires in specified service(s)
0x3FFF Request All Data
```

- Optional Sub-TLVs

  A set of zero or more TLVs, as follows:

  If the Request Type field is set to (0x00), then this field contains one or more PW ID TLV(s) or Generalized PW ID TLV(s). If the Request Type field is set to (0x01), then this field contains one or more Service Name TLV(s). If the Request Type field is set to (0x3FFF), then this field MUST be empty.

7.1.6. PW-RED Synchronization Data TLV

The PW-RED Synchronization Data TLV is used in the RG Application Data mesage. A pair of these TLVs is used by a device to delimit a set of TLVs that are sent in response to a PW-RED 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 PW-RED Synchronization Data TLVs are also used for unsolicited advertisements of complete PW-RED configuration and operational state data. In this case, the Request Number field MUST be set to 0.

This TLV has the following format:

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type=0x0018             |    Length                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Request Number            |     Flags                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0018 for "PW-RED Synchronization Data TLV"

- Length

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

- Request Number

  2 octets. Unsigned integer identifying the Request Number from the "PW-RED Synchronization Request TLV" which solicited this synchronization data response.

- Flags

  2 octets, response flags encoded as follows:

      0x00 Synchronization Data Start
      0x01 Synchronization Data End

7.2. Multi-chassis LACP (mLACP) Application TLVs

This section discusses the ICCP TLVs for Ethernet attachment circuit redundancy using the multi-chassis LACP (mLACP) application.

7.2.1. mLACP Connect TLV

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

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0030             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Protocol Version         |A|        Reserved             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Optional Sub-TLVs                          |
~                                                               ~
|                                                               |
+                                 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             ...                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0030 for "mLACP 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 an mLACP 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.

7.2.2. mLACP Disconnect TLV

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

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0031             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Optional Sub-TLVs                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0031 for "mLACP Disconnect TLV"

- Length

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

- Optional Sub-TLVs

  The only optional Sub-TLV defined for this version of the protocol is the "mLACP Disconnect Cause" TLV defined next:

7.2.2.1. mLACP Disconnect Cause TLV

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x003A             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Disconnect Cause String                  |
~                                                               ~
```

```
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x003A for "mLACP 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. Used for network management.

7.2.3. mLACP System Config TLV

The mLACP System Config TLV is sent in the RG Application Data message. This TLV announces the local node’s LACP System Parameters to the RG peers.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type=0x0032             |    Length                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          System ID                            |
   +                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                               |         System Priority       |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |    Node ID    |
   +-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0032 for "mLACP System Config TLV"

- Length

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

- System ID

  6 octets field encoding the System ID used by LACP as specified in [IEEE-802.3] section 43.3.2.

- System Priority

  2 octets encoding the LACP System Priority as defined in [IEEE-802.3] section 43.3.2.

- Node ID

  One octet, LACP node ID. Used to ensure that the LACP Port Numbers are unique across all devices in an RG. Valid values are in the range 0 - 7. Uniqueness of the LACP Port Numbers across RG members is ensured by encoding the Port Numbers as follows:

    - Most significant bit always set to 1
    - The next 3 most significant bits set to Node ID
    - Remaining 12 bits freely assigned by the system

7.2.4. mLACP Aggregator Config TLV

The mLACP Aggregator Config TLV is sent in the RG Application Data message. This TLV is used to notify RG peers about the local configuration state of an aggregator.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0036             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              ROID                             |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Aggregator ID           |    MAC Address                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Actor Key               |    Member Ports Priority      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     | Agg Name Len  |    Aggregator Name            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
~                                                               ~
|                                         ...                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0036 for "mLACP Aggregator Config TLV"

- Length

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

- ROID

  Defined in the ’ROID Encoding’ section above.

- Aggregator ID

  Two octets, LACP Aggregator Identifier as specified in [IEEE-802.3] section 43.4.6

- MAC Address

  Six octets encoding the Aggregator MAC address.

- Actor Key

  Two octets, LACP Actor Key for the corresponding Aggregator, as specified in [IEEE-802.3] section 43.3.5.

- Member Ports Priority

  Two octets, LACP administrative port priority associated with all interfaces bound to the Aggregator. This field is valid only when the "Flags" field has "Priority Set" asserted.

- Flags

  Valid values are:

  -i. Synchronized (0x01)

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

  -ii. Purge Configuration (0x02)

    Indicates that the Aggregator is no longer configured for mLACP operation.

  -iii. Priority Set (0x04)

    Indicates that the "Member Ports Priority" field is valid.

- Agg Name Len

  One octet, length of the "Aggregator Name" field in octets.

- Aggregator Name

  Aggregator name encoded in UTF-8 format, up to a maximum of 20 characters. Used for ease of management.

7.2.5. mLACP Port Config TLV

The mLACP Port Config TLV is sent in the RG Application Data message. This TLV is used to notify RG peers about the local configuration state of a port.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0033             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Port Number             |    MAC Address                |
+-------------------------------+                               +
|                                                               |
+---------------------------------------------------------------+
|       Actor Key               |     Port Priority             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Port Speed                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     | Port Name Len |         Port Name             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
~                                                               ~
|                                        ...                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0033 for "mLACP Port Config TLV"

- Length

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

- Port Number

  Two octets, LACP Port Number for the corresponding interface as specified in [IEEE-802.3] section 43.3.4. The Port Number MUST be encoded with the Node ID as was discussed above.

- MAC Address

  Six octets encoding the port MAC address.

- Actor Key

  Two octets, LACP Actor Key for the corresponding interface, as specified in [IEEE-802.3] section 43.3.5.

- Port Priority

  Two octets, LACP administrative port priority for the corresponding interface, as specified in [IEEE-802.3] section 43.3.4. This field is valid only when the "Flags" field has "Priority Set" asserted.

- Port Speed

  Four octets integer encoding the port’s current bandwidth in units of 1,000,000 bits per second. This field corresponds to the ifHighSpeed object of IF-MIB [RFC2863].

- Flags

  Valid values are:

  -i. Synchronized (0x01)

  Indicates that the sender has concluded transmitting all member link port configurations for a given Aggregator.

  -ii. Purge Configuration (0x02)

  Indicates that the port is no longer configured for mLACP operation.

  -iii. Priority Set (0x04)

  Indicates that the "Port Priority" field is valid.

- Port Name Len

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

- Port Name

  Port (interface) name encoded in UTF-8 format, up to a maximum of 20 characters.

7.2.6. mLACP Port Priority TLV

The mLACP Port Priority TLV is sent in the RG Application Data message. This TLV is used by a device to either advertise its operational Port Priority to other members in the RG, or to authoritatively request that a particular member of an RG change its port priority.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |U|F|   Type=0x0034             |    Length                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          OpCode               |          Port Number          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Aggregator ID         |    Last Port Priority         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |    Current Port Priority      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0034 for "mLACP Port Priority TLV"

- Length

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

- OpCode

  Two octets identifying the operational code-point for the TLV, encoded as follows:

      0x00 Local Priority Change Notification
      0x01 Remote Request for Priority Change

- Port Number

  2 octets field representing the LACP Port Number as specified in [IEEE-802.3] section 43.3.4. When the value of this field is 0, it denotes all ports bound to the Aggregator specified in the "Aggregator ID" field. When non-zero, the Port Number MUST be encoded with the Node ID as was discussed above.

- Aggregator ID

  Two octets, LACP Aggregator Identifier as specified in [IEEE-802.3] section 43.4.6

- Last Port Priority

  Two octets, LACP port priority for the corresponding interface, as specified in [IEEE-802.3] section 43.3.4. For local ports, this field encodes the previous operational value of port priority. For remote ports, this field encodes the operational port priority last known to the PE via notifications received from its peers in the RG.

- Current Port Priority

  Two octets, LACP port priority for the corresponding interface, as specified in [IEEE-802.3] section 43.3.4. For local ports, this field encodes the new operational value of port priority being advertised by the PE. For remote ports, this field specifies the new port priority being requested by the PE.

7.2.7. mLACP Port State TLV

The mLACP Port State TLV is used in the RG Application Data message. This TLV is used by a device to report its LACP port status to other members in the RG.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0035             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Partner System ID                        |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |     Partner System Priority   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Partner Port Number       |     Partner Port Priority     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Partner Key             | Partner State |  Actor State  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Actor Port Number        |           Actor Key           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Selected     |  Port State   |        Aggregator ID          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0035 for "mLACP Port State TLV"

- Length

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

- Partner System ID

  6 octets, the LACP Partner System ID for the corresponding interface, encoded as a MAC address as specified in [IEEE-802.3] section 43.4.2.2 item r.

- Partner System Priority

  2 octets field specifying the LACP Partner System Priority as specified in [IEEE-802.3] section 43.4.2.2 item q.

- Partner Port Number

  2 octets encoding the LACP Partner Port Number as specified in [IEEE-802.3] section 43.4.2.2 item u. The Port Number MUST be encoded with the Node ID as was discussed above.

- Partner Port Priority

  2 octets field encoding the LACP Partner Port Priority as specified in [IEEE-802.3] section 43.4.2.2 item t.

- Partner Key

  2 octets field representing the LACP Partner Key as defined in [IEEE-802.3] section 43.4.2.2 item s.

- Partner State

  1 octet field encoding the LACP Partner State Variable as defined in [IEEE-802.3] section 43.4.2.2 item v.

- Actor State

  1 octet encoding the LACP Actor's State Variable for the port as specified in [IEEE-802.3] section 43.4.2.2 item m.

- Actor Port Number

  2 octets field representing the LACP Actor Port Number as specified in [IEEE-802.3] section 43.3.4. The Port Number MUST be encoded with the Node ID as was discussed above.

- Actor Key

  2 octet field encoding the LACP Actor Operational Key as specified in [IEEE-802.3] section 43.3.5.

- Selected

  1 octet encoding the LACP 'Selected' variable, defined in [IEEE-802.3] section 43.4.8, as follows:

      0x00 SELECTED
      0x01 UNSELECTED
      0x02 STANDBY

- Port State

  1 octet encoding the operational state of the port as follows:
      0x00 Up
      0x01 Down
      0x02 Administrative Down
      0x03 Test (e.g. IEEE 802.3ah OAM Intrusive Loopback mode)

- Aggregator ID

  Two octets, LACP Aggregator Identifier to which this port is bound based on the outcome of the LACP selection logic.

7.2.8. mLACP Aggregator State TLV

The mLACP Aggregator State TLV is used in the RG Application Data message. This TLV is used by a device to report its Aggregator status to other members in the RG.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0037             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Partner System ID                        |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |     Partner System Priority   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Partner Key              |         Aggregator ID         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Actor Key                |   Agg State   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0037 for "mLACP Aggregator State TLV"

- Length

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

- Partner System ID

  6 octets, the LACP Partner System ID for the corresponding interface, encoded as a MAC address as specified in [IEEE-802.3] section 43.4.2.2 item r.

- Partner System Priority

  2 octets field specifying the LACP Partner System Priority as specified in [IEEE-802.3] section 43.4.2.2 item q.

- Partner Key

  2 octets field representing the LACP Partner Key as defined in [IEEE-802.3] section 43.4.2.2 item s.

- Aggregator ID

  Two octets, LACP Aggregator Identifier as specified in [IEEE-802.3] section 43.4.6

- Actor Key

  2 octet field encoding the LACP Actor Operational Key as specified in [IEEE-802.3] section 43.3.5.

- Agg State

  1 octet encoding the operational state of the Aggregator as follows:
      0x00 Up
      0x01 Down
      0x02 Administrative Down
      0x03 Test (e.g. IEEE 802.3ah OAM Intrusive Loopback mode)

7.2.9. mLACP Synchronization Request TLV

The mLACP 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. The following information can be requested:

- system configuration and/or state

- configuration and/or state for a specific port

- configuration and/or state for all ports with a specific LACP key

- configuration and/or state for all mLACP ports

- configuration and/or state for a specific aggregator

- configuration and/or state for all aggregators with a specific LACP key

- configuration and/or state for all mLACP aggregators

  The format of the TLV is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0038             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Request Number           |C|S|    Request Type           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Port Number / Aggregator ID  |             Actor Key         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0038 for "mLACP Synchronization Request TLV"

- Length

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

- Request Number

  2 octets. Unsigned integer uniquely identifying the request. Used to match the request with a response. The value of 0 is reserved for unsolicited synchronization, and MUST NOT be used in the mLACP 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 System Data
0x01   Request Aggregator Data
0x02   Request Port Data
0x3FFF Request All Data
```

- Port Number / Aggregator ID

  2 octets. When Request Type field is set to ’Request Port Data’, this field encodes the LACP Port Number for the requested port. When the Request Type field is set to ’Request Aggregator Data’, this field encodes the Aggregator ID of the requested Aggregator. When the value of this field is 0, it denotes that all ports (or Aggregators), whose LACP Key is specified in the "Actor Key" field, are being requested.

- Actor Key

  Two octets, LACP Actor key for the corresponding port or Aggregator. When the value of this field is 0 (and the Port Number/Aggregator ID field is 0 as well), it denotes that information for all ports or Aggregators in the system is being requested.

7.2.10. mLACP Synchronization Data TLV

The mLACP Synchronization Data TLV 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 being transmitted in response to an mLACP 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 mLACP Synchronization Data TLVs are also used for unsolicited advertisements of complete mLACP configuration and operational state data. The ’Request Number’ field MUST be set to 0 in this case. For such unsolicited synchronization, the PE MUST advertise all system, Aggregator and port information as done during the initialization sequence.

This TLV has the following format:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Type=0x0039             |    Length                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Request Number            |     Flags                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- U and F Bits

  Both are set to 0.

- Type

  set to 0x0039 for "mLACP Synchronization Data TLV"

- Length

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

- Request Number

  2 octets. Unsigned integer identifying the Request Number from the "mLACP Synchronization Request TLV" which solicited this synchronization data response.

- Flags

  2 octets, response flags encoded as follows:

      0x00 Synchronization Data Start
      0x01 Synchronization Data End

8. LDP Capability Negotiation

As requited in [RFC5561] the following TLV is defined to indicate the ICCP capability:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F| TLV Code Point=0x700      |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S| Reserved    |    Reserved   |  VER/Maj      |  Ver/Min      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

- U-bit

  SHOULD be 1 (ignore if not understood).

- F-bit

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

- TLV Code Point

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

- S-bit The State Bit indicates whether the sender is advertising or withdrawing the ICCP capability. The State bit is used as follows:

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

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

- Ver/Maj

  The major version revision of the ICCP protocol, this document specifies 1.0. This field is then set to 1

- Ver/Min

  The minor version revision of the ICCP protocol, this document specifies 1.0. This field is then set to 0

ICCP capability is advertised to a LDP peer if there is at least one RG enabled on the local PE.

9. Client Applications

9.1. Pseudowire Redundancy Application Procedures

This section defines the procedures for the Pseudowire Redundancy (PW-RED) Application.

9.1.1. Initial Setup

When an RG is configured on a system and multi-chassis pseudowire redundancy is enabled in that RG, the PW-RED application MUST send an "RG Connect" message with "PW-RED Connect TLV" to each PE that is a member of the same RG. The sending PE MUST set the A bit to 1 if it has already received a "PW-RED Connect TLV" from its peer; otherwise, the PE MUST set the A bit to 0. If a PE, that has sent the TLV with the A bit set to 0, receives a "PW-RED Connect TLV" from a peer, it MUST repeat its advertisement with the A bit set to 1. The PW-RED application connection is considered to be operational when both PEs have sent and received "PW-RED Connect TLVs" with the A bit set to 1. Once the application connection becomes operational, the two devices can start exchanging "RG Application Data" messages for the PW-RED application.

If a system receives an "RG Connect" message with "PW-RED Connect TLV" that has a differing Protocol Version, it must follow the procedures outlined in the "Application Versioning" section above.

When the PW-RED application is disabled on the device, or is unconfigured for the RG in question, the system MUST send an "RG Disconnect" message with "PW-RED Disconnect TLV".

9.1.2. Pseudowire Configuration Synchronization

A system MUST advertise its local PW configuration to other PEs that are members of the same RG. This allows the PEs to build a view of the redundant nodes and pseudowires that are protecting the same service instances. The advertisement MUST be initiated when the PW-RED application connection first comes up. To that end, the system sends "RG Application Data" messages with "PW-RED Config TLVs" as part of an unsolicited synchronization. A PE MUST use a pair of "PW-RED Synchronization Data TLVs" to delimit the set of TLVs that are being sent as part of this unsolicited advertisement.

In the case of a configuration change, a PE MUST re-advertise the most up to date information for the affected pseudowires.

As part of the configuration synchronization, a PE advertizes the ROID associated with the pseudowire. This is used to correlate the pseudowires that are protecting each other on different PEs. A PE also advertizes a priority value that is used to determine the precedence of a given pseudowire to assume the Active role in a redundant setup. Furthermore, a PE advertizes a Service Name that is global in the context of an RG and is used to identify which pseudowires belong to the same service. Finally, a PE also advertizes the pseudowire identifier as part of this synchronization.

9.1.3. Pseudowire Status Synchronization

The mechanism for synchronizing pseudowire state depends on whether or not an AC redundancy mechanism is in use. If an AC mechanism is in use, then on a given PE, the forwarding status of the PW (Active or Standby) is derived from the state of the associated AC(s). This simplifies the operation of the multi-chassis redundancy solution (Figure 1) and eliminates the possibility of deadlock conditions between the AC and PW redundancy mechanisms. The rules by which the PW state is derived from the AC state are as follows:

- VPWS

  For VPWS, there’s a single AC per service instance. If the AC is Active, then the PW status should be Active. If the AC is Standby, then the PW status should be Standby.

- VPLS

  For VPLS, there could be multiple ACs per service instance (i.e. VFI). If AT LEAST ONE AC is Active, then the PW status should be Active. If ALL ACs are Standby, then the PW status should be Standby.

In this case, the PW-RED application does not synchronize PW status across chassis, per se. Rather, the AC Redundancy application should synchronize AC status between chassis, in order to determine which AC (and subsequently which PE) is Active or Standby for a given service. When that is determined, each PE will then adjust its local PWs state according to the rules described above.

On the other hand, if an AC redundancy mechanism is not in use, then the PW-RED application is used to synchronize pseudowire state. This is done by sending the "PW-RED State TLV" whenever the pseudowire state changes on a PE. This includes changes to the local end as well as the remote end of the pseudowire.

A PE may request that its peer retransmit previously advertised PW-RED state. This is useful for instance when the PE is recovering from a soft failure. To request such retransmission, a PE MUST send a set of one or more "PW-RED Synchronization Request TLVs".

A PE MUST respond to a "PW-RED Synchronization Request TLV" by sending the requested data in a set of one or more PW-RED TLVs delimited by a pair of "PW-RED Synchronization Data TLVs". The TLVs comprising the response MUST be ordered such that the Synchronization Response TLV with the "Synchronization Data Start" flag precedes the various other PW-RED TLVs encoding the requested data. These, in turn, MUST precede the Synchronization Data TLV with the "Synchronization Data End" flag. It is worth noting that the response may span across multiple RG Application Data messages; however, the above TLV ordering MUST be retained across messages, and only a single pair of Synchronization Data TLVs must be used to delimit the response across all Application Data Messages.

A PE MAY re-advertise its PW-RED state in an unsolicited manner. This is done by sending the appropriate config and state TLVs delimited by a pair of "PW-RED Synchronization Data TLVs" and using a ’Request Number’ of 0.

While a PE has a pending synchronization request for a pseudowire or a service, it SHOULD silently ignore all TLVs for said pseudowire or service that are received prior to the synchronization response and which carry the same type of information being requested. This saves the system from the burden of updating state that will ultimately be overwritten by the synchronization response. Note that TLVs pertaining to other pseudowires or services are to continue to be processed per normal in the interim.

If a PE receives a synchronization request for a pseudowire or service that doesn’t exist or is not known to the PE, then it MUST trigger an unsolicited synchronization of all pseudowire information (i.e. replay the initialization sequence).

9.1.4. PE Node Failure

When a PE node detects that a remote PE, that is member of the same RG, has gone down, the local PE examines if it has redundant PWs for the affected services. If the local PE has the highest priority (after the failed PE) then it becomes the active node for the services in question, and subsequently activates its associated PWs.

9.2. Attachment Circuit Redundancy Application Procedures

9.2.1. Common AC Procedures

This section describes generic procedures for AC Redundancy applications, independent of the type of the AC (ATM, FR or Ethernet).

9.2.2. AC Failure

When the AC Redundancy mechanism on the Active PE detects a failure of the AC, it should send an ICCP Application Data message to inform the redundant PEs of the need to take over. The AC failures can be categorized into the following scenarios:

- Failure of CE interface connecting to PE

- Failure of CE uplink to PE

- Failure of PE interface connecting to CE

9.2.3. PE Node Failure

When a PE node detects that a remote PE, that is member of the same RG, has gone down, the local PE examines if it has redundant ACs for the affected services. If the local PE has the highest priority (after the failed PE) then it becomes the active node for the services in question, and subsequently activates its associated ACs.

9.2.4. PE Isolation

When a PE node detects that is has been isolated from the core network (i.e. all core facing interfaces/links are not operational), then it should instruct its AC Redundancy mechanism to change the status of any active ACs to Standby. The AC Redundancy application should then send ICCP Application Data messages in order to trigger failover to a standby PE.

9.2.5. Ethernet AC Procedures

9.2.6. Multi-chassis LACP (mLACP) Application Procedures

This section defines the procedures that are specific to the multi-chassis LACP (mLACP) application.

9.2.6.1. Initial Setup

When an RG is configured on a system and mLACP is enabled in that RG, the mLACP application MUST send an "RG Connect" message with "mLACP Connect TLV" to each PE that is member of the same RG. The sending PE MUST set the A bit to 1 in the said TLV if it has received a corresponding "mLACP Connect TLV" from its peer PE; otherwise, the sending PE MUST set the A bit to 0. If a PE receives an "mLACP Connect TLV" from its peer after sending the said TLV with the A bit set to 0, it MUST resend the TLV with the A bit set to 1. A system considers the mLACP application connection to be operational when it has sent and received "mLACP Connect TLVs" with the A bit set to 1. When the mLACP application connection between a pair of PEs is operational, the two devices can start exchanging "RG Application Data" messages for the mLACP application. This involves having each PE advertise its mLACP configuration and operational state in an unsolicited manner. A PE SHOULD subscribe to the following order when advertising its mLACP state upon initial application connection setup:

- Advertise system configuration
- Advertise Aggregator configuration
- Advertise port configuration
- Advertise Aggregator state
- Advertise port state

A PE MUST use a pair of "mLACP Synchronization Data TLVs" to delimit the entire set of TLVs that are being sent as part of this unsolicited advertisement.

If a system receives an "RG Connect" message with "mLACP Connect TLV" that has a differing Protocol Version, it MUST follow the procedures outlined in the "Application Versioning" section above.

After the mLACP application connection has been established, every PE MUST communicate its system level configuration to its peers via the use of "mLACP System Config TLV". This allows every PE to discover the Node ID and the locally configured System ID and System Priority values of its peers.

If a PE receives an "mLACP System Config TLV" from a remote peer advertising the same Node ID value as the local system, then the PE MUST respond with an "RG Notification Message" to NAK the "mLACP System Config TLV". The PE MUST suspend the mLACP application until a satisfactory "mLACP System Config TLV" is received from the peer. It SHOULD also raise an alarm to alert the operator. Furthermore, if a PE receives a NAK for an "mLACP System Config TLV" that it has advertised, the PE MUST suspend the mLACP application and SHOULD raise an alarm to alert the network operator of potential device mis-configuration.

If a PE receives an "mLACP System Config TLV" from a new peer advertising the same Node ID value as another existing peer with which the local system has an established mLACP Application connection, then the PE MUST respond to the new peer with an "RG

Notification Message" to NAK the "mLACP System Config TLV" and MUST ignore the offending TLV.

If the Node ID of a particular PE changes due to administrative configuration action, the PE MUST then inform its peers to purge the configuration of all previously advertised ports and/or aggregators, and MUST replay the initialization sequence by sending an unsolicited synchronization of: the system configuration, Aggregator configuration, port configuration, Aggregator state and port state.

It is necessary for all PEs in an RG to agree upon the System ID and System Priority values to be used ubiquitously. To achieve this, every PE MUST use the values for the two parameters that are supplied by the PE with the numerically lowest value (among RG members) of System Aggregation Priority. This guarantees that the PEs always agree on uniform values, which yield the highest System Priority.

When the mLACP application is disabled on the device, or is unconfigured for the RG in question, the system MUST send an "RG Disconnect" message with "mLACP Disconnect TLV".

9.2.6.2. mLACP Aggregator and Port Configuration

A system MUST synchronize the configuration of its mLACP enabled Aggregators and ports with other RG members. This is achieved via the use of "mLACP Aggregator Config TLVs" and "mLACP Port Config TLVs", respectively. An implementation MUST advertise the configuration of Aggregators prior to advertising the configuration of any of their associated member ports.

The PEs in an RG MUST all agree on the MAC address to be associated with a given Aggregator. It is possible to achieve this via consistent configuration on member PEs. However, in order to protect against possible misconfiguration, a system MUST use, for any given Aggregator, the MAC address supplied by the PE with the numerically lowest System Aggregation Priority in the RG.

A system that receives an "mLACP Aggregator Config TLV" with an ROID to Key association that is different from its local association MUST NAK the corresponding TLV and disable the Aggregator with the same ROID. Furthermore, it SHOULD raise an alarm to alert the operator. Similarly, a system that receives a NAK in response to a transmitted "mLACP Aggregator Config TLV" MUST disable the associated Aggregator and SHOULD raise an alarm to alert the network operator.

A system MAY enforce a restriction that all ports that are to be bundled together on a given PE share the same Port Priority value. If so, the system MUST advertise this common priority in the "mLACP Aggregator Config TLV" and assert the "Priority Set" flag in such TLV. Furthermore, the system in this case MUST NOT advertise individual Port Priority values in the associated "mLACP Port Config TLVs" (i.e. the "Priority Set" flag in these TLVs should be 0).

A system MAY support individual Port Priority values to be configured on ports that are to be bundled together on a PE. If so, the system MUST advertise the individual Port Priority values in the appropriate "mLACP Port Config TLVs", and MUST NOT assert the "Priority Set" flag in the corresponding "mLACP Aggregator Config TLV".

When the configurations of all ports for member links associated with a given Aggregator have been sent by a device, it asserts that fact by setting the "Synchronized" flag in the last port’s "mLACP Port Config TLV". If an Aggregator doesn’t have any candidate member ports configured, this is indicated by asserting the "Synchronized" flag in its "mLACP Aggregator Config TLV".

Furthermore, for a given port/Aggregator, an implementation MUST advertise the port/Aggregator configuration prior to advertising its state (via the "mLACP Port State TLV" or "mLACP Aggregator State TLV"). If a PE receives an "mLACP Port State TLV" or "mLACP Aggregator State TLV" for a port or Aggregator that it had not learned of before via an appropriate Port or Aggregator Config TLV, then the PE MUST request synchronization of the configuration and state of all mLACP ports as well as all mLACP Aggregators from its respective peer. If during a synchronization (solicited or unsolicited), a PE receives a State TLV for a port or Aggregator that it has not learned of before, then the PE MUST send a NAK for the offending TLV. The PE MUST NOT request re-synchronization in this case.

When mLACP is unconfigured on a port/Aggregator, a PE MUST send a "Port/Aggregator Config TLV" with the "Purge Configuration" flag asserted. This allows receiving PEs to purge any state maintained for the decommissioned port/Aggregator. If a PE receives a "Port/Aggregator Config TLV" with the "Purge Configuration" flag asserted, and the PE is not maintaining any state for that port/Aggregator, then it MUST silently discard the TLV.

9.2.6.3. mLACP Aggregator and Port Status Synchronization

PEs within an RG need to synchronize their state-machines for proper mLACP operation with a multi-homed device. This is achieved by having each system advertise its Aggregators and ports running state in "mLACP Aggregator State TLVs" and "mLACP Port State TLVs", respectively. Whenever any LACP parameter for an Aggregator or a port, whether on the Partner (i.e. multi-homed device) or the Actor (i.e. PE) side, is changed a system MUST transmit an updated TLV for the affected Aggregator and/or port. Moreover, when the administrative or operational state of an Aggregator or port changes, the system MUST transmit an updated Aggregator or port state TLV to its peers.

If a PE receives an Aggregator or port state TLV where the 'Actor Key' doesn't match what was previously received in a corresponding Aggregator or port config TLV, the PE MUST then request synchronization of the configuration and state of the affected Aggregator or port. If such a mismatch occurs between the config and state TLVs as part of a synchronization (solicited or unsolicited), then the PE MUST send a NAK for the state TLV. Furthermore, if a PE receives a port state TLV with the 'Aggregator ID' set to a value that doesn't map to some Aggregator that the PE had learned of via a previous Aggregator config TLV, then the PE MUST request synchronization of the configuration and state of all Aggregators and ports. If the above anomaly occurs during a synchronization, then the PE MUST send a NAK for the offending port state TLV.

A PE MAY request that its peer retransmit previously advertised state. This is useful for example when the PE is recovering from a soft failure and attempting to relearn state. To request such retransmissions, a PE MUST send a set of one or more "mLACP Synchronization Request TLVs".

A PE MUST respond to an "mLACP Synchronization Request TLV" by sending the requested data in a set of one or more mLACP TLVs delimited by a pair of "mLACP Synchronization Data TLVs". The TLVs comprising the response MUST be ordered in the RG Application Data message(s) such that the Synchronization Response TLV with the "Synchronization Data Start" flag precedes the various other mLACP TLVs encoding the requested data. These, in turn, MUST precede the Synchronization Data TLV with the "Synchronization Data End" flag. Note that the response may span across multiple RG Application Data messages, for example when MTU limits are exceeded; however, the above ordering MUST be retained across messages, and only a single pair of Synchronization Data TLVs MUST be used to delimit the response across all Application Data Messages.

A PE device MAY re-advertise its mLACP state in an unsolicited manner. This is done by sending the appropriate Config and State TLVs delimited by a pair of "mLACP Synchronization Data TLVs" and using a 'Request Number' of 0.

While a PE has a pending synchronization request for a system, Aggregator or port, it SHOULD silently ignore all TLVs for said system, Aggregator or port that are received prior to the synchronization response and which carry the same type of information being requested. This saves the system from the burden of updating state that will utlimately be overwritten by the synchronization response. Note that TLVs pertaining to other systems, Aggregators or ports are to continue to be processed per normal in this case.

If a PE receives a synchronization request for an Aggregator, port or Key that doesn’t exist or is not known to the PE, then it MUST trigger an unsolicited synchronization of all system, Aggregator and port information (i.e. replay the initialization sequence).

If a PE learns, as part of a synchronization operation from its peer, that the latter is advertising a Node ID value which is different from the value previously advertised, then the PE MUST purge all port/aggregator data previously learnt from that peer prior to the last synchronization.

9.2.6.4. Failure and Recovery

When a PE that is active for a multi-chassis link aggregation group encounters a fault, it SHOULD attempt to fail-over to a peer PE which hosts the same RO. To that effect, the faulty PE SHOULD lower its port priority (by using a larger numeric value) and advertise this change in the "mLACP Port Priority TLV". If the PE is not capable of lowering its own port priority any further, it SHOULD trigger a failover to the redundant PE by sending an "mLACP Port Priority TLV" in which it requests the redundant PE to raise the latter’s port priority to the maximum permitted in [IEEE802.3ad] (i.e. the smallest allowed numeric value) for the Aggregator in question. Furthermore, the PE SHOULD set its own port priority to the next smallest numeric value.

Upon recovery from a previous fault, a PE MAY reclaim active role for a multi-chassis link aggregation group if configured for revertive protection. Otherwise, the recovering PE may assume standby role when configured for non-revertive protection. In the revertive scenario, a PE SHOULD assume active role within the RG by sending an "mLACP Port Priority TLV" to the currently active PE, requesting that the latter change its port priority to a value that is lower (i.e. numerically larger) for the Aggregator in question.

If a system is operating in a mode where different ports of a bundle are configured with different Port Priorities, then the system MUST NOT advertise or request change of Port Priority values for aggregated ports collectively (i.e. by using a ’Port Number’ of 0 in the "mLACP Port Priority TLV"). This is to avoid ambiguity in the interpretation of the 'Last Port Priority' field.

If a PE receives an "mLACP Port Priority TLV" requesting a priority change for a port or Aggregator that is not local to the device, then the PE MUST re-advertise the local configuration of the system, as well as the configuration and state of all its mLACP ports and Aggregators.

If a PE receives an "mLACP Port Priority TLV" in which the remote system is advertising priority change for a port or Aggregator that the local PE had not learned of before via an appropriate Port or Aggregator Config TLV, then the PE MUST request synchronization of the configuration and state of all mLACP ports as well as all mLACP Aggregators from its respective peer.

10. Security Considerations

The security considerations described in [RFC5036] and [RFC4447] that apply to the base LDP specification, and to the PW LDP control protocol extensions apply to the capability mechanism described in this document.

The ICCP protocol is not intended to be applicable when the redundancy group spans PE in different administrative domains. Furthermore, implementations SHOULD provide a mechanism to select to which LDP peers the ICCP capability will be advertised, and from which LDP peers the ICCP messages will be accepted.

11. IANA Considerations

11.1. MESSAGE TYPE NAME SPACE

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

Message type	Description
0x0700	RG Connect Message
0x0701	RG Disconnect Message
0x0702	RG Notification Message
0x0703	RG Application Data Message

11.2. TLV TYPE NAME SPACE

This document use a new LDP TLV type, IANA already maintains a registry of name "TLV TYPE NAME SPACE" defined by [RFC5036]. The following value is suggested for assignment:

TLV Type	Description
0x700	ICCP capability TLV.
0x701	LDP TCP/IP Port TLV.

11.3. ICC RG Parameter Type Space

IANA needs to set up a registry of "ICC RG parameter type". These are 14-bit values. Parameter Type values 1 through 0x000F are specified in this document, Parameter Type values 0x0010 through 0x1FFF are to be assigned by IANA, using the "Expert Review" policy defined in [RFC5226]. Parameter Type values 0x2000 through 0x2FFF, 0x3FFF, and 0 are to be allocated using the IETF consensus policy defined in [RFC5226]. Parameter Type values 0x3000 through 0x3FFE are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in [RFC5226]. A Parameter Type description is required for any assignment from this registry. Additionally, for the vendor proprietary extensions range a citation of a person or company name is also required. A document reference should also be provided.

Initial ICC RG parameter type space value allocations are specified below:

Parameter Type	Description	Reference
0x0001	ICC Sender Name	[RFCxxxx]
0x0002	NAK TLV	[RFCxxxx]
0x0003	Requested Protocol Version TLV	[RFCxxxx]
0x0004	Disconnect Code TLV	[RFCxxxx]
0x0005	ICC RG ID TLV	[RFCxxxx]
0x0010	PW-RED Connect TLV	[RFCxxxx]
0x0011	PW-RED Disconnect TLV	[RFCxxxx]
0x0012	PW-RED Config TLV	[RFCxxxx]
0x0013	Service Name TLV	[RFCxxxx]
0x0014	PW ID TLV	[RFCxxxx]
0x0015	Generalized PW ID TLV	[RFCxxxx]
0x0016	PW-RED State TLV	[RFCxxxx]
0x0017	PW-RED Synchronization Request TLV	[RFCxxxx]
0x0018	PW-RED Synchronization Data TLV	[RFCxxxx]
0x0019	PW-RED Disconnect Cause TLV	[RFCxxxx]

0x0030	mLACP Connect TLV	[RFCxxxx]
0x0031	mLACP Disconnect TLV	[RFCxxxx]
0x0032	mLACP System Config TLV	[RFCxxxx]
0x0033	mLACP Port Config TLV	[RFCxxxx]
0x0034	mLACP Port Priority TLV	[RFCxxxx]
0x0035	mLACP Port State TLV	[RFCxxxx]
0x0036	mLACP Aggregator Config TLV	[RFCxxxx]
0x0037	mLACP Aggregator State TLV	[RFCxxxx]
0x0038	mLACP Synchronization Request TLV	[RFCxxxx]
0x0039	mLACP Synchronization Data TLV	[RFCxxxx]
0x003A	mLACP Disconnect Cause TLV	[RFCxxxx]

11.4. STATUS CODE NAME SPACE

This document use several new Status codes, IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by [RFC5036]. The following values is suggested for assignment: The "E" column is the required setting of the Status Code E-bit.

Range/Value	E	Description	Reference
0x00010001	0	Unknown ICCP RG	
0x00010002	0	ICCP Connection Count Exceeded	
0x00010003	0	ICCP Application Connection Count Exceeded	
0x00010004	0	ICCP Application not in RG	
0x00010005	0	Incompatible ICCP Protocol Version	
0x00010006	0	ICCP Rejected Message	
0x00010007	0	ICCP Administratively Disabled	
0x00010010	0	ICCP RG Removed	
0x00010011	0	ICCP Application Removed from RG	

12. Acknowledgments

The authors wish to acknowledge the important contributions of Dennis Cai, Neil McGill, Amir Maleki, Dan Biagini, Robert Leger, Sami Boutros, Neil Ketley and Mark Christopher Sains.

13. Normative References

[RFC5036] L. Andersson et al, "LDP Specification", RFC 5036, October 2007.

[RFC5561] "LDP Capabilities", RFC5561, July 2009.

[RFC4447] "Transport of Layer 2 Frames Over MPLS", Martini, L., et al., rfc4447 April 2006.

[IEEE-802.3] IEEE Std. 802.3-2005, "Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications", IEEE Computer Society, December 2005.

[RFC2863] K. McCloghrie, F. Kastenholz, "The Interfaces Group MIB", rfc2863, June 2000.

14. Informative References

[RFC5880] D. Katz, D. Ward, "Bidirectional Forwarding Detection", RFC5880, June 2010

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

15. Author’s Addresses

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

Samer Salam
Cisco Systems, Inc.
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1
Canada
e-mail: ssalam@cisco.com

Ali Sajassi
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
e-mail: sajassi@cisco.com

Matthew Bocci
Alcatel-Lucent
Grove House, Waltham Road Rd
White Waltham, Berks, UK. SL6 3TN
e-mail: matthew.bocci@alcatel-lucent.co.uk

Satoru Matsushima
Softbank Telecom
1-9-1, Higashi-Shinbashi, Minato-ku
Tokyo 105-7313, JAPAN
e-mail: satoru.matsushima@tm.softbank.co.jp

Thomas D. Nadeau
Huawei Technologies
2330 Central Expy
Santa Clara, CA 95050
USA
e-mail: thomas.nadeau@huawei.com

Expiration Date: April 2011

Internet Engineering Task Force
Internet Draft
Intended status: Standards Track
Expires: April 22, 2011

Luca Martini
George Swallow
Giles Heron
Cisco

Matthew Bocci
Alcatel-Lucent

October 22, 2010

Pseudowire Status for Static Pseudowires

draft-ietf-pwe3-static-pw-status-01.txt

Status of this Memo

Abstract

This document specifies a mechanism to signal Pseudowire (PW) status messages using an PW associated channel (ACh). Such a mechanism is suitable for use where no PW dynamic control plane exits, known as static PWs, or where a Terminating Provider Edge (T-PE) needs to send a PW status message directly to a far end T-PE. The mechanism allows PW OAM message mapping and PW redundancy to operate on static PWs.

Table of Contents

1. 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 [RFC2119].

2. Introduction

The default control plane for Pseudowire (PW) technology, as defined in [RFC4447], is based on LDP. However that document also describes a static provisioning mode without control plane. When a static PW is used, there is no method to transmit the status of the PW, or attachment circuit (AC) between the two PEs at each end of the PW. This document defines a method to transport the PW status codes defined in [RFC4447], sec 5.4.2, and [REDUNDANCY] in-band with the PW data using a generic associated channel [RFC5586].

3. 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

4. Applicability

The procedures described in this draft are intended for the case where PWs are statically configured. Where an LDP control plane exists, this MUST be used for signaling all PW status messages with the exception of those specified in [REDUNDANCY]. For [REDUNDANCY], the ’S-PE’ bypass mode described below MAY be used in the presence of an LDP control plane.

5. Pseudowire Status Operation

5.1. PW OAM Message

The PW status TLV as defined in [RFC4447] sec 5.4.2 is transported in a PW OAM message using the PW associated channel (ACH).

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 1|Version|   Reserved    | 0xZZ PW OAM Message           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          ACH TLV Header                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Refresh Timer         |  TLV Length   |A|   Flags     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                             TLVs                              ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: ACH PW OAM Message Packet Header.

The first 32 bits are the standard ACH header construct as defined in [RFC5586].

The first nibble (0001b) indicates the ACH instead of PW data. The version and the reserved values are both set to 0 as specified in [RFC4385].

The ACH TLV header is defined in [RFC5586] section 3.2, and contains the length of ACH TLVs. In this application the long word is set to 0 as there are no ACH TLVs.

The refresh timer is an unsigned integer and specifies refresh time in seconds with a range from 1 to 65535. The value 0 means that the refresh timer is set indefinitely, and the PW OAM message will never be refreshed, and will never timeout. This mode SHOULD NOT be used other then when specified in this document.

The TLV length field indicates the length of all PW OAM TLVs only.

The A flag bit is used to indicate an acknowledgment of the PW status TLV included. The rest of the flag bits are reserved and they must be set to 0 on transmit, and ignored upon receive. When the A bit is set, the refresh timer value is a requested timer value. PW OAM Message code point = 0xZZ. [ZZ to be assigned by IANA from the PW Associated Channel Type registry.]

TLV types for use in this message are allocated by IANA in the LDP registry named: "TLV TYPE NAME SPACE" .

5.2. Sending a PW Status Message

PW Status messages are indicated by sending in-band PW OAM messages for a particular PW containing the PW Status TLV defined in [RFC4447]. The PW Status TLV format is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Res|     PW Status (0x096A)    |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Status Code                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: PW Status Message Format.

The first 2 bits are reserved, and MUST be set to zero on transmit, and ignored on receive.

The PW Status TLV is prepended with an PW OAM message header and sent on the ACH of the PW to which the status update applies.

To clear a particular status indication, the PE needs to send a new PW OAM message containing a PW Status TLV with the corresponding bit cleared.

The procedures described in [SEGMENTED] that apply to an S-PE and PW using an LDP control plane also apply when sending PW status using the PE OAM channel. The OPTIONAL procedures using the S-PE TLV described in [SEGMENTED] can also be applied when sending PW status using the PE OAM channel.

The detailed message transmit, and receive procedures are specified in the next section. PW OAM Status Messages MUST NOT be used as a connectivity verification method.

5.3. PW OAM status message transmit and receive

Unlike the PW status procedures defined in [RFC4447] with this method there is no TCP/IP session, or session management. Therefore unlike in the TCP/IP case, where the message is sent only once, the PW OAM message containing the PW status TLV needs to be transmitted repeatedly to ensure reliable message delivery.

The PW OAM message containing a PW status TLV with a new status bit set, will be transmitted twice at an initial interval of one second. Subsequently the PW OAM message will be transmitted with an interval specified by the refresh timer value in the packet. Note that this value MAY be updated in the new PW OAM message packet, in which case the new refresh timer value becomes the new packet transmit interval.

The suggested default value for the refresh timer is 30 seconds.

When a PW OAM message containing a status TLV is received, a timer is started according to the refresh rate specified in the packet. If another non zero PW status message is not received within 3.5 times the specified timer value, the status condition will timeout in 3.5 times the last refresh timer value received, and the default status of zero is assumed on the PW. It is also a good practice to introduce some jitter in the delay between refresh transmissions, as long as the maximum jitter delay is within the prescribed maximum refresh time of 3.5 times the specified timer value for 3 consecutive refresh packets.

To clear a particular status fault the PE need only send an updated message with the corresponding bit cleared. If the PW status word is zero, the PW OAM message will be sent with the method described above, however it MUST be acknowledged with a packet with a timer value of zero. This will cause the PE sending the message to stop sending, and continue normal operation.

The message containing the clear status TLV is sent according to the same rules defined above.

5.3.1. Acknowledge of PW status

The PE receiving a PW OAM message containing a PW status message can acknowledge the PW status message by simply building an almost identical reply packet with the A bit set, and transmitting it on the PW ACH back to the source of the PW status message. The timer value set in the reply packet will then be used as the new transmit interval. If the sender PE of a PW status message receives an acknowledge for a particular message where the PW status TLV matches exactly the PW status TLV in the message that is currently being refreshed, the sender PE MUST use the new timer value received.

The suggested default value for the refresh timer value in the acknowledge packet is 600 seconds.

If the sender PE receives an acknowledge message that does not match the current active PW status message being sent, it simply ignores the acknowledgment packet.

If a PE that has a non zero status word for a particular PW, detect by any means that the peer PW has become unreachable, it will follow the standard procedures and consider that PW as having an additional status bit set. This would, normally trigger sending updates again, and canceling the acknowledge refresh timer state.

5.3.2. Applicable PW status Bits

In some situations it might not be useful or possible to transit a PW status message because the remote PE is not reachable. For example a PE that detects a local PSN TX fault condition, will be unable to transmit a PW OAM message with a PW status TLV reflecting that condition. The general rule is that a PE or S-PE should always attempt to send a PW status message.

5.4. MPLS Label Stack

With one exception, all PW OAM status messages are are sent to the adjacent PE across the PSN tunnel. in many cases the transmitting PE has no way to determine whether the adjacent PE is a S-PE, or a T-PE. This is a necessary behavior to preserve backward compatibility with PEs that do not understand MS-PWs. In the procedures described in this document there are two possible destinations for the PW OAM status messages: the adjacent PE, or the T-PE. Sending a PW status message directly to the T-PE is a enhanced method that is only applicable using PW OAM status messages sent in the PW ACH.

5.4.1. Label stack for a message destined to the next PE

A PE that needs to forward a PW OAM status message to the adjacent PE across the PSN tunnel, MUST set the PW label TTL field to 1. Furthermore if the control word is not in use on the particular PW, the PE MUST also place the GAL reserved label [RFC5586], below the PW label also with the TTL field set to 1.

5.4.2. Label stack for a message destined to the egress PE

This is also known as "S-PE bypass mode" see below. A T-PE that requires sending a PW OAM status message directly to the corresponding T-PE at the other end of the PW MUST set the TTL of the PW label to a value that is sufficient to reach the corresponding T-PE. This value will be greater then one, but will be set according to the local policy on the transmitting T-PE. Furthermore if the control word is not in use on the particular PW, the PE MUST also place the GAL reserved label [RFC5586], below the PW label with the TTL field set to 1.

5.5. S-PE bypass mode

S-PE bypass mode enables a T-PE to bypass all S-PEs that might be present along the MS-PW and to send a message directly to the remote T-PE. This is used for very fast message transmission in-band with the PW PDUs. This mode is OPTIONAL, and must be supported by both T-PEs to be enabled.

Note that this method MUST NOT be used to send messages which are permitted to originate at an S-PE, since otherwise race conditions could occur between messages sent via the control plane by S-PEs, and messages sent via the data plane by T-PEs.

Currently the only PW status codes which MAY be sent using the S-PE bypass procedure are:

0x00000002 - Local Attachment Circuit (ingress) Receive Fault
0x00000004 - Local Attachment Circuit (egress) Transmit Fault

Note that since "clear all failures" may be sent by an S-PE it MUST NOT be sent using the S-PE bypass mode.

When S-PE bypass mode is enabled, all PW Status TLVs received using this method have priority over PW Status TLVs sent via control protocols such as LDP [RFC4447].

5.5.1. S-PE bypass mode LDP flag bit

When a PW Segment along an MS-PW is using the LDP control protocol, a flag bit MUST be set in the interface parameters sub-TLV to indicate that the T-PE is requesting S-PE bypass status message mode. If the S-PE bypass mode LDP flag bit in the generic protocol flags interface parameter does not mach in the FEC advertisement for directions of a specific PW, that PW MUST NOT be enabled.

The interface parameter is defined as follow:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Type=0X16  |   Length=4    |R R R R R R R R R R R R R R R B|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Generic Protocol Flags.

- TLV Type.

  Type 0x16 - Generic Protocol Flags. Note: Value 0x16 suggested for assignment pending IANA allocation.

- Length

  TLV length always 4 octets.

- Flags

  Protocol flags, Bit B is set to request the S-PE bypass mode. Bits R are reserved for future use, and must be zero on transmission, and ignored on reception of this TLV.

5.5.2. S-PE bypass mode negotiation procedure

To be written in the next revision.

6. S-PE operation

The S-PE will operate according to the procedures defined in [SEGMENTED]. The following additional procedures apply to the case where a static PW segment is switched to a dynamic PW segment that uses LDP, and the case a static PW segment is switched to another static PW segment.

6.1. Static PW to another Static PW

The procedures that are described in [SEGMENTED] section 10 also apply to the case of a static PW switched to another static PW. The LDP header is simply replaced by the PE OAM header, otherwise the packet format will be identical. The information that is necessary to form a SP-PE TLV MUST be configured in the S-PE, or no S-PE TLV will be sent. The Document [SEGMENTED] defines a IANA registry named "Pseudowire Switching Point PE TLV Type". In order to support the static PW configuration and addressing scheme, a new code point is requested as follows:

Type	Length	Description
0x07	24	Static PW/MPLS-TP PW segment ID of last PW segment traversed

The format of this TLV is that of the "Static Pseudowire Sub-TLV" defined in [ON DEMAND].

6.2. Dynamic PW to Static PW or vice versa

The procedures that are described in [SEGMENTED] section 10 also apply to this situation. However if the PW label of the LDP controlled PW segment is withdrawn, by the adjacent PE, the S-PE will set the PW status code "0x00000001 - Pseudowire Not Forwarding" to the adjacent PW on the static PW segment.

The S-PE will only withdraw its label for the dynamic, LDP controlled, PW segment if the S-PE is un-provisioned.

7. Security Considerations

The security measures described in [RFC4447] and [SEGMENTED] are adequate for the proposed mechanism.

8. IANA Considerations

This document uses a new Associated Channel Type. IANA already maintains a registry of name "Pseudowire Associated Channel Types". A value of 0x0022 is suggested for assignment with TLVs. The description is "PW OAM Message".

This document uses a new Pseudowire Switching Point PE TLV Type. IANA already maintains a registry of name "Pseudowire Switching Point PE

TLV Type". A value of 0x07 is suggested for assignment. The description is "Static PW/MPLS-TP PW segment ID of last PW segment traversed".

This document uses a new interface parameter type. IANA already maintains a registry of name "Pseudowire Interface Parameters Sub-TLV type Registry". A value of 0x16 is suggested for assignment. The description is "Generic Protocol Flags".

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.

[SEGMENTED] Martini et.al. "Segmented Pseudo Wire", draft-ietf-pwe3-segmented-pw-18.txt, IETF Work in Progress, September 2010.

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

[REDUNDANCY] Muley et.al. "Preferential Forwarding Status bit definition", draft-ietf-pwe3-redundancy-bit-03.txt, IETF Work in Progress, May 2010.

[ON DEMAND] Bahadur et.al. "MPLS on-demand Connectivity Verification, Route Tracing and Adjacency Verification", draft-ietf-mpls-tp-on-demand-cv-01.txt, IETF Work in Progress, October 2010

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

Giles Heron
Cisco Systems
9-11 New Square
Bedfont Lakes
Feltham
Middlesex
TW14 8HA
United Kingdom
e-mail: giheron@cisco.com

Matthew Bocci
Alcatel-Lucent
Grove House, Waltham Road Rd
White Waltham, Berks, UK. SL6 3TN
e-mail: matthew.bocci@alcatel-lucent.co.uk

Copyright Notice

Expiration Date: April 2011

Network Working Group
Internet-Draft
Updates: 4447
Category: Standards Track
Expires: April 22, 2011

Lizhong Jin (ed.), ZTE
Raymond Key (ed.), Telstra
Simon Delord
Thomas Nadeau, Huawei
Vishwas Manral, IPInfusion

October 22, 2010

Pseudowire Control Word Negotiation Mechanism Analysis and Update
draft-jin-pwe3-cbit-negotiation-01.txt

Abstract

This draft describes the problem of control word negotiation mechanism specified in [RFC4447]. Based on the problem analysis, possible solutions and their potential shortcomings are also discussed.

Status of this Memo

Copyright Notice

Table of Contents

Conventions used in this document

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

1. Introduction

This draft describes the problem of control word negotiation mechanism specified in [RFC4447].

Based on the problem analysis, possible solutions and their potential shortcomings are also discussed.

2. Problem Statement

[RFC4447] section 6 describes the control word negotiation mechanism. Each PW endpoint has the capability of being configurable with a parameter that specifies whether the use of the control word is PREFERRED or NOT PREFERRED.

This negotiation mechanism will not work properly in the following case:

```
+-------+                    +-------+
|       |         PW         |       |
|  PE1  |====================|  PE2  |
|       |                    |       |
+-------+                    +-------+
```

Figure 1

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

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

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

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

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

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

3. Possible Solutions

The solution for this problem should be applicable to both SS-PW and MS-PW.

In this draft, possible solutions are discussed.

3.1. Option 1: Control Word Re-Negotiation by Label Request

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

The behavior of PE1 and PE2 should be as follows:

1. PE2 changes locally configured control word to PREFERRED.

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

3. When PE doing the CW changing operation, PE2 needs to send label request to PE1 although it already received the label mapping.

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

5. PE1 will send label mapping message with Cbit=1 again to PE2 (Note: PE1 SHOULD send label mapping with locally configured CW parameter).

6. PE2 receives the label mapping from PE1 and updates the remote label binding information.

7. PE2 will send label mapping to PE1 with CW=1.

It should be noted that the request message should be processed in ordered mode in MS-PW case. When S-PE receives a label request message from a remote, it should advertise the request message to the other remote PE. This is necessary since S-PE does not have full information of interface parameter field in the FEC advertisement.

By sending label request message, PE2 will get the configured CW parameter from peer PE1 from receiving label mapping message. By using the new CW parameter from label mapping message sent by peer PE1 and locally configured CW, PE2 will determine the control word parameter according to [RFC4447] section 6.

3.2. Option 2: Make CW Non-Configurable

The second solution is to change the control word to be not configurable, and default value is PREFERRED which can be degraded to NOT PREFERRED by negotiation automatically. The negotiation mechanism can still follow the procedure described in [RFC4447] section 6.

There is explicit requirement from some service providers to allow control word to be configurable. This option will not fulfill their need.

3.3. Option 3: Manual Configuration Process for CW

The third solution is to abandon the control word negotiation mechanism described in [RFC4447], and use a new simple mechanism.

When receiving the CW bit from peer PE, local PE should simply compare the control word with local configuration (PREFERRED or not-PREFERRED). Only when the control word configured on both end-points of PW is PREFERRED, the PW will be UP with CW = 1, otherwise the PW will be UP with CW = 0 and the node with CW PREFERRED will automatically degrade to CW not-PREFERRED.

It is important to note that this control word negotiation mechanism is not interoperable with the old mechanism defined in [RFC4447].

3.4. Option 4: Make CW Capability Mandatory

This option is to make CW capability mandatory. The PW will only be in operation UP when both PW end-points support control word capability.

We should consider some side effect while making CW capability mandatory, which will be analyzed in future.

3.5. Extra Considerations

The possible CW negotiation for multi-segment PW as well as potential complications with FEC129 will be covered in later version of this document.

Backward compatibility issues will be further discussed in later version of this document.

4. Security Considerations

This will be added in later version of this document.

5. IANA Considerations

This will be no IANA request for this document.

6. Acknowledgements

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

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

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

Authors' Addresses

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

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

Simon Delord
Email: simon.delord@gmail.com

Thomas Nadeau
Huawei
Email: Thomas.Nadeau@huawei.com

Vishwas Manral
IPInfusion
Email: vishwas@ipinfusion.com

PWE3 Working Group
Internet-Draft
Intended Status: Standards Track
Expires: April 2011
S. Kini
D. Sinicrope
Ericsson
October 18, 2010

Pseudowire Virtual Circuit Connectivity Verification (VCCV): An Inband Control Channel using offset

draft-kini-pwe3-inband-cc-offset-00.txt

Status of this Memo

Copyright Notice

Abstract

Pseudowires need an inband control channel (CC) to do VCCV such that OAM and data packets follow the same path. However most PW deployments are without a Control Word (CW) and hence are unable to use the inband CC as defined in RFC5085. This document defines a simple extension to the TTL expiry CC (Type 3) to do inband VCCV. This can be used even without a CW.

Table of Contents

1. Introduction

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

2. Conventions used in this document

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

3. Problem Statement

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

4. Solution

This document defines a new VCCV CC. It is an extension of the TTL Expiry VCCV (Type 3) defined in [VCCV]. In this CC the associated channel starts at a fixed offset after the PW label. This CC is henceforth referred to as Inband-offset VCCV (Type TBA). A fixed number of bytes between the PW label and the start of the associated channel can be used to emulate flow header information and are henceforth referred to as a "pseudo flow header". A VCCV message with a pseudo flow header will follow the same path as that taken by a data packet of the flow, as long as any multipath forwarding decision taken by the intermediate LSRs do not look beyond the pseudo flow header. A pseudo flow header length of 64 bytes is expected to meet the requirements of all current implementations and also meet the requirements of deployments (both current and in the foreseeable future). If a size other than 64 is needed then it can be configured or signaled as an attribute of the PW. The content of the pseudo flow header is set according to the flow that needs an OAM function such as connectivity verification (CV). E.g. if the encapsulation consists of an IP packet following the PW label, then the pseudo flow header would be the IP header of a flow.

5. Security Considerations

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

6. IANA Considerations

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

7. Future work

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

8. References

8.1. Normative References

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

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

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

[ID.PPW-EIM] Kini, S., et al, "Encapsulation Methods for Transport of packets over an MPLS PSN - efficient for IP/MPLS", draft-kini-pwe3-pkt-encap-efficient-ip-mpls-00 (work in progress), July 2010.

8.2. Informative References

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

Authors' Addresses

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

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

MPLS Working Group S. Kini
Internet-Draft D. Sinicrope
Intended Status: Standards Track Ericsson
Expires: April 2011 October 25, 2010

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

draft-kini-pwe3-pkt-encap-efficient-ip-mpls-01.txt

Status of this Memo

Copyright Notice

Abstract

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

Table of Contents

1. Introduction

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

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

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

2. Conventions used in this document

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

3. Scope

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

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

4. Network Reference Model

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

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

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

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

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

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

5. Solution

This solution does not use a data link layer header (such as Ethernet) on the PPW to transport IP/MPLS packets. This reduces the overhead bytes for such packets. There are implementations that look beyond the MPLS label stack for an IP packet. For non IP/MPLS packets, whenever there is a potential for such a condition, an IP encapsulation (with GRE) is used. Thus ECMP based on looking for an IP packet beyond the MPLS stack will work correctly and not re-order any flows. To prevent the GRE encapsulated packets from having IP address conflicts with the IP address space of the customer’s network, a non-routable IP address (in the 127/8 range) is used. The details of the packet encapsulation are in section 5.1. The adaptation of PE-bound and CE-bound traffic is explained in section 5.2.

5.1. Encapsulation format on the PPW

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

5.1.1. IP packets

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

```
+-------------------------------------------------+
|PSN Tunnel & PSN Physical Headers                | m octets
|-------------------------------------------------|
|PW Label (S=0 if FAT-PW label present, else S=1)| 4
|-------------------------------------------------|
|Optional FAT-PW label  S=1                       | 4
|-------------------------------------------------|
|IP v4/v6 packet                                  | n octets
|                                                 |
+-------------------------------------------------+
```

Figure 1 IPv4/v6 packet encapsulated into PPW without CW

```
+-------------------------------------------------+
|PSN Tunnel & PSN Physical Headers                | m octets
|-------------------------------------------------|
|PW Label (S=0 if FAT-PW label present, else S=1)| 4
|-------------------------------------------------|
|Optional FAT-PW label  S=1                       | 4
|-------------------------------------------------|
|Control Word with Flags bits 6,7 set to 01       | 4
|-------------------------------------------------|
|IP v4/v6 packet                                  | n octets
|                                                 |
+-------------------------------------------------+
```

Figure 2 IPv4/v6 packet encapsulated into PPW with CW

5.1.2. MPLS packet

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

```
+-------------------------------------------------+
|PSN Tunnel & PSN Physical Headers                | m octets
|-------------------------------------------------|
|PW Label (S=0 if FAT-PW label present, else S=1)| 4
|-------------------------------------------------|
|Optional FAT-PW label S=1                        | 4
|-------------------------------------------------|
|Control Word with Flags bits 6,7 set to 10       | 4
|-------------------------------------------------|
|MPLS Packet                                      | n octets
|                                                 |
+-------------------------------------------------+
```

Figure 3 MPLS packet encapsulated into PPW with CW

```
+-------------------------------------------------+
|PSN Tunnel & PSN Physical Headers                | m octets
|-------------------------------------------------|
|PW Label S=0                                     | 4
|-------------------------------------------------|
|Optional FAT-PW label S=0                        | 4
|-------------------------------------------------|
|MPLS Packet                                      | n octets
|                                                 |
+-------------------------------------------------+
```

Figure 4 MPLS packet encapsulated into PPW without CW

5.1.3. An arbitrary protocol

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

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

```
   +-------------------------------------------------+
   |PSN Tunnel & PSN Physical Headers                | m octets
   |-------------------------------------------------|
   |PW Label (S=0 if FAT-PW label present, else S=1)| 4
   |-------------------------------------------------|
   |Optional FAT-PW label S=1                        | 4
   |-------------------------------------------------|
   |IPv4 header (GRE Delivery header)                | 20
   |  IPv4 protocol field=47(GRE)                    |
   |  TTL=1                                          |
   |                                                 |
   |  Dst Addr 127/8                                 |
   |-------------------------------------------------|
   | GRE Header                                      | 8
   |                                                 |
   +-------------------------------------------------+
   | GRE Payload Packet - any arbitrary protocol     | n octets
   |                                                 |
   +-------------------------------------------------+
```

Figure 5 An arbitrary protocol packet encapsulated into PPW without CW

```
   +-------------------------------------------------+
   |PSN Tunnel & PSN Physical Headers                | m octets
   |-------------------------------------------------|
   |PW Label (S=0 if FAT-PW label present, else S=1)| 4
   |-------------------------------------------------|
   |Optional FAT-PW label S=1                        | 4
   |-------------------------------------------------|
   |Control Word with Flags bits 6,7 set to 00       | 4
   |-------------------------------------------------|
   |Ethernet frame of an arbitrary protocol          | n octets
   |                                                 |
   +-------------------------------------------------+
```

Figure 6 An arbitrary protocol packet encapsulated into PPW with CW

5.2. Traffic adaptation

5.2.1. PE-bound

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

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

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

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

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

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

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

5.2.2. CE-bound

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

      ii. If the destination IPv6 address is broadcast/multicast then the destination MAC address of the Ethernet frame is determined accordingly. Else if the destination IPv6 address is unicast then PPW-EIM-DMAC is used.

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

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

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

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

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

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

5.3. QoS considerations

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

5.4. PW Types

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.5. Control Word

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

5.5.1. Characteristics without CW

PPW-EIM (without CW) is not susceptible to re-ordering flows within the PPW. It can also take advantage of ECMP implementations that examine the first nibble after the MPLS label stack to determine whether the labeled packet is an IP packet. Such implementations are widely available today and will correctly identify the IP flow in the PPW. Even the flows of non IP/MPLS protocols will not be re-ordered as long as the same source and destination IP addresses are used in the GRE Delivery header for the life of the PPW. Hence a CW is not necessary for PPW-EIM to prevent flow re-ordering. This can also obviate the need for [FAT-PW] within PPW-EIM and thereby save on processing power at ingress to identify the flow (through packet classification) and add the flow-label. When an ECMP based on the label stack is required (and available), then [FAT-PW] must be used with PPW-EIM. An important benefit of not adding a CW and/or flow-label is that the difference in packet size between the access network and the PSN is further reduced by up to 8 bytes (compared with [PWE3-ETH]) and hence there is less chance for fragmentation of jumbo IP/MPLS packets.

5.5.2. PPW-EIM-CW

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

- 00 indicates payload is any protocol encapsulated in an Ethernet frame

- 01 indicates payload is IP

- 10 indicates payload is MPLS

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

5.6. Signaling extensions

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

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

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|   Local MAC addr (TBA)    |            Length=6           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Local MAC address                      |
|                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

U bit: Unknown bit. This bit MUST be set to 1. If the MAC address format is not understood, then the TLV is not understood and MUST be ignored.

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

5.7. Implementation considerations

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

6. PSN MTU requirements

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

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

7. Security Considerations

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

8. IANA Considerations

IANA needs to allocate values for the following:

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

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

9. Conclusion

PPW-EIM has the following useful advantages:

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

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

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

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

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

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

inspection/flow-based applications (including ECMP).

10. References

10.1. Normative References

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

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

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

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

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

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

10.2. Informative References

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

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

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

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

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

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

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

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

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

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

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

11. Acknowledgments

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

Appendix A: Example

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

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

```
        +-------+                                     +-------+
        |       |   AC   +---+  PSN   +---+  AC    |       |
        |  R1   |-------|PE1|-------|PE2|-------|  R2   |
        |       |   E    +---+   L    +---+   E    |       |
        +-------+                                     +-------+

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

Figure 7 Router inter-connect using PPW

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

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

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

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

```
+-------------------------------------------------+
| Dest MAC addr AllL2ISs 01-80-C2-00-00-14        |    4
|                       +-------------------------|
|                       |                         |    4
+-----------------------+                         |
| Src MAC addr     00-01-02-03-04-05              |    4
+-----------------------+-------------------------|
| TPID=0x8100           | VID=0x5 PCP=111 CFI=0   |    4
+-----------------------+-------------------------|
| Length= n+3           | LLC = 0xFE 0xFE         |    4
+-----------+-------------------------------------+
| SNAP=0x03 | NLPID=0x83|                         |    4
+-----------+-----------+                         |
| ISIS L2 LAN-IIH                                 |    n-3 octets
|                                                 |
+-------------------------------------------------+
```

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

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

```
+-------------------------------------------------+
|PSN Physical layer headers                       |    m octets
+-------------------------------------------------+
|PW Demultiplexer Label=0x4321 S=1 TC=0x7         |    4
|-------------------------------------------------|
|IPv4 header (GRE Delivery header)                |    20
| IPv4 protocol field=47(GRE)                     |
| TTL=0, Checksum=<computed>                      |
| Src Addr 127.0.0.1                              |
| Dst Addr 127.0.0.1                              |
|-------------------------------------------------|
| GRE Header            Protocol Type=0x8100      |    8
| Checksum=<computed>                             |
+-------------------------------------------------+
| GRE Payload Packet - frame E                    |    n+22 octets
|                                                 |
+-------------------------------------------------+
```

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

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

```
+-------------------------------------------------+
| Dest MAC addr    10-11-12-13-14-15              |     4
|                        +------------------------|
|                        |                        |     4
+------------------------+                        |
| Src MAC addr     00-01-02-03-04-05              |     4
+------------------------+------------------------|
| TPID=0x8100            | VID=0x5 PCP=000 CFI=0  |     4
+------------------------+------------------------|
| EtherType=0x800        |                        |     4
+------------------------+                        |
| IP packet                                       |     n-2 octets
|                                                 |
+-------------------------------------------------+
```

Figure 10 IP packet from R1 to R2 on AC

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

```
+-------------------------------------------------+
|PSN Physical layer headers                       | m octets
+-------------------------------------------------+
|PW Demultiplexer Label=0x4321 S=1 TC=0x0         | 4
+-------------------------------------------------+
| IP packet                                       | n octets
|                                                 |
+-------------------------------------------------+
```

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

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

```
         R1                                        R2
      +-------+                                +-------+
      |CE1|   |                                |   |CE2|
      +---|   |             +---+              |   |---|
      |   |PE1|-------------| P |--------------|PE2|   |
      |   .   |     L1      +---+      L2      |   .   |
      |   .   |                                |   .   |
      +-------+                                +-------+

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

Figure 12 CE interconnect when co-existing with PE

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

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

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

```
+-------------------------------------------------+
|PSN Tunnel and Physical layer headers            |    m octets
+-------------------------------------------------+
|PW Demultiplexer Label=0x4321 S=1 TC=0x7         |    4
|-------------------------------------------------|
|IPv4 header (GRE Delivery header)                |    20
| IPv4 protocol field=47(GRE)                     |
| TTL=1, Checksum=<computed>                      |
| Src Addr 127.0.0.1                              |
| Dst Addr 127.0.0.1                              |
|-------------------------------------------------|
| GRE Header          Protocol Type=Length=n      |    8
| Checksum=<computed>                             |
+-------------------------------------------------+
| GRE Payload Packet - IIH                        |    n octets
|                                                 |
+-------------------------------------------------+
```

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

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

```
+-------------------------------------------------+
|PSN Tunnel and Physical layer headers            | m octets
+-------------------------------------------------+
|PW Demultiplexer Label=0x4321 S=1 TC=0x0         |4
+-------------------------------------------------+
| IP packet                                       | n octets
|                                                 |
+-------------------------------------------------+
```

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

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

```
+-------------------------------------------------+
|PSN Tunnel and Physical layer headers            | m octets
+-------------------------------------------------+
|PW Demultiplexer Label=0x4321 S=0 TC=0x0         |4
+-------------------------------------------------+
| MPLS packet                                     | n octets
|                                                 |
+-------------------------------------------------+
```

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

Authors' Addresses

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

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

Network Working Group
Internet Draft
Updates (if published): RFC5586
Intended status: Standards Track

Han Li
China Mobile

Luca Martini
Cisco System

Jia He
Huawei

Feng Huang
Alcatel-Lucent

Expires: March 2011 September 14, 2010

Using the Generic Associated Channel Label for Pseudowire in MPLS-TP
draft-lm-pwe3-mpls-tp-gal-in-pw-00.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on March 14, 2011.

Copyright Notice

Abstract

This document describes the requirements for using the Generic Associated Channel Label (GAL) in Pseudowires (PWs) in MPLS-TP networks, and provides an update to the description of GAL usage in [RFC5586] by removing the restriction that is imposed on using GAL for PWs especially in MPLS-TP environments.

.

Table of Contents

1. Introduction

[RFC5586] generalizes the associated control channel mechanism of [RFC5085] to be used for Sections, Label Switched Paths (LSPs), and Pseudowires (PWs) in MPLS networks. [RFC5085] defines the Associated Channel Header (ACH), and [RFC5586] generalizes this for use in the Generic Associated Channel (G-ACh).

[RFC5586] defines a generalized label-based exception mechanism using the Generic Associated Channel Label (GAL) to work together with the ACH for use with LSPs but places restrictions on GAL usage with PWs.

This document removes the restriction imposed by [RFC5586].

2. Conventions used in this document

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

2.1. Terminology

ACH Associated Channel Header

CW Control Word

G-ACh Generic Associated Channel

GAL G-ACh Label

MPLS-TP MPLS Transport Profile

OAM Operation, Administration, and Maintenance

3. GAL Usage for MPLS-TP PW

According to the MPLS-TP requirement document [RFC5654], it is necessary that MPLS-TP mechanisms and capabilities be able to interoperate with the existing IETF MPLS [RFC3031] and IETF PWE3 [RFC3985] architectures appropriate. [RFC5586] differentiates between the usage of the GAL with PWs in MPLS and MPLS-TP environments in section 4.2 as follows:

> In MPLS-TP, the GAL MUST be used with packets on a G-ACh on LSPs, Concatenated Segments of LSPs, and with Sections, and MUST NOT be used with PWs.

This indicates that the GAL can be used for MPLS-TP LSPs and Sections, but not for PWs using an MPLS-TP PSN.

However, there is no restriction imposed on the usage of the GAL in MPLS PWs, which is described immediately afterwards in the same section of [RFC5586] (Section 4.2):

> However, in other MPLS environments, this document places no restrictions on where the GAL may appear within the label stack or its use with PWs.

The inconsistency between the usage of the GAL with MPLS PWs and MPLS-TP PWs may cause unnecessary implementation differences and is in disagreement with the MPLS-TP requirements.

Therefore, this document specifies that the GAL can be used with packets on a G-ACh on LSPs, Concatenated Segments of LSPs, Sections, and PWs in both MPLS and MPLS-TP environments without discrimination.

[RFC5586] is updated by removing the restrictions on using GAL for PW as follows:

- Section 1 (Introduction) in [RFC5586], the original text:

  > The GAL mechanism is defined to work together with the ACH for LSPs and MPLS Sections.

  is replaced by:

  > The GAL mechanism is defined to work together with the ACH for LSPs and MPLS Sections, and for PWs.

- Section 4.2. (GAL Applicability and Usage) in [RFC5586], the original text:

  > In MPLS-TP, the GAL MUST be used with packets on a G-ACh on LSPs, Concatenated Segments of LSPs, and with Sections, and MUST NOT be used with PWs. It MUST always be at the bottom of the label stack (i.e., S bit set to 1). However, in other MPLS environments, this document places no restrictions on where the GAL may appear within the label stack or its use with PWs.

  is replaced by:

  > In MPLS-TP, the GAL MUST be used with packets on a G-ACh on LSPs, Concatenated Segments of LSPs, and with Sections, and MAY be used with PWs. It MUST always be at the bottom of the label stack (i.e., S bit set to 1). However, in other MPLS environments, this document places no restrictions on where the GAL may appear within the label stack.

4. Security Considerations

No further security considerations than [RFC5586].

5. IANA Considerations

There are no IANA actions required.

6. Acknowledgments

The authors would like to thank Luyuan Fang, Adrian Farrel, Haiyan Zhang, Guanghui Sun, Italo Busi, Matthew Bocci for their contributions to this work.

The authors would also like to thank the authors of [RFC5586] and people who were involved in the development of [RFC5586].

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

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

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

7.2. Informative References

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

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

8. Authors' Addresses

Han Li
China Mobile Communications Corporation
Email: lihan@chinamobile.com

Luca Martini
Cisco Systems, Inc.
Email: lmartini@cisco.com

Jia He
Huawei Technologies Co., Ltd.
Email: hejia@huawei.com

Feng Huang
Alcatel-Lucent shanghai Bell
Email: feng.f.huang@alcatel-sbell.com.cn

Intellectual Property

MPLS Working Group F. Zhang, Ed.
Internet-Draft B. Wu, Ed.
Intended status: Standards Track ZTE Corporation
Expires: April 28, 2011 E. Bellagamba, Ed.
A. Takacs
Ericsson
X. Dai
M. Xiao
ZTE Corporation
October 25, 2010

LDP Extensions for Proactive OAM Configuration of Dynamic MPLS-TP PW
draft-zhang-mpls-tp-pw-oam-config-03

Abstract

This document specifies extensions to the LDP protocol to configure and control proactive OAM functions, suitable for dynamic SS-PW and MS-PW.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

MPLS PWs are defined in [RFC3985] and [RFC5659], which provide for emulated services over an MPLS Packet Switched Network (PSN). MPLS Transport Profile (MPLS-TP) describes a profile of MPLS that enables operational models typical in transport networks, while providing additional OAM, survivability and other maintenance functions not previously supported by IP/MPLS, including PW. The corresponding requirements are defined in [I-D.ietf-mpls-tp-oam-requirements].

[I-D.ietf-mpls-tp-oam-framework] describes how MPLS-TP OAM mechanisms are operated to meet transport requirements, categorized into proactive and on-demand monitoring. Proactive monitoring is typically configured at transport path creation time, either be carried out periodically and continuously or act on certain events such as alarm signals. In contract on-demand monitoring is initiated manually and for a limited amount of time, usually for operations such as diagnostics to investigate into a defect condition.

NMS or LSP Ping [I-D.absw-mpls-lsp-ping-mpls-tp-oam-conf] are used to configure these OAM functionalities if a control plane is not instantiated. But if the control plane is used, it must support to the configuration and modification of OAM maintenance points as well as the activation/deactivation of OAM when the transport path or transport service is established or modified [RFC5654].

This document specifies extensions to the LDP protocol to negotiate PW OAM capabilities, configure and bootstrap proactive PW OAM functions, suitable for SS-PW and MS-PW. Configuration of OAM entities for MS-PW SPME will be added in the future, and P2MP PW is out of the scope of this document.

1.1. Analysis of existing PW OAM Configuration

1.1.1. MPLS PW OAM Functions

Before MPLS-TP standards, PW OAM functions are implemented by [RFC5085], [RFC5885], [RFC4447] and [I-D.ietf-pwe3-static-pw-status]. [RFC5085] defines CV(connectivity verification),which belongs to on-demand PW monitoring. [RFC5885] defines proactive connectivity connection and PW/AC status notification. [RFC4447] and [I-D.ietf-pwe3-static-pw-status] give some other ways of PW/AC status notification.

1.1.2. VCCV

The goal of VCCV is to verify and further diagnose PW forwarding path. The extension to LDP is signaling VCCV capabilities to a peer

PE.

The extension to LDP is signaling VCCV LSP ping/ICMP ping capabilities to a peer PE.

1.1.3. VCCV BFD

[RFC5885] specifies four CV types for BFD by combining two types of encapsulation with two types of functionality. When multiple BFD CV Types are advertised, it also describes how to select one to use.

The extension to LDP is to signal VCCV BFD capabilities to a peer PE, and activate BFD protocol after PW is established. If the BFD parameters(such as sending interval) need to be modified, BFD itself will handle it.

1.1.4. PW Status

PW status codes provides a mechanism to signal the status of PW, or AC failure between the two PEs at each end of the PW. When PW control plane exists, the PW status TLV is carried in the initial Label Mapping message or Notification message to signal all PW status messages.

The extension to LDP is to signal PW status capabilities to a peer PE, and activate PW status notification function after PW is established. So when a event occurs, an update PW status will be sent.

1.1.5. Conclusion

In summary, IP/MPLS PW OAM functions and relation with control plane/ NMS is described in the table. This document will not replace or deprecate this; e.g.,VCCV capability advertisement and PW status negotiation for MPLS networks.

		LDP	LSP Ping	NMS
On-demand MPLS PW OAM	VCCV LSP ping	Capability negotiation		Capability configuration& Bootstrapping
	VCCV ICMP ping	Capability negotiation		Capability configuration& Bootstrapping
Proactive OAM	VCCV BFD	Capability negotiation& Bootstrapping		Capability configuration& Bootstrapping
	PW status	Capability negotiation& Bootstrapping		Capability configuration& Bootstrapping

Figure 1: IP/MPLS PW OAM functions

1.2. Analysis of PW OAM Configuration Extended by MPLS-TP

1.2.1. CC-CV-RDI

[I-D.ietf-mpls-tp-cc-cv-rdi] has been chosen to be the basis of pro-active MPLS-TP OAM functions. Because VCCV BFD currently has no CV function, it SHOULD evolve with [I-D.ietf-mpls-tp-cc-cv-rdi] to provide this function in TP environment. The use of the VCCV control channel provides the context, based on the MPLS-PW label, required to bind and bootstrap the BFD session to a particular PW (FEC) so local discriminator values are not exchanged; please refer to the analysis in [I-D.ietf-mpls-tp-oam-analysis] and [RFC5885]. However, in order to identify certain extreme cases of mis-connectivity and fulfill the requirements that the BFD mechanism MUST be the same for LSP, (MS-)PW and Section as well as for SPME, BFD MAY still need to use Discriminator values to identify the connection being verified at both ends of the PW. The discriminator values can be statically configured, or signaled via LSP Ping or LDP extensions defined in this document.

Timer negotiation, such as TX/RX interval is performed in subsequent BFD control messages [RFC5880], but it also can be gotten by control plane signaling [I-D.ietf-mpls-tp-oam-framework].

The source MEP-ID does not need to be carried, for they can be deduced from the advertised FEC (129) TLV, as described in [I-D.ietf-mpls-tp-identifiers].

PHB, which identifies the per-hop behavior of BFD packet, SHOULD be configured as well. This permits the verification of correct operation of QoS queuing as well as connectivity.

In conclusion, the configuration of VCCV BFD by control plane is not necessary, but for consistent operation of transport path and section, it SHOULD be an option.

1.2.2. PM Loss/Delay

[I-D.frost-mpls-tp-loss-delay]specifies mechanisms for performance monitoring of PWs, in particular it specifies loss and delay measurements.

For proactive LM, the transmission rate and PHB associated with the LM OAM packets originating from a MEP need be negotiated with the other MEP. LM OAM packets should be transmitted with the same PHB class that the LM is intended to measure.

Just like LM, Both one way and two way mode of proactive DM need the two MEPs nodes of PW to negotiate the measure interval and PHB value of OAM packets.

1.2.3. FMS

[I-D.ietf-mpls-tp-fault]specifies fault management signals with which a server PW can notify client PWs about various fault conditions to suppress alarms or to be used as triggers for actions in the client PWs. The following signals are defined: Alarm Indication Signal (AIS), Link Down Indication (LDI) and Lock Reporting (LKR).

For each MEP of each MEG, enabling/disabling the generation of FMS packets, the transmitted period and PHB SHOULD be configured. This can be done independently, and the values of configured parameters can be different, but for easy maintenance, these setting SHOULD be consistent.

In conclusion, the configuration of FMS by control plane is not necessary, but for easy maintenance, it SHOULD be an option also.

1.2.4. On-demand OAM functions

The extended on-demand OAM functions MAY need capability negotiation in the initialized LDP mapping message. However, On-demand PW OAM functions are expected to be carried out by directly accessing network nodes via a management interface; hence configuration and control of on-demand PW OAM functions are out-of-scope for this document.

1.2.5. Conclusion

According to the analysis above, LDP extensions to the LDP protocol to negotiate PW OAM capabilities, configure and bootstrap proactive PW OAM functions, such as, CC-CV-RDI, PM Loss/Delay, FMS. In this way OAM setup is bound to connection establishment signaling, avoiding two separate management/configuration steps (connection setup followed by OAM configuration) which would increases delay, processing and more importantly may be prune to mis-configuration errors.

Furthermore, LSP ping can be used to configure the proactive PW OAM function extended by MPLS-TP also, suitable for dynamic and static PW. For reference, the following table describes the different scope of different proactive OAM bootstrapping schemes of dynamic PW.

		LDP	LSP Ping	NMS
Proactive OAM	CC/CV/RDI	Capability negotiation& Function configuration& Bootstrapping	Function configuration& Bootstrapping	Capability configuration& Function configuration& Bootstrapping
	FMS	Capability negotiation& Function configuration& Bootstrapping	Function configuration& Bootstrapping	Capability configuration& Function configuration& Bootstrapping
	PM Loss/Delay	Capability negotiation& Function configuration& Bootstrapping	Function configuration& Bootstrapping	Capability configuration& Function configuration& Bootstrapping

Figure 2: MPLS-TP PW OAM functions

2. Conventions Used in This Document

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

3. MPLS-TP PW OAM Capability Advertisement

When a PW is first set up, the PEs MUST attempt to negotiate the usage of what kind of OAM functions. Up to now, there are PW status negotiation and VCCV capability advertisement. For the newly extended OAM function by MPLS-TP, such as PM loss/delay and FMS, the capability negotiation is accomplished as follows: A PE that supports the MPLS-TP PW OAM capability MUST include MPLS-TP PW OAM capability TLV in the initial Label Mapping message, following the PW Status TLV or VCCV parameter field in Interface Parameters TLV. If the extended on-demand OAM functions also need capability negotiation, just follow the same rules.

4. PW OAM Configuration Procedures

A PE may play active or passive role in the signaling of the PW. There exist two situations:

a) Active/active "C Both PEs of a PW are active (SS-PW), they select PW OAM configuration parameters and send with the Label Mapping message to each other independently.

b) Active/passive "C One PE is active and the others are passive (MS-PW). The active/passive role election is defined in Section 7.2.1 of [I-D.ietf-pwe3-segmented-pw] and applies here, this document does not define any new role election procedures.

The general rules of OAM configuration procedures are mostly identical between MS-PW and SS-PW, except that SS-PW does not need to configure MIP function and the Mapping message are sent out independently. This section takes MS-PW as an example to describe the general OAM configuration procedures. As for SS-PW, there may be some collisions of PW OAM configuration parameters, and these specific differences would be addressed in section 6.

4.1. Establishment of OAM Entities and Functions

Assuming there is one PW needs to be setup between T-PE1 and T-PE2, across S-PE1 and S-PE2. OAM functions must be setup and enabled in the appropriate order so that spurious alarms can be avoided.

```
          +-------+         +-------+         +-------+         +-------+
          |       |         |       |         |       |         |       |
          |      A|---------|B     C|---------|D     E|---------|F      |
          |       |         |       |         |       |         |       |
          +-------+         +-------+         +-------+         +-------+
            T-PE1             S-PE1             S-PE2             T-PE2
```

Figure 3: MS-PW OAM configuration scheme

Fist of all, T-PE1 MUST setup the OAM sink function to be prepared to receive OAM messages but MUST suppress any OAM alarms (e.g., due to missing or unidentified OAM messages). The Mapping message MUST be sent with the "OAM Alarms Enabled" cleared, "OAM MEP Entities desired" set and "OAM MIP Entities desired" set in the MPLS-TP PW OAM capability TLV.

When the Mapping message arrives at the down receivers, such as S-PE1, S-PE2 and T-PE2, they MUST establish and configure OAM entities according to the OAM information provided in mapping message. If this is not possible, a Label Release message SHOULD be sent and neither the OAM entities nor the PW SHOULD be established. If OAM entities are established successfully, the middle points (S-PE1 and S-PE2) MUST forward the Mapping message downstream, the endpoint (T-PE2) MUST set the OAM Source function and MUST be prepared to Send OAM messages.

The same rules are applied to the reverse direction (from T-PE2 to T-PE1), that is to say, T-PE2 needs to setup the OAM sink function to be prepared to receive OAM messages but MUST suppress any OAM alarms (e.g., due to missing or unidentified OAM messages). The Mapping message MUST be sent with the "OAM Alarms Enabled" cleared, "OAM MEP Entities desired" set, "OAM MIP Entities desired" set in the MPLS-TP PW OAM capability TLV. When T-PE1 receives the Mapping message, it completes any pending OAM configuration and enables the OAM source function to send OAM messages.

After this round, OAM entities are established and configured for the PW and OAM messages MAY already be exchanged, and OAM alarms can now be enabled. The T-PE nodes(T-PE1 and T-PE2), while still keeping OAM alarms disabled send a Notification message with "OAM Alarms Enabled" PW status flag set, and enable the OAM alarms after processing the Notification message. Data plane OAM is now fully functional, by the way, the MPLS-TP PW OAM Configuration TLV is not needed to be carried in the Notification message.

The PW may be setup with OAM entities right away with the first signaling, as described above, but a PW may be signaled and established without OAM configuration first, and OAM entities may be added later. This can be done by sending Notification message with the related configuration parameters subsequently.

4.2. Adjustment of OAM Parameters

There may be a need to change the parameters of an already established and configured OAM function during the lifetime of the PW. To do so the T-PE nodes need to send Notification message with the updated parameters. OAM parameters that influence the content and timing of OAM messages and identify the way OAM defects and alarms are derived and generated. Hence, to avoid spurious alarms, it is important that both sides, OAM sink and source, are updated in a synchronized way. First, the alarms of the OAM sink function should be suppressed and only then should expected OAM parameters be adjusted. Subsequently, the parameters of the OAM source function can be updated. Finally, the alarms of the OAM sink side can be enabled again.

In accordance with the above operation, T-PE1 MUST send Notification message with "OAM Alarms Enabled" cleared and including the updated MPLS-TP PW OAM Configuration TLV corresponding to the new parameter settings. The initiator (T-PE1) MUST keep its OAM sink and source functions running unmodified, but it MUST suppress OAM alarms after the updated Notification message is sent. The receiver (T-PE2) MUST first disable all OAM alarms, then update the OAM parameters according to the information in the Notification message and reply with a Notification message acknowledging the changes by including the MPLS-TP PW OAM Configuration TLV. Note that the receiving side has the possibility to adjust the requested OAM configuration parameters and reply with and updated MPLS-TP PW OAM Configuration TLV in the Notification message, reflecting the actually configured values. However, in order to avoid an extensive negotiation phase, in the case of adjusting already configured OAM functions, the receiving side SHOULD NOT update the parameters requested in the Notification message to an extent that would provide lower performance than what has been configured previously.

The initiator (T-PE1) MUST only update its OAM sink and source functions after it received the Notification message. After this Notification messages that exchange (in both directions) the OAM parameters are updated and OAM is running according the new parameter settings. However OAM alarms are still disabled, a subsequent Notification messages exchanges with "OAM Alarms Enabled" flag set are needed to enable OAM alarms again.

4.3. Deleting OAM Entities

In some cases it may be useful to remove some or all OAM entities and functions from one PW without actually tearing down the connection. To avoid any spurious alarm, the following procedure should be followed:

The T-PE nodes disable OAM alarms and SHOULD send Notification message each other with "OAM Alarms Enabled" cleared but unchanged OAM configuration and without the MPLS-TP PW OAM Configuration TLV. After that, T-PE1 (T-PE2) SHOULD delete OAM source functions, then send Notification message with "OAM MEP Entities desired" and "OAM MIP Entities desired" cleared. While T-PE2 (T-PE1) deletes OAM sink function when it receives the Notification message with "OAM MEP Entities desired" cleared, S-PE1 and S-PE2 delete MIP configuration when they receive the Notification message with "OAM MIP Entities desired" cleared.

Alternatively, if only some OAM functions need to be removed, the T-PE node sends the Notification message with the updated OAM Configuration TLV. Changes between the contents of the previously signaled OAM Configuration TLV and the currently received TLV represent which functions SHOULD be removed/added.

5. LDP extensions

Below, extensions to LDP are defined in order to configure MPLS-TP PW OAM functionalities during the PW setup.

5.1. MPLS-TP PW OAM capability TLV

The format of the MPLS-TP PW OAM Capability TLV is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0|  MPLS-TP PW OAM Capability |           Length              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|E|I|A|      MPLS-TP PW OAM Capability Flags              |F|D|L|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: MPLS-TP PW OAM Capability TLV

Currently defined OAM Capability Flags are:

0	PM Loss supported
1	PM Delay supported
2	FMS supported
29	OAM Alarms Enabled
30	OAM MIP entities desired
31	OAM MEP entities desired

One bit (0, IANA to assign): "PM Loss supported" is allocated.

One bit (1, IANA to assign): "PM delay supported" is allocated.

One bit (2, IANA to assign): "FMS supported" is allocated.

One bit (31, IANA to assign): "OAM MEP entities desired" is allocated. If the "OAM MEP entities desired" bit is set it is indicating that the establishment of OAM MEP entities are required at the endpoints of the signaled PW. If the establishment of MEPs is not supported, a Label Release message MUST be sent. If the "OAM MEP entities desired" bit is set and additional parameters are needed to be configured on the OAM entities, an "MPLS-TP PW OAM Configuration TLV" may be included in the Mapping or Notification message.

One bit (30, IANA to assign): "OAM MIP entities desired" is allocated. This bit can only be set if the "OAM MEP entities desired" bit is set. If the "OAM MIP entities desired" bit is set, it is indicating that the establishment of OAM MIP entities is required at every transit node of the signaled PW. If the establishment of a MIP is not supported, a Label Release message MUST be sent.

One bit (29, IANA to assign): "OAM Alarms Enabled" is allocated. This bit can only be set if the "OAM MEP entities desired" bit is set. If the "OAM Alarms Enabled" bit is set, it is indicating that the T-PE needs to enable OAM alarms. If the establishment of a MIP is not supported, a Label Release message MUST be sent.

[Editor notes]: If the MPLS-TP equipments support all the PW OAM functions defined and the OAM capability negotiation is not needed, this MPLS-TP PW OAM capability TLV just use to configure MEP/MIP entities and enable/disable OAM alarms.

5.2. MPLS-TP PW OAM Configuration TLV

The "OAM Configuration TLV", defined in [I-D.ietf-ccamp-oam-configuration-fwk], is depicted in the following figure. It may be carried in the Mapping and Notification messages, just following the PW Status TLV.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0|      Type (2) (IANA)     |           Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    OAM Type   |                Reserved                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                           sub-TLVs                            ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: MPLS-TP PW OAM Configuration TLV

OAM type: indicates a new type: the MPLS-TP PW OAM Configuration TLV (IANA to assign). If this type is not supported, a Label Release message MUST be sent. The specific OAM functions are specified in the "Function Flags" sub-TLV as depicted in [I-D.ietf-ccamp-oam-configuration-fwk], and the additional corresponding sub-TLVs are defined in section 3.2 of [I-D.ietf-ccamp-rsvp-te-mpls-tp-oam-ext].

For active/active signaling, if the flags in the "MPLS-TP PW OAM Function Flags sub-TLV" are different in the two Mapping message, the two PEs nodes can compare the node IDs. Label Withdraw message MUST be sent by the PE with lower ID, then it sends the Label Mapping message again with the same flags carried in the received Label Mapping message.

5.2.1. BFD Configuration TLV

BFD Configuration TLV follows the same TLV structure defined for RSVP-TE in section 3.3 of [I-D.ietf-ccamp-rsvp-te-mpls-tp-oam-ext].

For active/active signaling, if the flags of "BFD Configuration TLV" are different in the two Mapping message, similarly Label Withdraw message MUST be sent by the PE with lower ID. Then it sends the Label Mapping message again with the same flags carried in the "BFD configuration TLV" of the received Label Mapping message. If the flags of "BFD Configuration TLV" are the same, but the values of "Negotiation Timer parameters sub-TLV" are different, both the PE nodes MUST adopt the bigger interval and detection time multiplier.

5.2.2. MPLS-TP PW PM Loss TLV

MPLS-TP PW PM Loss TLV follows the same TLV structure defined for RSVP-TE in section 3.4 of [I-D.ietf-ccamp-rsvp-te-mpls-tp-oam-ext].

For active/active signaling, if the flags of "MPLS-TP PW OAM PM Loss TLV" are different in the two Mapping message, similarly Label Withdraw message MUST be sent by the PE with lower ID. Then it sends the Label Mapping message again with the same flags carried in the "MPLS-TP PW OAM PM Loss TLV" of the received Label Mapping message. If the flags of "MPLS-TP PW OAM PM Loss TLV" are the same, but the Measurement Interval and Loss Threshold are different, both the PE nodes MUST adopt the bigger values.

5.2.3. MPLS-TP PW PM Delay TLV

MPLS-TP PW PM Delay TLV follows the same TLV structure defined for RSVP-TE in section 3.5 of [I-D.ietf-ccamp-rsvp-te-mpls-tp-oam-ext].

For active/active signaling, if the flags of "MPLS-TP PW OAM PM Delay TLV" are different in the two Mapping message, similarly Label Withdraw message MUST be sent by the PE with lower ID. Then it sends the Label Mapping message again with the same flags carried in the "MPLS-TP PW OAM PM Delay TLV" of the received Label Mapping message. If the flags of "MPLS-TP PW OAM PM Delay TLV" are the same, but the Measurement Interval and Delay Threshold are different, both the PE nodes MUST adopt the bigger values.

5.2.4. MPLS-TP PW FMS TLV

MPLS-TP PW FMS TLV follows the same TLV structure defined for RSVP-TE in section 3.6 of [I-D.ietf-ccamp-rsvp-te-mpls-tp-oam-ext].

For active/active signaling, if the flags of "MPLS-TP PW OAM FMS TLV" are different in the two Mapping message, similarly Label Withdraw message MUST be sent by the PE with lower ID. Then it sends the Label Mapping message again with the same flags carried in the "MPLS-TP PW OAM FMS TLV" of the received Label Mapping message.

Notes: CSF are overlapped with PW Status TLV, and the field of Refresh Timer is not needed.

6. IANA Considerations

6.1. LDP TLV Types

This document specifies the following new LDP TLV types:

- o MPLS-TP PW OAM Capability TLV;
- o MPLS-TP PW OAM Configuration TLV;

Sub-TLV types to be carried in the "MPLS-TP PW OAM Configuration TLV":

- o MPLS-TP PW OAM Function Flags sub-TLV;
- o BFD Configuration sub-TLV;
- o MPLS-TP PW PM Loss sub-TLV;
- o MPLS-TP PW PM Delay sub-TLV;
- o MPLS-TP PW FMS sub-TLV;

Sub-TLV types to be carried in the "BFD Configuration sub-TLV":

- o Local Discriminator sub-TLV;
- o Negotiation Timer Parameters sub-TLV.

6.2. LDP Status Code

TBD.

7. Security Considerations

TBD.

8. Acknowledgement

The authors would like to thank Thomas Nadeau for his valuable comments.

9. Normative references

[I-D.absw-mpls-lsp-ping-mpls-tp-oam-conf]
Bellagamba, E., Andersson, L., Skoldstrom, P., and D. Ward, "Configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions Using LSP Ping", draft-absw-mpls-lsp-ping-mpls-tp-oam-conf-00 (work in progress), July 2010.

[I-D.frost-mpls-tp-loss-delay]
Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for the MPLS Transport Profile", draft-frost-mpls-tp-loss-delay-02 (work in progress), June 2010.

[I-D.ietf-ccamp-oam-configuration-fwk]
Takacs, A., Fedyk, D., and H. Jia, "OAM Configuration Framework and Requirements for GMPLS RSVP-TE", draft-ietf-ccamp-oam-configuration-fwk-03 (work in progress), January 2010.

[I-D.ietf-ccamp-rsvp-te-mpls-tp-oam-ext]
Bellagamba, E., Andersson, L., Skoldstrom, P., Ward, D., and A. Takacs, "Configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions Using RSVP-TE", draft-ietf-ccamp-rsvp-te-mpls-tp-oam-ext-03 (work in progress), July 2010.

[I-D.ietf-mpls-tp-cc-cv-rdi]
Allan, D., Drake, J., Swallow, G., Boutros, S., Sivabalan, S., and D. Ward, "Proactive Connectivity Verification, Continuity Check and Remote Defect indication for MPLS Transport Profile", draft-ietf-mpls-tp-cc-cv-rdi-02 (work in progress), October 2010.

[I-D.ietf-mpls-tp-fault]
Swallow, G., Fulignoli, A., Vigoureux, M., Boutros, S., Ward, D., Bryant, S., and S. Sivabalan, "MPLS Fault Management OAM", draft-ietf-mpls-tp-fault-02 (work in progress), July 2010.

[I-D.ietf-mpls-tp-identifiers]
Bocci, M. and G. Swallow, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-02 (work in progress), July 2010.

[I-D.ietf-mpls-tp-lsp-ping-bfd-procedures]
Bahadur, N., Aggarwal, R., Ward, D., Nadeau, T., Sprecher, N., and Y. Weingarten, "LSP-Ping and BFD encapsulation over ACH", draft-ietf-mpls-tp-lsp-ping-bfd-procedures-01 (work in progress), August 2010.

[I-D.ietf-mpls-tp-oam-analysis]
Sprecher, N., Bellagamba, E., and Y. Weingarten, "MPLS-TP OAM Analysis", draft-ietf-mpls-tp-oam-analysis-02 (work in progress), July 2010.

[I-D.ietf-mpls-tp-oam-framework]
Allan, D., Busi, I., Niven-Jenkins, B., Fulignoli, A., Hernandez-Valencia, E., Levrau, L., Sestito, V., Sprecher, N., Helvoort, H., Vigoureux, M., Weingarten, Y., and R. Winter, "Operations, Administration and Maintenance

Framework for MPLS- based Transport Networks",
draft-ietf-mpls-tp-oam-framework-09 (work in progress),
October 2010.

[I-D.ietf-mpls-tp-oam-requirements]
Vigoureux, M. and D. Ward, "Requirements for OAM in MPLS Transport Networks",
draft-ietf-mpls-tp-oam-requirements-06 (work in progress),
March 2010.

[I-D.ietf-pwe3-dynamic-ms-pw]
Martini, L., Bocci, M., Balus, F., Bitar, N., Shah, H., Aissaoui, M., Rusmisel, 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.

[I-D.ietf-pwe3-redundancy]
Muley, P., "Pseudowire (PW) Redundancy",
draft-ietf-pwe3-redundancy-03 (work in progress),
May 2010.

[I-D.ietf-pwe3-segmented-pw]
Martini, L., Nadeau, T., Metz, C., Bocci, M., Aissaoui, M., Balus, F., and M. Duckett, "Segmented Pseudowire",
draft-ietf-pwe3-segmented-pw-18 (work in progress),
September 2010.

[I-D.ietf-pwe3-static-pw-status]
Martini, L., Swallow, G., and M. Bocci, "Pseudowire Status for Static Pseudowires",
draft-ietf-pwe3-static-pw-status-00 (work in progress),
February 2010.

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

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

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

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label

Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC5003] Metz, C., Martini, L., Balus, F., and J. Sugimoto, "Attachment Individual Identifier (AII) Types for Aggregation", RFC 5003, September 2007.

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

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

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

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

[RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

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

Authors’ Addresses

Fei Zhang (editor)
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Yuhuatai District,Nanjing 210012
P.R.China

Phone: +86 025 52877612
Email: zhang.fei3@zte.com.cn

Bo Wu (editor)
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Yuhuatai District,Nanjing 210012
P.R.China

Phone: +86 025 52877276
Email: wu.bo@zte.com.cn

Elisa Bellagamba (editor)
Ericsson
Farogatan 6
Kista, 164 40
Sweden

Phone: +46 761440785
Email: elisa.bellagamba@ericsson.com

Attila Takacs
Ericsson
Laborc u. 1.
Budapest, 1037
Hungary

Email: attila.takacs@ericsson.com

Xuehui Dai
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Yuhuatai District,Nanjing 210012
P.R.China

Phone: +86 025 52877612
Email: dai.xuehui@zte.com.cn

Min Xiao
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Yuhuatai District,Nanjing 210012
P.R.China

Phone: +86 025 52877612
Email: xiao.min2@zte.com.cn