

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 30, 2014

S. Sivabalan
S. Boutros
Cisco Systems, Inc.
H. Shah
Ciena Corp.
September 30, 2013

MAC Address Withdrawal over Static Pseudowire
draft-boutros-l2vpn-mppls-tp-mac-wd-02.txt

Abstract

This document specifies a mechanism to signal MAC address withdrawal notification using PW Associated Channel (ACH). Such notification is useful when statically provisioned PWs are deployed in VPLS/H-VPLS environment.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on March 26, 2014.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

Table of Contents

1. Introduction	2
2. Terminology	2
3. MAC Withdraw OAM Message	3
4. Operation	4
4.1. Operation of Sender	4
4.2. Operation of Receiver	5
5. IANA Considerations	5
6. References	5
6.1. Normative References	5
6.2. Informative References	6
Authors' Addresses	6

1. Introduction

An LDP-based MAC Address Withdrawal Mechanism is specified in [RFC4762] to remove dynamically learned MAC addresses when the source of those addresses can no longer forward traffic. This is accomplished by sending an LDP Address Withdraw Message with a MAC List TLV containing the MAC addressed to be removed to all other PEs over LDP sessions. When the number of MAC addresses to be removed is large, empty MAC List TLV may be used. [MAC-OPT] describes an optimized MAC withdrawal mechanism which can be used to remove only the set of MAC addresses that need to be re-learned in H-VPLS networks. The solution also provides optimized MAC Withdrawal operations in PBB-VPLS networks.

A PW can be signaled via LDP or can be statically provisioned. In the case of static PW, LDP based MAC withdrawal mechanism cannot be used. This is analogous to the problem and solution described in [RFC4762] where PW OAM message has been introduced to carry PW status TLV using in-band PW Associated Channel. In this document, we propose to use PW OAM message to withdraw MAC address(es) learned via static PW.

2. Terminology

The following terminologies are used in this document:

ACK: Acknowledgement for MAC withdraw message.

LDP: Label Distribution Protocol.

MAC: Media Access Control.

PE: Provide Edge Node.

MPLS: Multi Protocol Label Switching.

PW: PseudoWire.

PW OAM: PW Operations, Administration and Maintenance.

TLV: Type, Length, and Value.

VPLS: Virtual Private LAN Services.

3. MAC Withdraw OAM Message

LDP provides a reliable packet transport for control plackets for dynamic PWs. This can be contrasted with static PWs which rely on re-transmission and acknowledgments (ACK) for reliable OAM packet delivery as described in [RFC6478]. The proposed solution for MAC withdrawal over static PW also relies on re-transmissions and ACKs. However, ACK is mandatory. A given MAC withdrawal notification is sent as a PW OAM message, and the sender keeps re-transmitting the message until it receives an ACK for that message. Once a receiver successfully remove MAC address(es) in response to a MAC address withdraw OAM message, it should not unnecessarily remove MAC address(es) upon getting refresh message(s). To facilitate this, the proposed mechanism uses sequence number, and defines a new TLV to carry the sequence number.

The format of the MAC address withdraw OAM message is shown in Figure 1. The PW OAM message header is exactly the same as what is defined in [RFC6478]. Since the MAC withdrawal PW OAM message is not refreshed forever. A MAC address withdraw OAM message MUST contain a "Sequence Number TLV" otherwise the entire message is dropped. It MAY contain MAC Flush Parameter TLVs defined in [MAC-OPT] when static PWs are deployed in H-VPLS and PBB-VPLS scenarios. The first 2 bits of the sequence number TLV are reserved and MUST be set to 0 on transmit and ignored on receipt.

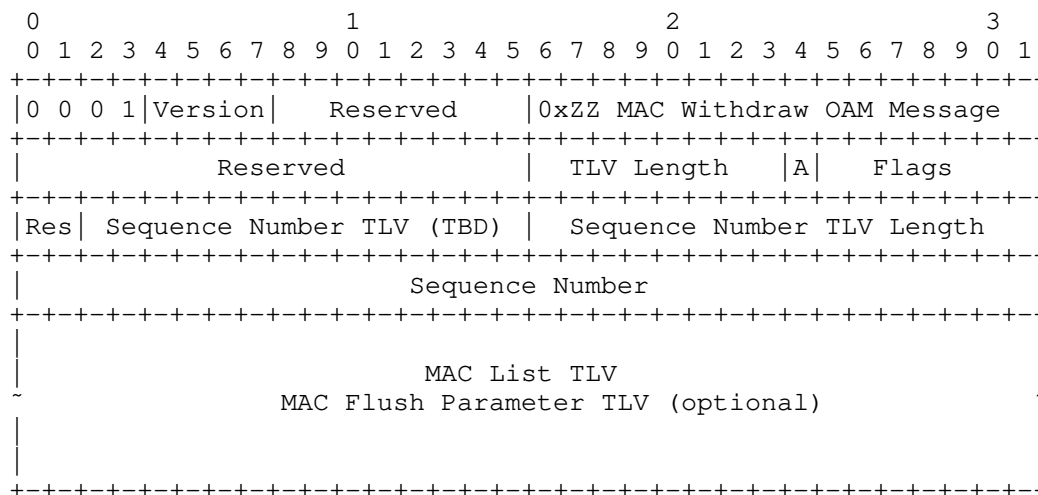


Figure 1: MAC Address Withdraw PW OAM Packet Format

In this section, MAC List TLV and MAC Flush Parameter TLV are collectively referred to as "MAC TLV(s)". The processing rules of MAC List TLV are governed by [RFC4762], and the corresponding rules of MAC Flush Parameter TLV are governed by [MAC-OPT].

"TLV Length" is the total length of all TLVs in the message, and "Sequence Number TLV Length" is the length of the sequence number field.

A single bit (called A-bit) is set to indicate if a MAC withdraw message is for ACK. Also, ACK does not include MAC TLV(s).

Only half of the sequence number space is used. Modular arithmetic is used to detect wrapping of sequence number. When sequence number wraps, all MAC addresses are flushed and the sequence number is reset.

4. Operation

This section describes how the initial MAC withdraw OAM messages are sent and retransmitted, as well as how the messages are processed and retransmitted messages are identified.

4.1. Operation of Sender

Each PW is associated with a counter to keep track of the sequence number of the transmitted MAC withdrawal messages. Whenever a node

sends a new set of MAC TLVs, it increments the transmitted sequence number counter, and include the new sequence number in the message.

The sender expects an ACK from the receiver within a time interval which we call "Retransmit Time" which can be either a default or configured value. If the ACK arrives within the Retransmit Time, the sender assumes that the message transmission is successful. Otherwise, it retransmits the message with the same sequence number as the original message.

4.2. Operation of Receiver

Each PW is associated with a register to keep track of the sequence number of the MAC withdrawal message received last. Whenever a MAC withdrawal message is received, and if the sequence number on the message is greater than the value in the register, the MAC address(es) contained in the MAC TLV(s) is/are removed, and the register is updated with the received sequence number. The receiver sends an ACK whose sequence number is the same as that in the received message.

If the sequence number in the received message is smaller than or equal to the value in the register, the MAC TLV(s) is/are not processed. However, an ACK with the received sequence number MUST be sent as a response. The receiver processes the ACK message as an acknowledgement for all the MAC withdraw messages sent up to the sequence number present in the ACK message and terminates retransmission.

As mentioned above, since only half of the sequence number space is used, the receiver MUST use modular arithmetic to detect wrapping of the sequence number.

5. IANA Considerations

The proposed mechanism requests IANA to assign new channel type (recommended value 0x0028) from the registry named "Pseudowire Associated Channel Types". The description of the new channel type is "Pseudowire MAC Withdraw OAM Channel".

IANA needs to create a new registry for Pseudowire Associated Channel TLVs, and create an entry for "Sequence Number TLV". The recommended value is 0x0001.

6. References

6.1. Normative References

- [MAC-OPT] Dutta, P., Balus, F., Stokes, O., and G. Calvinac, "LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS", draft-ietf-l2vpn-vpls-ldp-mac-opt-08.txt (work in progress), February 2013.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC6478] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", RFC 6478, May 2012.

6.2. Informative References

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

Authors' Addresses

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

Email: msiva@cisco.com

Sami Boutros
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
US

Email: sboutros@cisco.com

Himanshu Shah
Ciena Corp.
3939 North First Street
San Jose, CA 95134
US

Email: hshah@ciena.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2014

WQ. Cheng
L. Wang
H. Li
China Mobile
K. Liu
Huawei Technologies Co., Ltd.
S. Davari
Broadcom Corporation
October 21, 2013

MPLS-TP PWE3 dual-homed protection (MPDP)
draft-cheng-mpls-tp-pwe3-dual-homed-protection-00

Abstract

This document presents the requirements for Dual-homed protection in the MPLS-TP networks and defines a protocol that can protect the failure of an attachment circuit (AC) or the failure of a provider edge (PE) node or the failure of a pseudowire (PW) in the packet-switched network (PSN).

Status of This Memo

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

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

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

This Internet-Draft will expire on April 24, 2014.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Application scenarios of Dual-Homed protection	3
2.1. One-side Dual-Homing topology	4
2.2. Two-side Dual-Homing topology	4
3. PWE3 dual-homed protection mechanism	5
3.1. Multi-chassis PW protection	5
3.2. Multi-chassis LAG	7
4. Three point-switch collaboration	8
5. Formal Syntax	9
6. Security Considerations	9
7. Acknowledgments	9
8. Author's Addresses	9
9. References	9
Authors' Addresses	10

1. Introduction

The linear protection and Ring protection mechanisms for MPLS-TP is described in RFC 6378, RFC 6974 and other IETF drafts. These mechanisms work within the PSN and provide fast recovery when link failure or P node failure occurs. However, they are unable to protect against the failure of a PE node or the failure of an attachment circuit.

The PW redundancy solution which is defined by RFC 6718 requires separate mechanisms to recover the PE and AC link failure from the PSN failure.

The operators need an end-to-end network's survivability for guaranteed services, so the protection mechanisms for AC, PE and failure within PSN are all needed. In order to meet the requirement, multiple layers and across nested recovery domains protection should be deployed. It raised the following issues:

- o longer recovery time, because hold-off time should be set to avoid race scenarios, which makes the switching time longer.
- o lower bandwidth efficiency because of multi-layer protections.

- o extra configuration makes the operation and maintenance more complicated.

- o if RFC 6718 is used, the AC link failure will result in protection switching performed within PSN, that means the failure of an AC are propagated to the remote PEs on the other side of the network.

In order to improve on RFC 6718, dual-homed protection mechanism should meet the following requirements

O Using a single layer protection for PSN, AC and PE failure

PWE3 Dual-Homed protection needs to recover PE failures, tunnel failures within PSN and AC link failures through a single layer protection mechanism so that the multi-layer protection can be avoided.

O Independent failure recovery

The principle of independent failure recovery is that the protection switching is solely performed within the network domain where the failure takes place. For instance, When a failure in a an AC happens, there is no need to inform the remote PE about the failure and there is no need to change the PW and path in the PSN.

O To deploy dual-homed network protection, as far as protocols which PE previously support, such as linear protection protocol, can be reused, upgrading remote PEs should be avoided.

According to RFC5654 "2.5.6. Topology-Specific Recovery Mechanisms", "MPLS-TP MAY support recovery mechanisms that are optimized for specific network topologies. These mechanisms MUST be interoperable with the mechanisms defined for arbitrary topology (mesh) networks to enable the protection of end-to-end transport paths ",this document presents a single layer dual-homed protection to meet those requirements.

2. Application scenarios of Dual-Homed protection

The application scenarios of Dual-homed protection can be classified into a One-side Dual-Homing topology and a Two-side Dual-homing topology.

2.1. One-side Dual-Homing topology

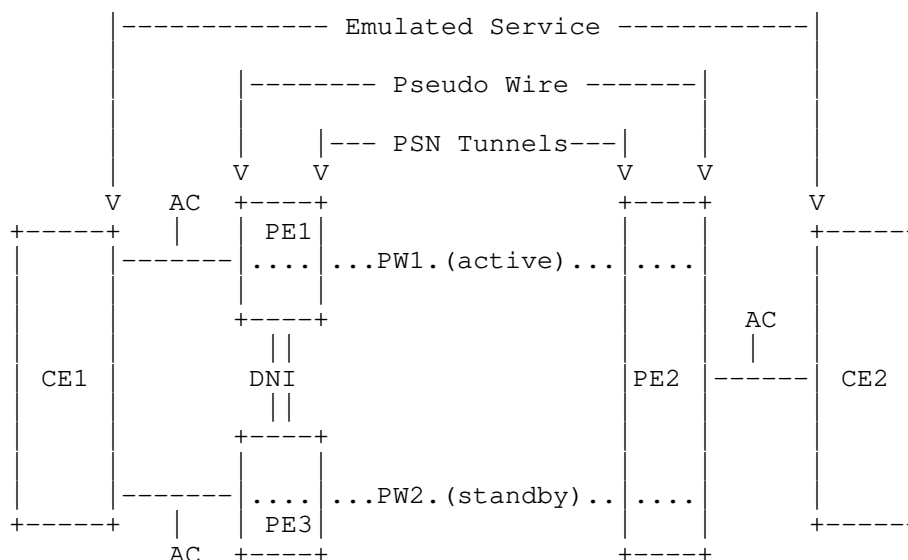
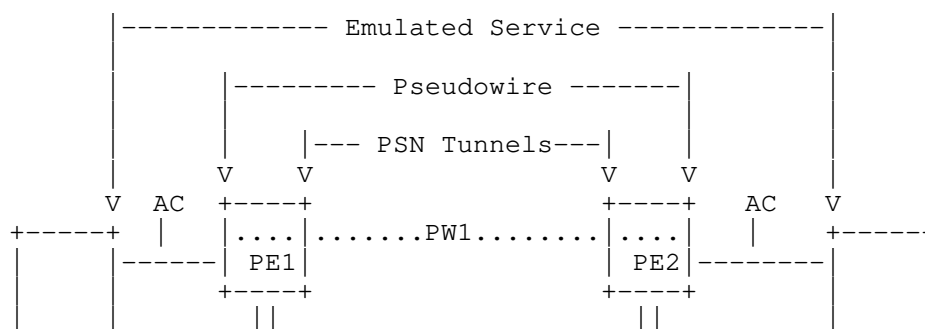


Figure 1 One-side PW Dual-Homing protection

Figure 1 illustrates the network scenario of one-side CE dual-homing topology. The dual-homing gateways for CE1 are PE1 and PE3, while PE2 is the single-homing for CE2. This scheme protects the node failures of PE1 and PE3 and the link failures between PE1 and CE1/PE3 and CE1. This scheme can be used in back hauling application scenarios. For example, nodeB serves for CE2 while RNC serves for CE1. PE2 works as an access layer MPLS-TP device while PE1 and PE3 works as a pair of core layer MPLS-TP devices.

2.2. Two-side Dual-Homing topology



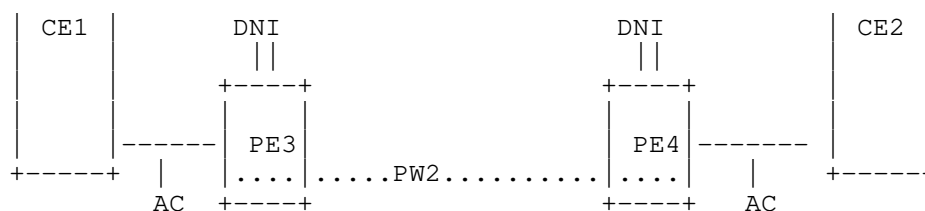


Figure 2 dual-side PW Dual-Homing protection

Figure 2 illustrates the network scenario of two-side CE dual-homing. The dual-homing nodes are PE1 and PE3 for CE1, and PE2 and PE4 for CE2, respectively. This scenario protects the PE1/PE3 and PE2/PE4 nodes failures. It also protects the links failure between PE1 and CE1, PE3 and CE1, PE2 and CE2, and PE4 and CE2. Meanwhile, dual-homing protection needs to handle the recovery of PSN Tunnel failure as well. As for broadband services provider, this scenario is mainly used in services for important business customers. Here, CE1 and CE2 can be regarded as service access points.

3. PWE3 dual-homed protection mechanism

In a PWE3 dual-homed protection mechanism, Multi-chassis PW protection is used between PEs, Multi-chassis LAG is used between dual-homed PE node group.

3.1. Multi-chassis PW protection

RFC6738(MPLS Transport Profile (MPLS-TP) Linear Protection) and ITU-T G.8131 have defined linear protection mechanism for MPLS-TP network. The PEs of working PW are the same as its protecting PW and the protection switching mechanism is running on each PE.

Dual-homed PW protection mechanism keeps the same protection switching mechanism with linear protection in the Remote PE (PE3 in figure 3) and the working PW and protecting PW are terminated in Dual-homed PEs (PE1 and PE2 in figure 3) respectively in order to protect Dual-homed PEs. The protection switching mechanism will be detailed in the following chapters.

Dual nodes interconnection (DNI) PW is set up between two dual-homed PE nodes, and it is used to bridge traffic when failure occurs.

Messages between two dual-homed node include channel status notify message and protection group status notify message. The format of these messages is TBD.

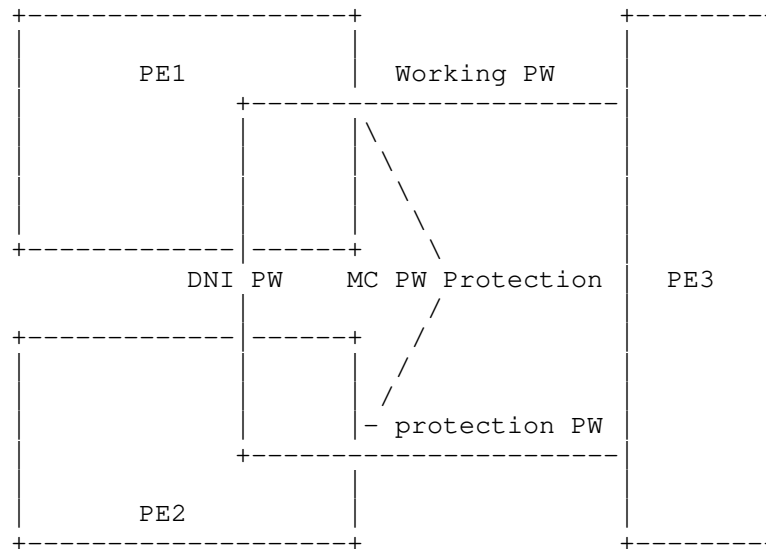


Figure 3 MC PW linear protection mechanism

The failure scenarios are listed as follows:

O One direction Failure of Working PW (PE1 to PE3):

PE3 detects failure and sends PSC or APS message in protection PW. When PE2 receives the failure information, it will exchange a switching message with PE3. At last, PE2 and PE3 will switch the traffic to the protection PW. PE2 will periodically send PE1 MC PW protection group status messages, and then PE1 will execute the switching according to the status of the MC PW protection.

O One direction Failure of Working PW (PE3 to PE1):

PE1 will detect the failure and send PE2 a MC PW protection message to notify work PW failure through DNI PW. PE2 will execute the switch with PE3 based on PSC or APS. PE2 will periodically send PE1 MC PW protection group status messages, and PE1 will execute switching according to the status of the MC PW protection.

O Bi-direction Failure of Working PW (Between PE3 and PE1):

Both of PE1 and PE3 will detect link fault respectively. PE3 executes switching to protection PW based on APS or PSC. PE1 sends PE2 a MC PW protection message to notify working PW failure through DNI PW, and then PE2 will execute switching to protection PW. PE2

will periodically send PE1 MC PW protection group status messages, and PE1 will execute switching according to the status of the MC PW protection.

O Working PE failure:

PE3 will detect failure, and send PSC or APS message in the protection PW. After PE2 exchanges the switching message with PE3, PE2 and PE3 will switch traffic to the protection PW.

3.2. Multi-chassis LAG

LAG(Link Aggregation Group) and LACP(Link Aggregation control protocol) is defined in IEEE802.1ax. LAG is used to expand bandwidth and protect link failure.

DRNI (Distributed Resilient Network Interconnect) and DRCP (Distributed Relay Control Protocol) is defined in IEEE P802.1AX-REV-D2.2. DRNI expands the concept of Link Aggregation so that, at either one or both ends of a Link Aggregation Group, the single Aggregation System is replaced by a Portal, composed from one to three Portal Systems.

In the document, DRNI is used to protect Ethernet link failure between CE and PE and dual-homed node Failure on the CE side as shown in figure 4.

O Working PE failure: Detailed signaling between PE1, PE2, CE is TBD

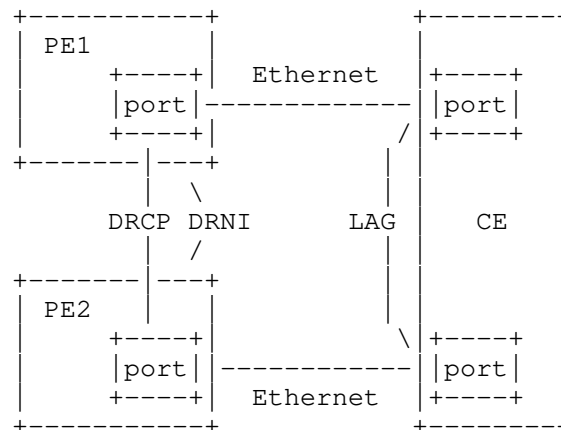


Figure 4 DRNI mechanism

4. Three point-switch collaboration

In dual-homed PE nodes, the protection mechanism in PSN network (MC PW protection) and protection mechanism between PE and CE (DRNI) should cooperate to ensure an appropriate switch operation.

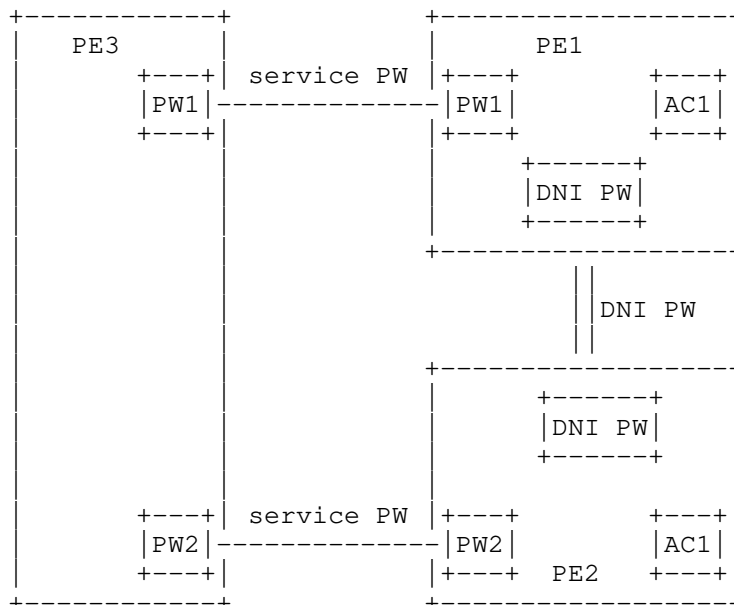


Figure 6 three point status machine

Service PW is the PW which carry service between dual-homed PE and remote PE. Service PW status is decided by MC PW protection mechanism.

DNI PW is the bridge PW between two dual-homed PE nodes. It is used to bridge traffic when PSN tunnel or AC failure occurs.

AC status is the status of AC port between dual-homed PW and CE, being either active or standby. It is decided by DRNI mechanism.

srv PW	AC	PW fwd	DNI fwd	AC fwd
Active	Active	to AC	to PW	to PW

Active	Standby	to DNI	to PW	to PW
+-----+	+-----+	+-----+	+-----+	+-----+
Standby	Active	to AC	to AC	to DNI
+-----+	+-----+	+-----+	+-----+	+-----+
Standby	Standby	to DNI	to AC	to DNI
+-----+	+-----+	+-----+	+-----+	+-----+

Figure 7 three point status machine in dual-homed nodes

The principle of three point status machine in dual-homed nodes:

O If AC status is active, establish connection from service PW to AC.
If AC status is standby, establish connection from service PW to DNI PW.

O If service PW is active, establish connection from AC to service PW. If service PW status is standby, establish connection from AC to DNI PW.

O If service PW is active, establish connection from DNI PW to service PW. If service PW status is standby, establish connection from DNI PW to AC.

5. Formal Syntax

None

6. Security Considerations

None

7. Acknowledgments

None

8. Author's Addresses

None

9. 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.

- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.
- [RFC6718] Muley, P., Aissaoui, M., and M. Bocci, "Pseudowire Redundancy", RFC 6718, August 2012.
- [RFC6974] Weingarten, Y., Bryant, S., Ceccarelli, D., Caviglia, D., Fondelli, F., Corsi, M., Wu, B., and X. Dai, "Applicability of MPLS Transport Profile for Ring Topologies", RFC 6974, July 2013.

Authors' Addresses

Weiqiang Cheng
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: chengweiqiang@chinamobile.com

Lei Wang
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: Wangleiyj@chinamobile.com

Han Li
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: Lihan@chinamobile.com

Kai Liu
Huawei Technologies Co., Ltd.
Huawei base, Bantian, Longgang District
Shenzhen 518129
China

Email: alex.liukai@huawei.com

Shahram Davari
Broadcom Corporation
3151 Zanker Road
San Jose, CA 95134-1933

Email: davari@broadcom.com

PWE3
Internet-Draft
Intended status: Informational
Expires: April 24, 2013

YJ. Stein
RAD Data Communications
D. Black
EMC Corporation
B. Briscoe
BT
October 21, 2012

PW Congestion Considerations
draft-ietf-pwe3-congcons-01

Abstract

Pseudowires (PWs) have become a common mechanism for tunneling traffic, and may be found competing for network resources both with other PWs and with non-PW traffic, such as TCP/IP flows. It is thus worthwhile specifying under what conditions such competition is safe, i.e., the PW traffic does not significantly harm other traffic or contribute more than it should to congestion. We conclude that PWs transporting responsive traffic behave as desired without the need for additional mechanisms. For inelastic PWs (such as TDM PWs) we derive a bound under which such PWs consume no more network capacity than a TCP flow.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 24, 2013.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
2. PWs Comprising Elastic Flows	4
3. PWs Comprising Inelastic Flows	5
4. Security Considerations	12
5. IANA Considerations	12
6. Informative References	13
Appendix A. Loss Probabilities for TDM PWs	14
Appendix B. Effect of Packet Loss on Voice Quality for TDM PWs .	16
Authors' Addresses	19

1. Introduction

A pseudowire (PW) is a construct for tunneling a native service over a Packet Switched Network (PSN) (see [RFC3985]), such as IPv4, IPv6, or MPLS. The PW packet encapsulates a unit of native service information by prepending the headers required for transport in the particular PSN (which must include a demultiplexer field to distinguish the different PWs) and preferably the 4 byte PWE3 control word. PWs have no bandwidth reservation mechanism, meaning that when multiple PWs are transported in parallel there is no defined means for guaranteeing network resources for any particular PW. This competition for resources may translate to a particular PW not being able to deliver the QoS required to emulate the native service. For example, MPLS-TE enables achieving a particular desired allocation of resources between multiple LSPs; however, when multiple Ethernet PWs are placed in a single MPLS tunnel, there is no way to similarly divide resources amongst them (although DiffServ QoS prioritization may be available for PWs). The use of PWs in service provider MPLS networks is well understood and will not be discussed further here.

While PWs are most often placed in MPLS tunnels, there are several mechanisms that enable transporting PWs over an IP infrastructure. These include:

- TDM PWs ([RFC4553][RFC5086][RFC5087]) that define UDP/IP encapsulations,
- L2TPv3 PWs,
- MPLS PWs directly over IP according to RFC 4023 [RFC4023],
- MPLS PWs over GRE over IP according to RFC 4023 [RFC4023].

Whenever PWs are transported over IP, they may compete with congestion-responsive flows (e.g., TCP flows). Hence in order to prevent congestion collapse the PWs MUST behave in a fashion that does not cause undue damage to the throughput of such congestion-responsive flows [RFC2914].

At first glance one may think that this would require a PW transported over IP to be considered as a single flow, on a par with a single TCP flow. Were we to accept this tenet, we would require a PW to back off under congestion to consume no more bandwidth than a single TCP flow under such conditions (see [RFC5348]). However, since PWs may carry traffic from many users, it makes more sense to consider each PW to be equivalent to multiple TCP flows. We will discuss whether PWs consisting of elastic flows need a back-off strategy in Section 2.

TDM PWs ([RFC4553][RFC5086][RFC5087]) represent inelastic constant bit-rate (CBR) flows that may require lower or higher throughput than that consumed by an otherwise-unconstrained TCP flow would under the same network conditions. In any case a TDM PW is not able to respond

to congestion in a TCP-like manner; on the other hand, the total bandwidth they consume remains constant and does not increase to consume additional bandwidth as TCP rates back off. If the bandwidth consumed by a TDM PW is considered detrimental, the only available remedy is to completely shut down the PW. Such a shutdown would impact multiple users, and the service restoration time would in general be lengthy. We will discuss when the shutdown of inelastic PWs can be avoided in Section 3.

2. PWs Comprising Elastic Flows

In this section we consider Ethernet PWs that primarily carry congestion-responsive traffic. We will show that we automatically obtain the desired congestion avoidance behavior, and that additional mechanisms are not needed.

Let us assume that an Ethernet PW aggregating several TCP flows is flowing alongside several TCP/IP flows. Each Ethernet PW packet carries a single Ethernet frame that carries a single IP packet that carries a single TCP segment. Thus, if congestion is signaled by an intermediate router dropping a packet, a single end-user TCP/IP packet is dropped, whether or not that packet is encapsulated in the PW.

The result is that the individual TCP flows inside the PW experience the same drop probability as the non-PW TCP flows. Thus the behavior of a TCP sender (retransmitting the packet and appropriately reducing its sending rate) is the same for flows directly over IP and for flows inside the PW. In other words, individual TCP flows are neither rewarded nor penalized for being carried over the PW. On the other hand, the PW does not behave as a single TCP flow; it will consume the aggregated bandwidth of its component flows, and backs off much less sharply than a single flow would.

We claim that this is precisely the desired behavior. Any fairness considerations should be applied to the individual TCP flows, and not to the aggregate. Were individual TCP flows rewarded for being carried over a PW, this would create an incentive to create PWs for no operational reason. Were individual flows penalized, there would be a deterrence that could impede pseudowire deployment.

There have been proposals to add additional TCP-friendly mechanisms to PWs, for example by carrying PWs over DCCP. In light of the above arguments, it is clear that this would force the PW to behave as a single flow, rather than N flows, and penalize the constituent TCP flows. In addition, the individual TCP flows would still back off due to their end points being oblivious to the fact that they are

carried over a PW. This will further degrade the flow's throughput as compared to a non-PW-encapsulated flow. Thus, such additional mechanisms contradict the behavior previously described as desirable.

3. PWs Comprising Inelastic Flows

TDM PWs ([RFC4553][RFC5086][RFC5087]) are more problematic than the elastic PWs of the previous section. Being constant bit-rate (CBR), they can not be made responsive to congestion. On the other hand, being CBR, they also do not attempt to capture additional bandwidth when TCP flows back off.

Since a TDM PW continuously consumes a constant amount of bandwidth, if the bandwidth occupied by a TDM PW endangers the network as a whole, the only recourse is to shut it down, denying service to all customers of the TDM native service. We should mention in passing that under certain conditions it may be possible to reduce the bandwidth consumption of a TDM PW. A prevalent case is that of a TDM native service that carries voice channels that may not all be active. Using the AAL2 mode of [RFC5087] (perhaps along with connection admission control) can enable bandwidth adaptation, at the expense of more sophisticated native service processing (NSP).

In the following we will show that for many cases of interest a TDM PW, treated as a single flow, will behave in a reasonable manner without any additional mechanisms. We will focus on structure-agnostic TDM PWs [RFC4553] although our analysis can be readily applied to structure-aware PWs (see Appendix A).

There are two network parameters relevant to our discussion, namely the one-way delay D and the loss probability p . The one-way delay of a native TDM service consists of the physical time-of-flight plus 125 microseconds for each TDM switch traversed. This is very small as compared to PSN network-crossing latencies. Many protocols and applications running over TDM circuits thus require low delay, and we need thus only consider delays of up to about 32 milliseconds.

The TDM PW RFCs specify the egress behavior upon experiencing packet loss. Structure-agnostic transport has no alternative to outputting an "all-ones" AIS pattern towards the TDM circuit, which if long enough in duration is recognized by the receiving TDM device as a fault indication (see Appendix A). International standards place stringent limits on the number of such faults tolerated. Calculations presented in the appendix show that only loss probabilities in the realm of fractions of a percent are relevant for structure-agnostic transport (see Appendix A).

Structure-aware transport regenerates frame alignment signals thus hiding AIS indications resulting from infrequent packet loss. Furthermore, for TDM circuits carrying voice channels the use of packet loss concealment algorithms is possible (such algorithms have been previously described for TDM PWs). However, even structure-aware transport ceases to provide a useful service at about 2 percent loss probability.

RFC 5348 on TCP Friendly Rate Control (TFRC) [RFC5348] provides the following simplified formula for throughput that is used as the basis for TFRC's sending rate control.

$$X_{\text{Bps}} = \frac{S}{R \left(\sqrt{2p/3} + 12 \sqrt{3p/8} p (1+32p^2) \right)}$$

where

X_{Bps} is average sending rate in Bytes per second,
 S is the segment (packet payload) size in Bytes,
 R is the round-trip time in seconds,
 p is the loss probability.

We can use this formula to determine when a TDM PW consumes no more bandwidth than a TCP flow between the same endpoints would consume under the same conditions. Replacing the round-trip delay with twice the one-way delay D , setting the bandwidth to that of the TDM service BW, and the segment size to be the TDM fragment TDM plus 4 Bytes to account for the PWE3 control word, we obtain the following condition for a TDM PW.

$$D < \frac{(TDM + 4)}{BW f(p) / 4}$$

where

D is the one-way delay,
 TDM is the TDM segment size in Bytes,
 BW is TDM service bandwidth in bits per second,
 $f(p) = \sqrt{2p/3} + 12 \sqrt{3p/8} p (1+32p^2)$.

One may view this condition as defining a safe operating envelope for a TDM PW, as a TDM PW that consumes no more bandwidth than a TCP flow would not affect congestion more than were it to be TCP traffic. Under this condition it should hence be safe to mix the TDM PW with congestion-responsive traffic such as TCP, without causing significant additional congestion problems. Were the TDM PW to consume significantly more bandwidth a TCP flow, it could contribute disproportionately to congestion, and its mixture with congestion-

responsive traffic may be inappropriate.

We derived the condition assuming steady-state conditions, and thus two caveats are in order. First, the condition does not specify how to treat a TDM PW that initially satisfies the condition, but is then faced with a deteriorating network environment. In such cases one additionally needs to analyze the reaction times of the responsive flows to congestion events. Second, the derivation assumed that the TDM PW was competing with long-lived TDM flows, because under this assumption it was straightforward to obtain a quantitative comparison with something widely considered to offer a safe response to congestion. Short-lived TCP flows may find themselves disadvantaged as compared to a long-lived TDM PW satisfying the condition. These dynamic cases will be considered in future versions of this draft.

The results are displayed in the accompanying figures (available only in the PDF version of this document). TCP compatible behavior is obtained for the area under curves appropriate for each TDM fragment size.

We see in Figure 1 that a TDM PW carrying an E1 native service (2.048 Mbps) satisfies the condition for all parameters of interest if each packet carries at least $S=512$ Bytes of TDM data. For the SAToP default of 256 Bytes, as long as the one-way delay is less than 10 milliseconds, the loss probability can exceed 0.3 percent. For packets containing 128 or 64 Bytes the constraints are more troublesome, but there are still parameter ranges where the TDM PW consumes less than a TCP flow under similar conditions. Similarly, Figure 2 demonstrates that an E3 native service (34.368 Mbps) with the SAToP default of 1024 Bytes of TDM per packet satisfies the condition for delays up to about 5 milliseconds.

Note that violating the condition for a short amount of time is not sufficient justification for shutting down the TDM PW. While TCP flows react within a round trip time, PW commissioning and decommissioning are time consuming processes that should only be undertaken when it becomes clear that the congestion is not transient. Future versions of this draft will provide guidance as to when a TDM PW should be terminated.

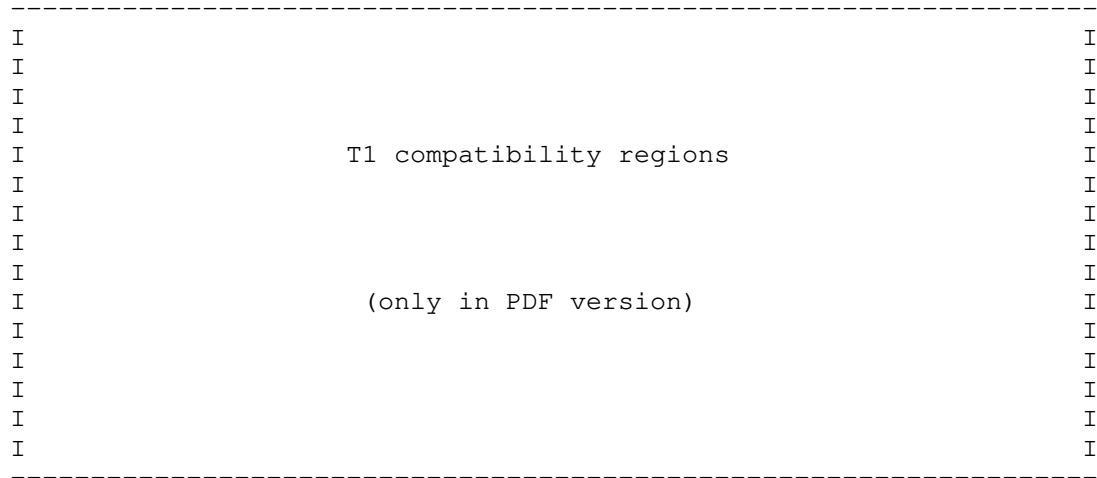


Figure 1 TCP Compatibility areas for T1 SAToP

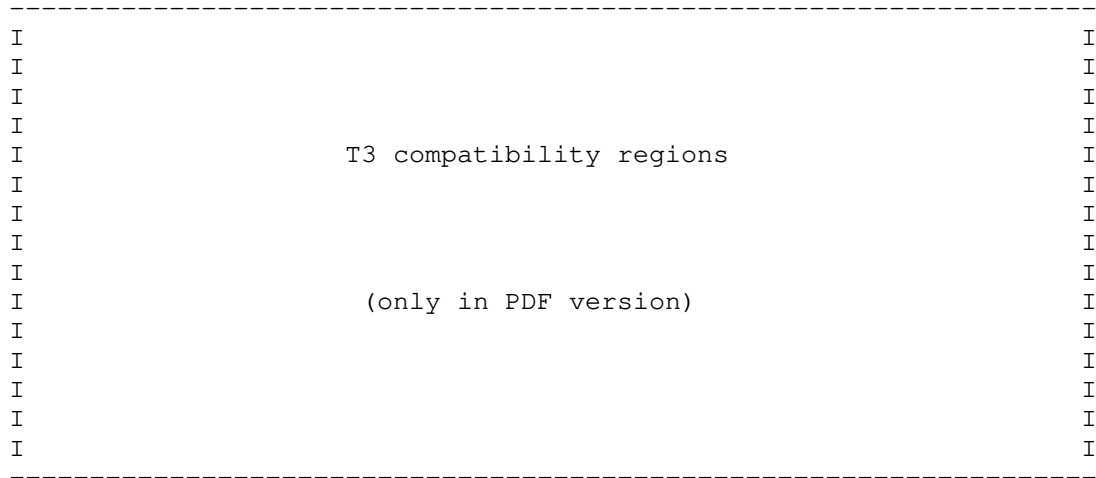


Figure 4 TCP Compatibility areas for T3 SAToP

4. Security Considerations

This document does not introduce any new congestion-specific mechanisms and thus does not introduce any new security considerations above those present for PWs in general.

5. IANA Considerations

This document requires no IANA actions.

6. Informative References

- [RFC2914] Floyd, S., "Congestion Control Principles", BCP 41, RFC 2914, September 2000.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.
- [RFC4553] Vainshtein, A. and YJ. Stein, "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", RFC 4553, June 2006.
- [RFC5086] Vainshtein, A., Sasson, I., Metz, E., Frost, T., and P. Pate, "Structure-Aware Time Division Multiplexed (TDM) Circuit Emulation Service over Packet Switched Network (CESoPSN)", RFC 5086, December 2007.
- [RFC5087] Stein, Y(J)., Shashoua, R., Insler, R., and M. Anavi, "Time Division Multiplexing over IP (TDMoIP)", RFC 5087, December 2007.
- [RFC5348] Floyd, S., Handley, M., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 5348, September 2008.
- [G775] International Telecommunications Union, "Loss of Signal (LOS), Alarm Indication Signal (AIS) and Remote Defect Indication (RDI) defect detection and clearance criteria for PDH signals", ITU Recommendation G.775, October 1998.
- [G826] International Telecommunications Union, "Error Performance Parameters and Objectives for International Constant Bit Rate Digital Paths at or above Primary Rate", ITU Recommendation G.826, December 2002.
- [P862] International Telecommunications Union, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", ITU Recommendation G.826, February 2001.
- [I-D.stein-pwe3-tdm-packetloss]
Stein, Y(J). and I. Druker, "The Effect of Packet Loss on Voice Quality for TDM over Pseudowires", October 2003.

Appendix A. Loss Probabilities for TDM PWs

ITU-T Recommendation G.826 [G826] specifies limits on the Errored Second Ratio (ESR) and the Severely Errored Second Ratio (SESR). For our purposes, we will simplify the definitions and understand an Errored Second (ES) to be a second of time during which a TDM bit error occurred or a defect indication was detected. A Severely Errored Second (SES) is an ES second during which the Bit Error Rate (BER) exceeded one in one thousand (10^{-3}). Note that if the error condition AIS was detected according to the criteria of ITU-T Recommendation G.775 [G826] a SES was considered to have occurred. The respective ratios are the fraction of ES or SES to the total number of seconds in the measurement interval.

For both E1 and T1 TDM circuits, G.826 allows ESR of 4% (0.04), and SESR of 1/5% (0.002). For E3 and T3 the ESR must be no more than 7.5% (0.075), while the SESR is unchanged.

Focusing on E1 circuits, the ESR of 4% translates, assuming the worst case of isolated exactly periodic packet loss, to a packet loss event no more than every 25 seconds. However, once a packet is lost, another packet lost in the same second doesn't change the ESR, although it may contribute to the ES becoming a SES. Assuming an integer number of TDM frames per PW packet, the number of packets per second is given by $\text{packets per second} = 8000 / (\text{frames per packet})$, where prevalent cases are 1, 2, 4 and 8 frames per packet. Since for these cases there will be 8000, 4000, 2000, and 1000 packets per second, respectively, the maximum allowed packet loss probability is 0.0005%, 0.001%, 0.002%, and 0.004% respectively.

These extremely low allowed packet loss probabilities are only for the worst case scenario. In reality, when packet loss is above 0.001%, it is likely that loss bursts will occur. If the lost packets are sufficiently close together (we ignore the precise details here) then the permitted packet loss rate increases by the appropriate factor, without G.826 being cognizant of any change. Hence the worst-case analysis is expected to be extremely pessimistic for real networks. Next we will go to the opposite extreme and assume that all packet loss events are in periodic loss bursts. In order to minimize the ESR we will assume that the burst lasts no more than one second, and so we can afford to lose no more than packet per second packets in each burst. As long as such one-second bursts do not exceed four percent of the time, we still maintain the allowable ESR. Hence the maximum permissible packet loss rate is 4%. Of course, this estimate is extremely optimistic, and furthermore does not take into consideration the SESR criteria.

As previously explained, a SES is declared whenever AIS is detected.

There is a major difference between structure-aware and structure-agnostic transport in this regards. When a packet is lost SAToP outputs an "all-ones" pattern to the TDM circuit, which is interpreted as AIS according to G.775 [G775]. For E1 circuits, G.775 specifies for AIS to be detected when four consecutive TDM frames have no more than 2 alternations. This means that if a PW packet or consecutive packets containing at least four frames are lost, and four or more frames of "all-ones" output to the TDM circuit, a SES will be declared. Thus burst packet loss, or packets containing a large number of TDM frames, lead SAToP to cause high SESR, which is 20 times more restricted than ESR. On the other hand, since structure-aware transport regenerates the correct frame alignment pattern, even when the corresponding packet has been lost, packet loss will not cause declaration of SES. This is the main reason that SAToP is much more vulnerable to packet loss than the structure-aware methods.

For realistic networks, the maximum allowed packet loss for SAToP will be intermediate between the extremely pessimistic estimates and the extremely optimistic ones. In order to numerically gauge the situation, we have modeled the network as a four-state Markov model, (corresponding to a successfully received packet, a packet received within a loss burst, a packet lost within a burst, and a packet lost when not within a burst). This model is an extension of the widely used Gilbert model. We set the transition probabilities in order to roughly correspond to anecdotal evidence, namely low background isolated packet loss, and infrequent bursts wherein most packets are lost. Such simulation shows that up to 0.5% average packet loss may occur and the recovered TDM still conform to the G.826 ESR and SESR criteria.

Appendix B. Effect of Packet Loss on Voice Quality for TDM PWs

Packet loss in voice traffic can cause in gaps or artifacts that result in choppy, annoying or even unintelligible speech. The precise effect of packet loss on voice quality has been the subject of detailed study in the VoIP community, but VoIP results are not directly applicable to TDM PWs. This is because VoIP packets typically contain over 10 milliseconds of the speech signal, while multichannel TDM packets may contain only a single sample, or perhaps a very small number of samples.

The effect of packet loss on TDM PWs has been previously reported [I-D.stein-pwe3-tdm-packetloss]. In that study it was assumed that each packet carried a single sample of each TDM timeslot (although the extension to multiple samples is relatively straightforward and does not drastically change the results). Four sample replacement algorithms were compared, differing in the value used to replace the lost sample:

1. replacing every lost sample by a preselected constant (e.g., zero or "AIS" insertion),
2. replacing a lost sample by the previous sample,
3. replacing a lost sample by linear interpolation between the previous and following samples,
4. replacing the lost sample by STatistically Enhanced INterpolation (STEIN).

Only the first method is applicable to SAToP transport, as structure awareness is required in order to identify the individual voice channels. For structure aware transport, the loss of a packet is typically identified by the receipt of the following packet, and thus the following sample is usually available. The last algorithm posits the LPC speech generation model and derives lost samples based on available samples both before and after each lost sample.

The four algorithms were compared in a controlled experiment in which speech data was selected from English and American English subsets of the ITU-T P.50 Appendix 1 corpus [P.50App1] and consisted of 16 speakers, eight male and eight female. Each speaker spoke either three or four sentences, for a total of between seven and 15 seconds. The selected files were filtered to telephony quality using modified IRS filtering and downsampled to 8 KHz. Packet loss of 0, 0.25, 0.5, 0.75, 1, 2, 3, 4 and 5 percent were simulated using a uniform random number generator (bursty packet loss was also simulated but is not reported here). For each file the four methods of lost sample replacement were applied and the Mean Opinion Score (MOS) was estimated using PESQ [P862]. Figure 5 depicts the PESQ-derived MOS for each of the four replacement methods for packet drop probabilities up to 5%.

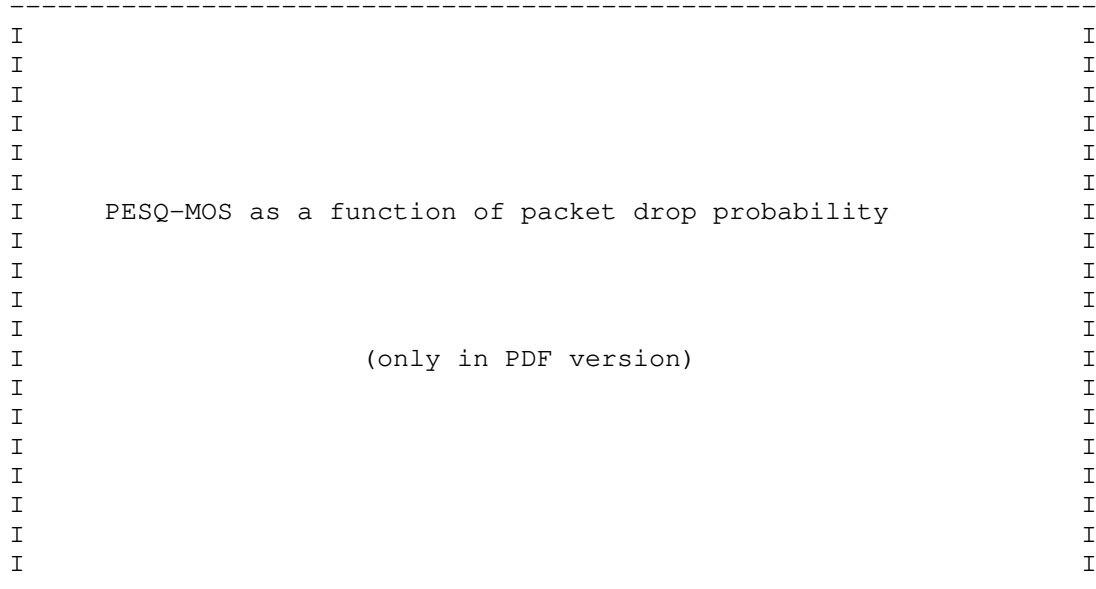


Figure 5 PESQ derived MOS as a function of packet drop probability

For all cases the MOS resulting from the use of zero insertion is less than that obtained by replacing with the previous sample, which in turn is less than that of linear interpolation, which is slightly less than that obtained by statistical interpolation.

Unlike the artifacts speech compression methods may produce when subject to buffer loss, packet loss here effectively produces additive white impulse noise. The subjective impression is that of static noise on AM radio stations or crackling on old phonograph records. For a given PESQ-derived MOS, this type of degradation is more acceptable to listeners than choppiness or tones common in VoIP.

If MOS>4 (full toll quality) is required, then the following packet drop probabilities are allowable:

- zero insertion - 0.05 %
- previous sample - 0.25 %
- linear interpolation - 0.75 %
- STEIN - 2 %

If MOS>3.75 (barely perceptible quality degradation) is acceptable, then the following packet drop probabilities are allowable:

- zero insertion - 0.1 %
- previous sample - 0.75 %
- linear interpolation - 3 %
- STEIN - 6.5 %

If MOS>3.5 (cell-phone quality) is tolerable, then the following packet drop probabilities are allowable:

- zero insertion - 0.4 %
- previous sample - 2 %
- linear interpolation - 8 %
- STEIN - 14 %

Authors' Addresses

Yaakov (Jonathan) Stein
RAD Data Communications
24 Raoul Wallenberg St., Bldg C
Tel Aviv 69719
ISRAEL

Phone: +972 (0)3 645-5389
Email: yaakov_s@rad.com

David L. Black
EMC Corporation
176 South St.
Hopkinton, MA 69719
USA

Phone: +1 (508) 293-7953
Email: david.black@emc.com

Bob Briscoe
BT
B54/77, Adastral Park
Martlesham Heath
Ipswich IP5 3RE
UK

Phone: +44 1473 645196
Email: bob.briscoe@bt.com
URI: <http://bobbbriscoe.net/>

Internet Working Group

Y. Jiang

Internet Draft

Y. Luo

Huawei

Intended status: Standards Track

Expires: April 2014

October 21, 2013

Multi-chassis PON Protection in MPLS
draft-jiang-pwe3-mc-pon-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/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 21, 2013.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

While MPLS is deployed further and further to the access network, a converging network edge point which provides both MPLS and PON access capability appears. To provide resiliency for its services, multi-homing is needed to support PON access in MPLS. This document describes the multi-chassis PON protection architecture in MPLS and also proposes the ICCP extension to support it.

Table of Contents

1.	Conventions used in this document	2
2.	Terminology	3
3.	Introduction	3
3.1.	Multi-chassis PON Application TLVs	5
3.1.1.	PON Connect TLV	5
3.1.2.	PON Disconnect TLV	6
3.1.3.	PON Configuration TLV	6
3.1.4.	PON State TLV	7
4.	Dual Homing protection procedures	8
4.1.	Protection procedure upon PON interface failures	9
4.2.	Protection procedure upon PW failures	9
4.3.	Protection procedure upon the working OLT failure	9
5.	Security Considerations	10
6.	IANA Considerations	10
7.	References	10
7.1.	Normative References	10
7.2.	Informative References	10
8.	Acknowledgments	10
	Authors' Addresses	11

1. Conventions used in this document

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

2. Terminology

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

ICCP Inter-Chassis Communication Protocol

OLT Optical Line Termination

ONU Optical Network Unit

MPLS Multi-Protocol Label Switching

PON Passive Optical Network

3. Introduction

MPLS is extending further and further to the edge of networks, for example, the seamless MPLS use cases as described in [SEAMLESS], and the MS-PW with PON access use case as described in [RFC6456], all show that MPLS is approaching the access networks.

Passive Optical Network (PON) can provide high bandwidth of 1Gbps or even 10Gbps, and provide support of access for dozens to more than one hundred subscribers at the same time. A huge number of PON access networks have been deployed over the last few years with the wide spread of FTTx technology.

With the fast growth of mobile data traffic, more and more LTE small cells and Wi-Fi hotspots will be deployed in the future. How to backhaul a large number of small cells or hotspots will pose a great challenge to mobile service providers.

PON access technology has the following advantages:

- saving trunk fibers with its point-to-multipoint physical topology;
- High bandwidth capability up to 10Gbps;
- Low Total Cost of Ownership (TCO).

PON also provides synchronization features, e.g., SyncE and IEEE1588 functionality, which can fulfill synchronization needs of mobile backhaul services. Some optical layer of protection mechanisms, such as Type B protection and Type C protection are also specified [G983.1] to avoid single point of failure in the access.

Therefore, PON may play a greater role in the access end for the mobile backhaul networks. Providing OLTs with MPLS functionality further facilitates multi-service convergence.

Type B protection architecture is an economical PON resiliency mechanism, where the working OLT and the working link between the working splitter and the working OLT (i.e., the working fiber) is protected by a redundant protection OLT and a redundant fiber between the working splitter and the protection OLT. This is different from the more complex and costly Type C protection architecture where working splitter and the working fibers from ONUs to the working splitter are further protected. Figure 1 demonstrates a typical scenario of Type B PON protection.

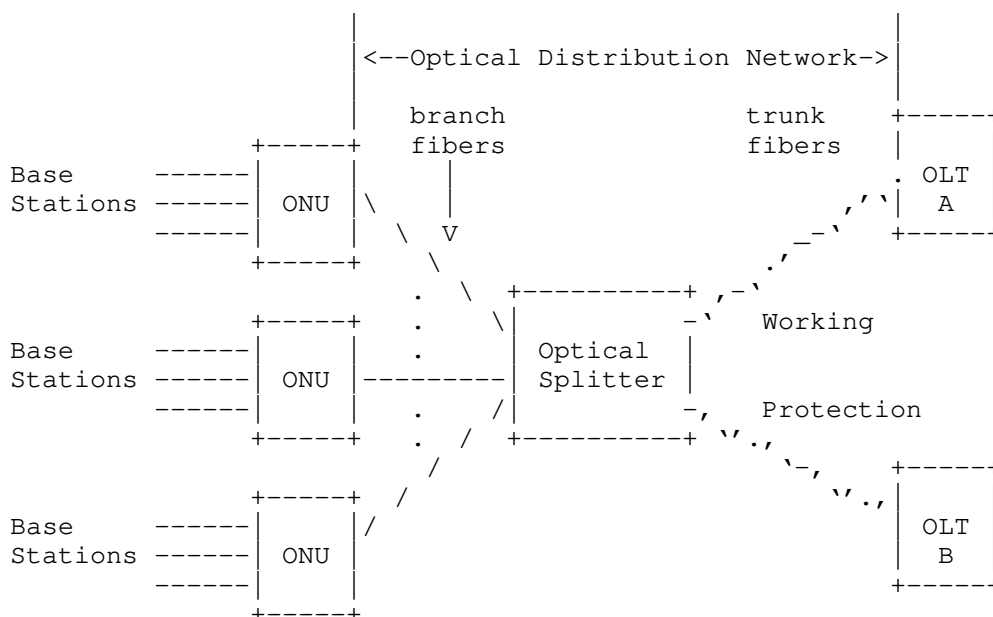


Figure 1 Type B PON protection Architecture

Though the above PON architecture provides redundancy in its physical topology, some standard mechanisms are needed to exchange PON link status and network status between OLTs in a Redundancy Group (RG) so that protection and restoration can be done reliably, especially when the OLTs also support MPLS. Thus there is a need for Multi-chassis PON protection protocol in MPLS.

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

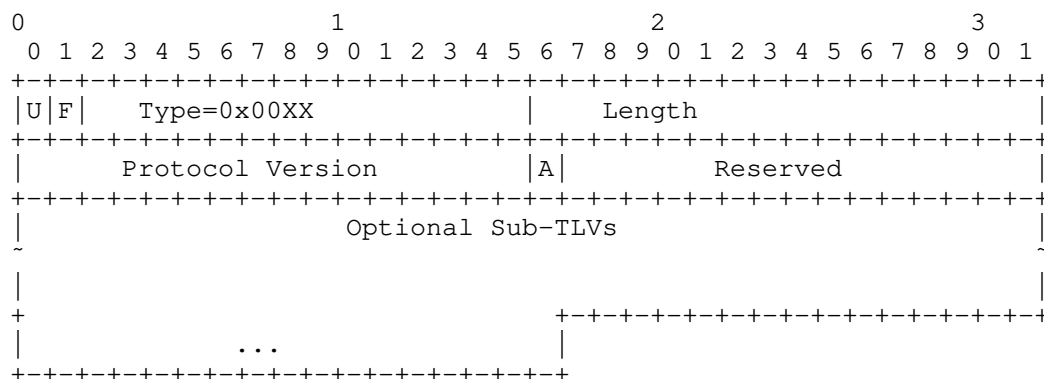
This document proposes the extension of ICCP to support Multi-chassis PON protection in MPLS.

3.1. Multi-chassis PON Application TLVs

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

3.1.1. PON Connect TLV

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



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

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

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

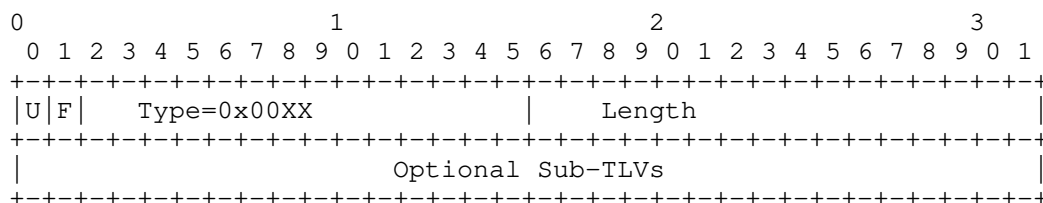
- Protocol Version, the version of this PON specific protocol for the purposes of inter-chassis communication. This is set to 0x0001.

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

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

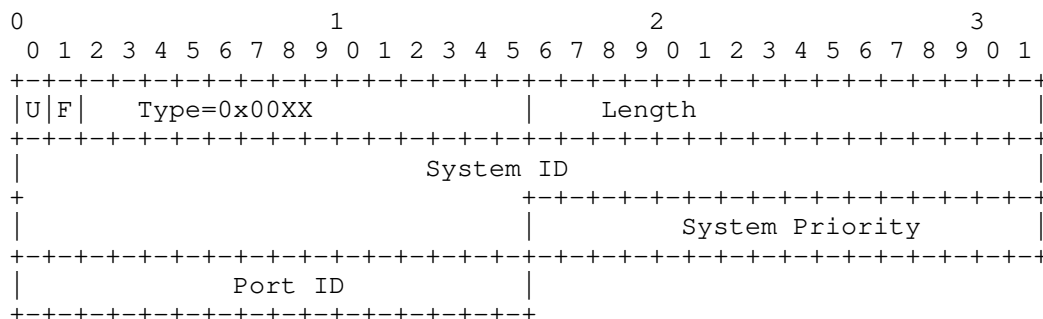
3.1.2. PON Disconnect TLV

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



- U and F Bits, both are set to 0.
- Type, set to 0x00XX for "PON Disconnect TLV".
- Length, Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.
- Optional Sub-TLVs, there are no optional Sub-TLVs defined for this version of the protocol.

3.1.3. PON Configuration TLV

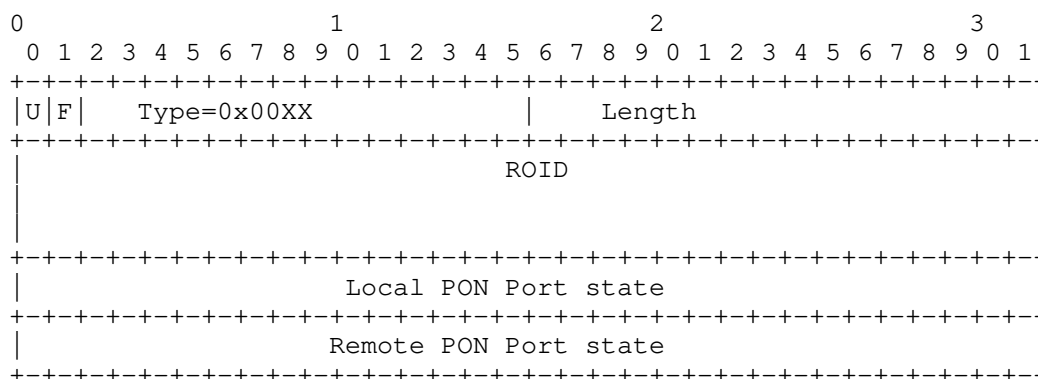


- U and F Bits, both are set to 0.
- Type, set to 0x00XX for "PON Configuration TLV".

- Length, Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.
- System ID, 6 octets encoding the System ID used by the OLT, which is a MAC address.
- System Priority, 2 octets encoding the System Priority.
- Port ID, 2 octets PON Port ID.

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

3.1.4. PON State TLV



- U and F Bits, both are set to 0.
- Type, set to 0x00XX for "PON State TLV"
- Length, Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.
- ROID, as defined in the ROID section of [ICCP].
- Local PON Port State, the status of the local PON port as determined by the sending OLT (PE). The last bit is defined as Fault indication of the PON Port associated with this PW.
- Remote PON Port State, the status of the remote PON port as determined by the remote peer of the sending OLT (PE). The last bit is defined as Fault indication of the PON Port associated with this PW.

4. Dual Homing protection procedures

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

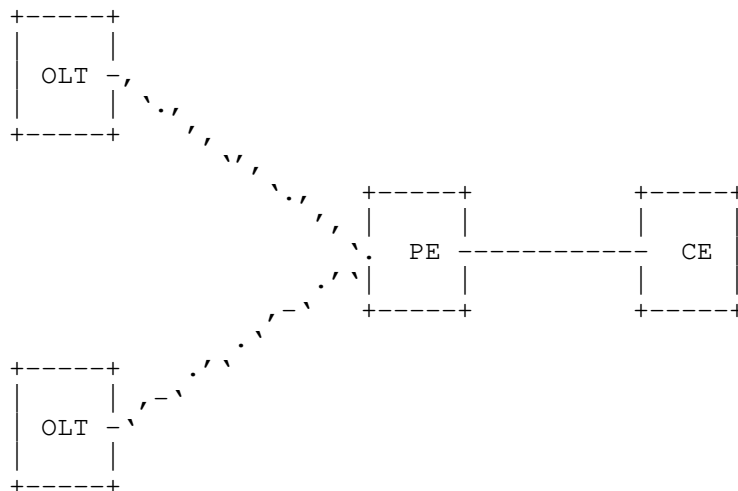


Figure 2 An MPLS network with a single PE

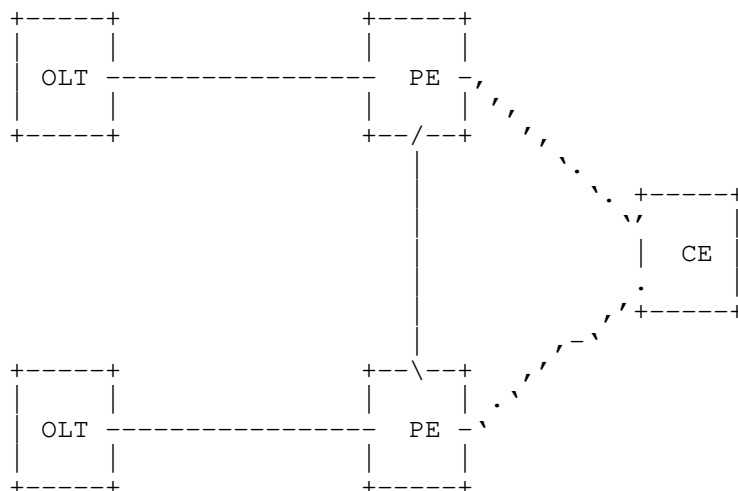


Figure 3 An MPLS network with dual home PEs

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

4.1. Protection procedure upon PON interface failures

When a fault is detected on a working PON link, a working OLT MUST turn off its associated PON interface and MUST send an LDP notification message with a forward defect indication and with the Request Switchover bit being set to its peer PE on the remote end of the PW. At the same time, the working OLT MUST send an ICCP message with PON State TLV to notify the backup OLT of the PON fault. Upon receiving a PON state TLV where Local PON Port state is False, an OLT in the protection mode MUST activate the protection PON link in the protection group.

4.2. Protection procedure upon PW failures

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

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

Upon receiving a PON state TLV where Local PON Port state is False, the backup OLT MUST activate its optical interface to the backup fiber. At the same time, the backup OLT MUST send a PW redundancy message to the remote PE, so that traffic can be switched to the backup PW.

4.3. Protection procedure upon the working OLT failure

If the backup OLT lost connection to the working OLT, it MUST activate its optical interface to the back fiber and activate the specific backup PW upon receiving a PW redundancy message from its remote PE with the Request Switchover bit being set, so that traffic can be reliably switched to the protection link and the backup PW.

5. Security Considerations

Security considerations as described in [ICCP] apply.

6. IANA Considerations

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

0x00X0	PON Connect TLV
0x00X1	PON Disconnect TLV
0x00X2	PON Configuration TLV
0x00X3	PON State TLV

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

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

7.2. Informative References

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

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

[ICCP] Martini, L. and et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11, Work in progress

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

8. Acknowledgments

TBD.

Authors' Addresses

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

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

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: January 13, 2014

Mingui Zhang
Peng Zhou
Huawei
July 12, 2013

ICCP Application TLVs for VPN Route Label Sharing
draft-zhang-pwe3-iccp-label-sharing-00.txt

Abstract

This document defines TLVs under Inter-Chassis Communication Protocol (ICCP) to include a new application: Label Sharing for Fast PE Protection. Egress PEs in the same Redundant Group utilize the ICCP connection to negotiate the "VPN route label" and the "BGP next hop" for each VPN.

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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Conventions used in this document	3
1.2. Terminology	3
2. Label Sharing TLVs in ICCP	3
2.1. Label Sharing Connect TLV	3
2.2. Label Sharing Disconnect TLV	4
2.2.1. Label Sharing Disconnect Cause TLV	5
2.3. Label Sharing Application Data TLVs	6
2.3.1. Service Name TLV	7
2.3.2. VPN Label TLV	7
2.3.3. vNH TLV	8
3. Security Considerations	9
4. IANA Considerations	10
5. References	10
5.1. Normative References	10
5.2. Informative References	10
Author's Addresses	11

1. Introduction

It's common for Service Providers (SPs) to connect one CE to multiple PEs for the sake of reliability. In [LS], this feature is leveraged to realize a method for fast PE protection. There, egress PEs in the same Redundant Group (RG) share the same "VPN route label" for one VPN. These egress PEs use a virtual Next Hop (vNH) as their "BGP next hop". Primary and backup LDP LSP tunnels ended at the vNH are set up using IGP FRR [LFA] [MRT]. When the PLR redirects the failure affected packet to the backup egress PE, the VPN route label encapsulated in the packet can be recognized by the backup egress PE and the packet will be delivered naturally.

This document extends ICCP to include the "label sharing" method as a new application. The connection of ICCP is leveraged to synchronize the label and BGP next hop of each VPN for the PEs in one RG. TLVs are defined in the next section.

1.1. Conventions used in this document

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

1.2. Terminology

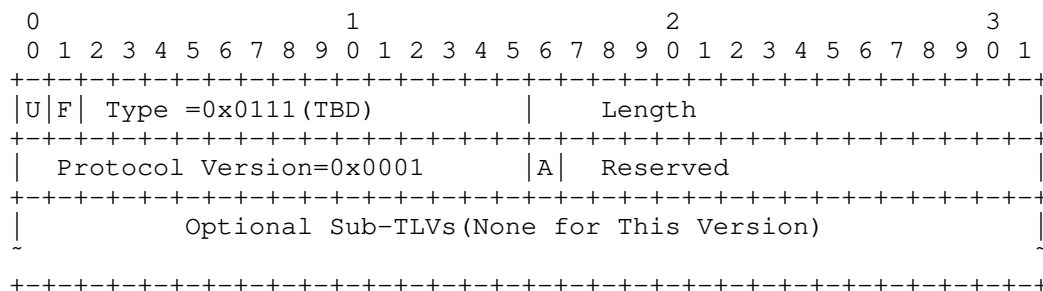
vNH: virtual Next Hop
FRR: Fast ReRouting
PLR: Point of Local Repair

2. Label Sharing TLVs in ICCP

This section specifies the ICCP Connect, Disconnect and Application Data TLVs to be used by egress PEs for the label sharing application.

2.1. Label Sharing Connect TLV

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



- U and F Bits

Both are set to 0.

- Type

set to 0x0111 (TBD) for "Label Sharing 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 Label Sharing Connect TLV from the recipient. Otherwise, set to 0.

- Reserved

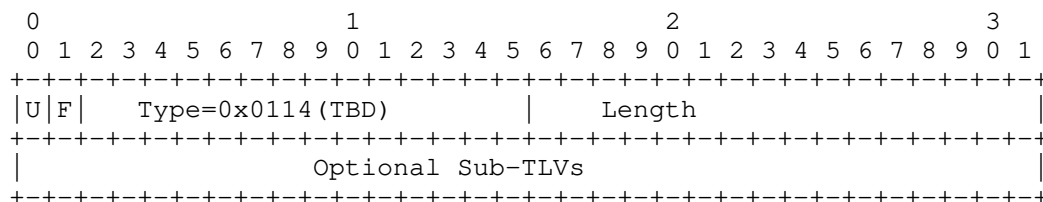
Reserved for future use.

- Optional Sub-TLVs

There are no optional Sub-TLVs defined for this version of the protocol.

2.2. Label Sharing Disconnect TLV

This TLV is included in an RG Disconnect Message as the "Disconnect Code TLV" (See Section 6.3 of [ICCP]). It indicates that the connection for the Label Sharing application is to be terminated.



- U and F Bits

Both are set to 0.

- Type

set to 0x0114 (TBD) for "Label Sharing Disconnect TLV"

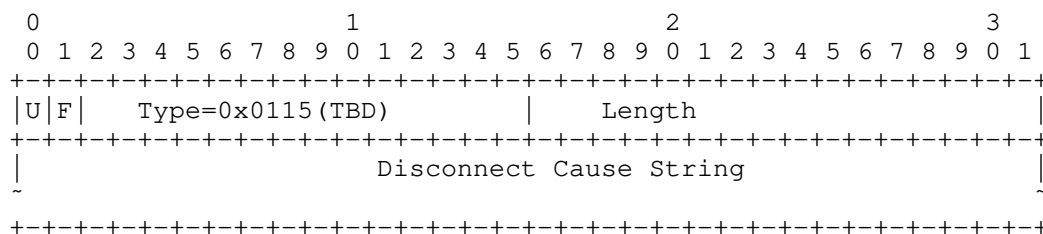
- Length

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 "Label Sharing Disconnect Cause" TLV defined next:

2.2.1. Label Sharing Disconnect Cause TLV



- U and F Bits

Both are set to 0.

- Type

set to 0x0115 (TBD) for "Label Sharing Disconnect Cause TLV"

- Length

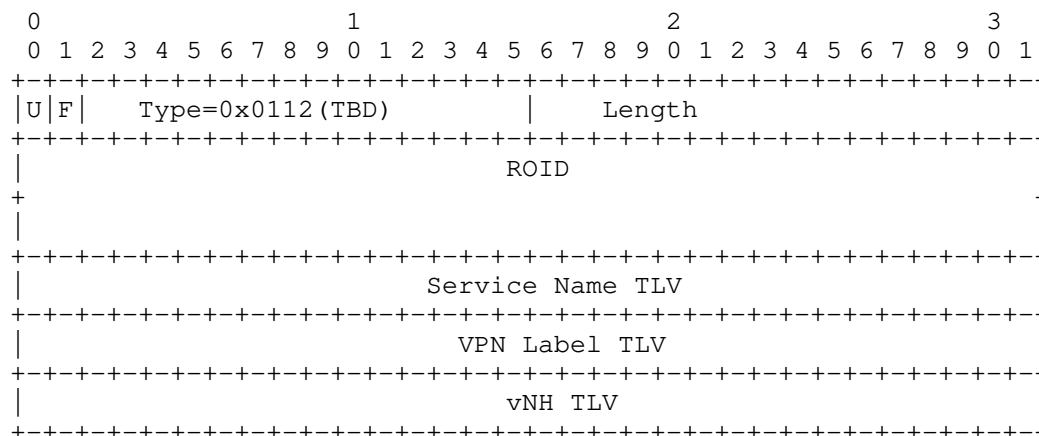
Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.

- Disconnect Cause String

Variable length string specifying the reason for the disconnect. Used for network management.

2.3. Label Sharing Application Data TLVs

The following TLVs are included in the RG Application Data message to deliver the information that need be synchronized among RG members.



- U and F Bits

Both are set to 0.

- Type

set to 0x0112 (TBD) for "Label Sharing Information TLV"

- Length

Length of the MAC address, which is 6 octets.

- ROID

As defined in the ROID section of [ICCP].

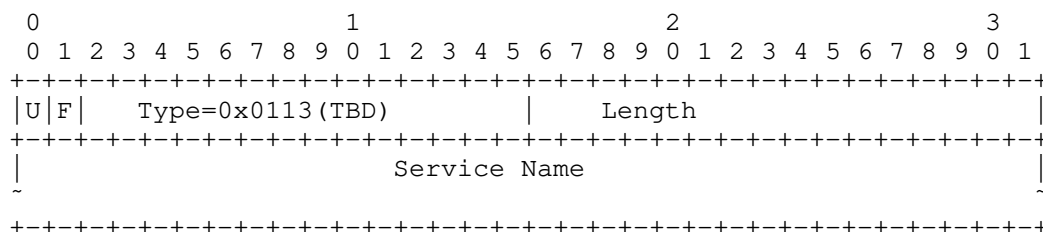
- Sub-TLVs

i Service Name TLV

ii VPN Label TLV

iii vNH TLV

2.3.1. Service Name TLV



- U and F Bits

Both are set to 0.

- Type

set to 0x0113 (TBD) for "Service Name TLV"

- Length

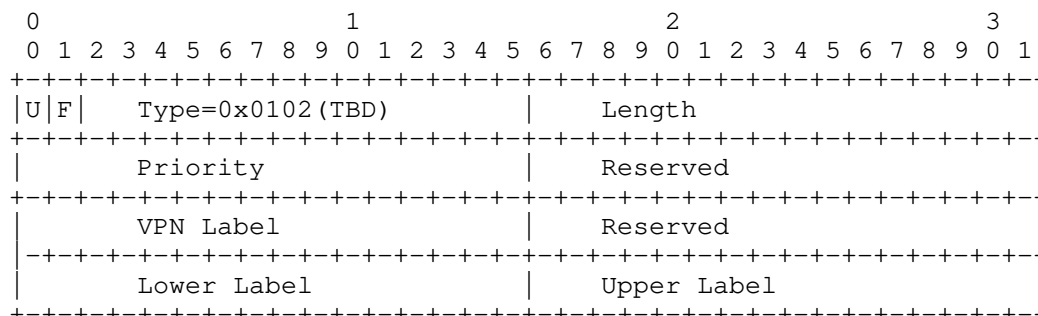
Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.

- Service Name

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

2.3.2. VPN Label TLV

The PE with the highest priority (with its MAC address as the tiebreaker) assigns the shared VPN label for a VPN. In a well configured network, PEs in the same RG will be configured to have the same range of VPN labels for sharing. When the ranges of the VPN labels are different, the VPN label is chosen from the intersection of the ranges.



- U and F Bits

Both are set to 0.

- Type

set to 0x0112 (TBD) for "VPN Label TLV"

- Length

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

- Priority

The priority that the sender has for the VPN label in this TLV. When there are more than one sender who has the highest priority, the MAC address of the sender used as the tiebreaker.

- Reserved

Reserved for future use.

- VPN Label

The VPN label to be shared among the RG.

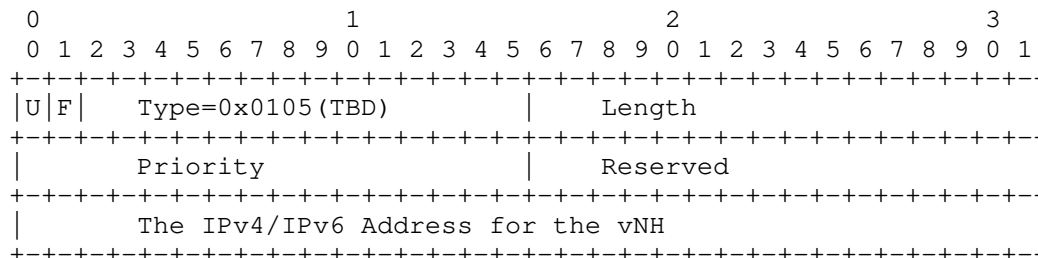
- Lower/Upper Label

The lower/upper bound of a valid VPN label.

2.3.3. vNH TLV

When a VPN route is distributed to ingress PEs by BGP, the IP address of the vNH will be used as the BGP next hop. Thus, tunnels terminated at the vNH will be set up. The PE with the highest priority (with its

MAC address as the tiebreaker) determines the IP address of the vNH.



- U and F Bits

Both are set to 0.

- Type

set to 0x0105 (TBD) for "Service Name TLV"

- Length

Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields. Lengths for the IPv4 and IPv6 Addresses TLVs are different.

- Priority

The priority that the sender has for the IPv4/IPv6 address for the vNH in this TLV. When there are more than one sender who has the highest priority, the MAC address of these senders will be used as the tiebreaker.

- Reserved

Reserved for future use.

- IPv4/IPv6 Address for the vNH

The IPv4/IPv6 address that the sender wants the vNH to use. The IPv4/IPv6 address of vNH TLV sent out by sender with the highest priority will be used as the IPv4/IPv6 address of the vNH by all the PEs in the same RG.

3. Security Considerations

This document raises no new security issues.

4. IANA Considerations

The types used by the application TLVs defined in Section 3 should be assigned.

5. References

5.1. Normative References

[ICCP] L. Martini, S. Salam, et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11.txt, work in progress.

[LS] M. Zhang, P. Zhou, "Label Sharing for Fast PE Protection", draft-zhang-l3vpn-label-sharing-00.txt, work in progress.

5.2. Informative References

[LFA] Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

[MRT] A. Atlas, Ed., R. Kebler, et al, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-02.txt, work in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Peng Zhou
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: Jewpon.zhou@huawei.com

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: April 15, 2014

Mingui Zhang
Huaifeng Wen
Huawei
October 12, 2013

STP Application of ICCP
draft-zhang-pwe3-iccp-stp-00.txt

Abstract

Inter-Chassis Communication Protocol (ICCP) supports the inter-chassis redundancy mechanism which achieves high network availability.

In this document, the PEs in a Redundant Group (RG) running ICCP are used to offer multi-homed connectivity to Spanning Tree Protocol (STP) networks. The ICCP TLVs for the STP application are defined, therefore PEs from the RG can make use of these TLVs to synchronize the state and configuration data.

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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
1.1. Conventions used in this document	3
1.2. Terminology	3
2. The Use Case Scenario	3
2.1. Virtual Root Bridge	4
3. Spanning Tree Protocol Application TLVs	4
3.1. STP Connect TLV	4
3.2. STP Disconnect TLV	5
3.2.1. STP Disconnect Cause TLV	6
3.3. STP Config TLVs	7
3.3.1. STP System Config	7
3.3.2. STP Topology Changed Instances	8
3.3.3. STP CIST Root Time	8
3.3.4. STP MSTI Root Time	9
3.3.5. STP Region Name	10
3.3.6. STP Revision Level	11
3.3.7. STP Instance Priority	11
3.3.8. STP Configuration Digest	12
3.4. STP Synchronization Request TLV	13
3.5. STP Synchronization Data TLV	14
4. Security Considerations	15
5. IANA Considerations	15
6. References	15
6.1. Normative References	15
6.2. Informative References	15
Author's Addresses	17

1. Introduction

Inter-Chassis Communication Protocol (ICCP) specifies a multi-chassis redundant mechanism, which enables PEs located in multi-chassis to act as a single Redundant Group (RG).

When a bridge network running Spanning Tree Protocol (STP) is connected to a RG, the RG members should pretend to be a single root bridge to participate the operations of the STP. STP relevant information need be exchanged and synchronized among the RG members. ICCP TLVs for the Spanning Tree Protocol application are specified for this purpose.

1.1. Conventions used in this document

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

1.2. Terminology

STP: Spanning Tree Protocol
MSTP: Multiple Spanning Tree Protocol
DSLAM: Digital Subscriber Line Access Multiplexer
MST: Multiple Spanning Trees
CIST: Common and Internal Spanning Tree
MSTI: Multiple Spanning Tree Instance
BPDU: Bridge Protocol Data Unit

In this document, unless otherwise explicitly noted, when the term STP is used, it also covers MSTP.

2. The Use Case Scenario

It is a common case that an RG is connected to a bridge network where STP is running. For example, geographically dispersed DSLAMs of a Broadband Network may be connected by an RG. These DSLAMs constitute a typical STP network. For the sake of network resilience, it is reasonable to connect each RG member to this bridge network. The scenario in Figure 2.1 illustrates this kind of connection.

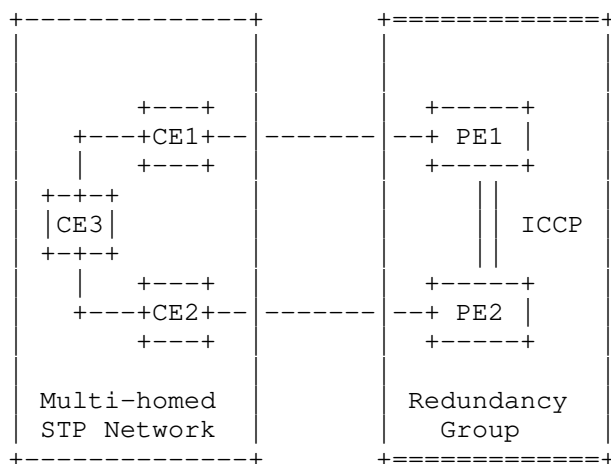


Figure 2.1: A STP network is multi-homed to an Redundant Group.

2.1. Virtual Root Bridge

With ICCP, the whole RG will be virtualized to be a single bridge. The RG pretends that the ports connected to the STP network are from the same bridge. All these ports emit configuration BPDU with the highest root priority to trigger the construction of the spanning tree. In this way, the STP will always broken a loop within the multi-homed STP network.

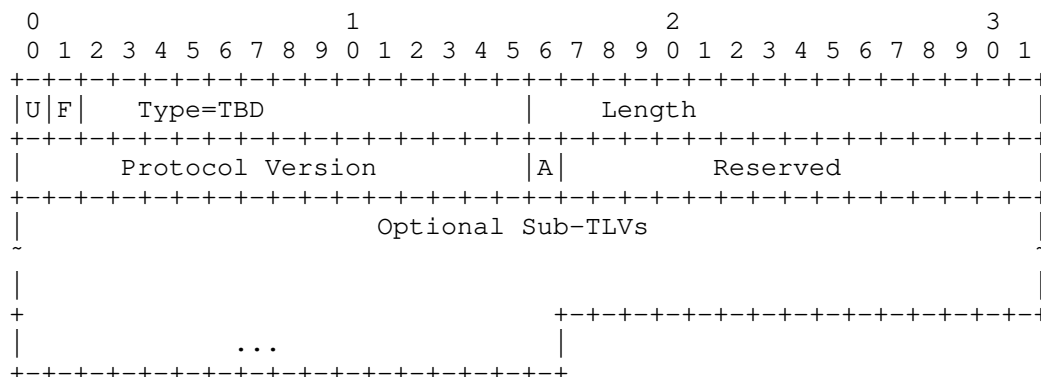
Each RG member has its BridgeIdentifier (the MAC address). The least significant one is elected as the BridgeIdentifier of the 'virtualized root bridge'.

3. Spanning Tree Protocol Application TLVs

This section discusses the ICCP TLVs for the Spanning Tree Protocol application.

3.1. STP Connect TLV

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



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP 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 STP Connect TLV from the recipient. Otherwise, set to 0.

- Reserved

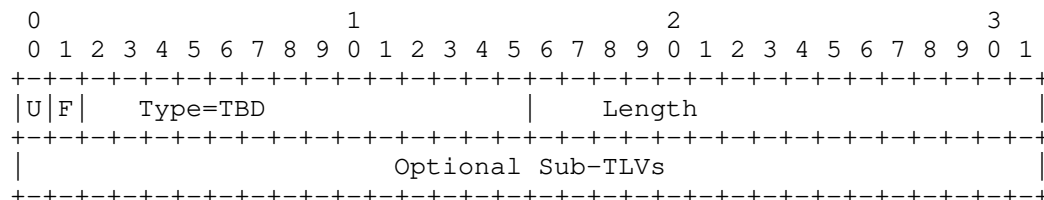
Reserved for future use.

- Optional Sub-TLVs

There are no optional Sub-TLVs defined for this version of the protocol.

3.2. STP Disconnect TLV

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



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Disconnect TLV"

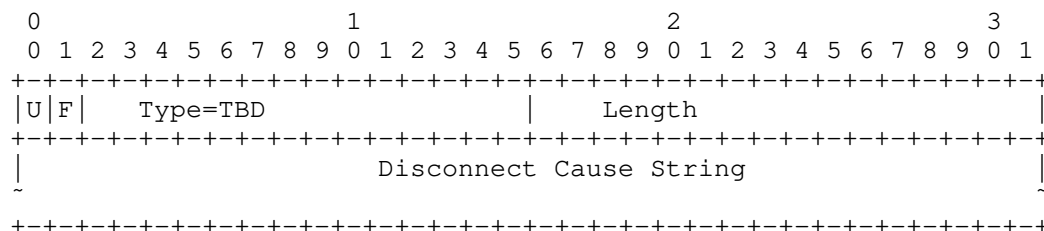
- Length

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 "STP Disconnect Cause" TLV defined next:

3.2.1. STP Disconnect Cause TLV



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Disconnect Cause TLV"

- Length

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

- Disconnect Cause String

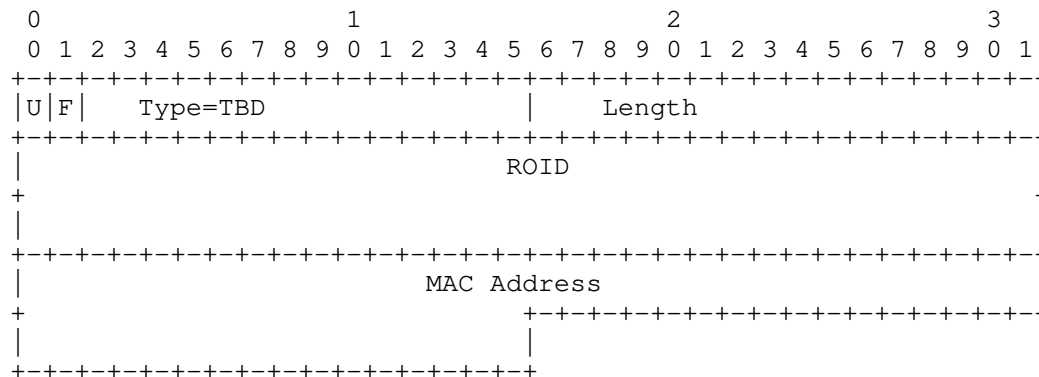
Variable length string specifying the reason for the disconnect. Used for network management.

3.3. STP Config TLVs

The STP Config TLVs are sent in the RG Application Data message. When a STP Config TLV is received by a peering RB member, it SHOULD synchronize the configuration information contained in the TLV. TLVs specified from section 3.3.1 through section 3.3.9 contains such kind of configuration information.

3.3.1. STP System Config

This TLV announces the local node's STP System Parameters to the RG peers.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP System Config"

- Length

Length of the MAC address, which is 6 octets.

-ROID

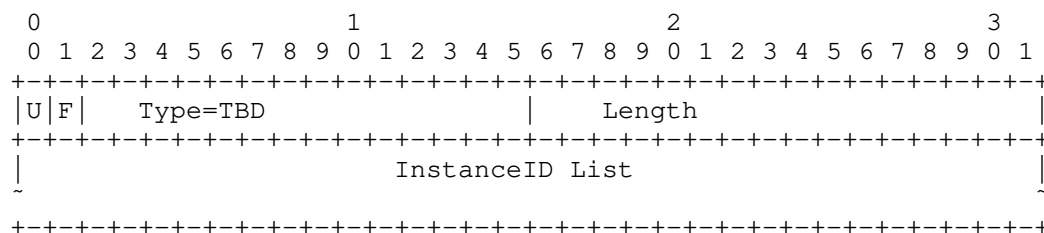
As defined in the ROID section of [ICCP].

- MAC Address

The MAC address of the sender. This MAC address is set to the BridgeIdentifier of the sender, as defined in [802.1q] section 13.23.2. The the least significant unsigned BridgeIdentifier is used as the MAC address of the Virtual Root Bridge mentioned in Section 2.1.

3.3.2. STP Topology Changed Instances

This TLV is used to report the Topology Changed Instances to other members in the RG. The receiver RG member SHOULD enforce the Topology Change to its port connected to the STP network, including the flush out of MAC addresses relevant to the instances listed in this TLV.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Topology Changed Instances"

- Length

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

- InstanceID List

The list of the instances whose topology is changed as indicated by the Topology Change Notification (TCN) Messages as specified in [802.1q] section 13.14.

3.3.3. STP CIST Root Time

This TLV is used to report the Value of CIST Root Time 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=TBD   |   Length   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   MaxAge   |   MessageAge   |   FwdDelay   |   HelloTime   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| RemainingHops |
+---+---+---+---+---+---+

```

- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP CIST Root Time"

- Length

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

- MaxAge

The Maximum Age of this TLV.

- MessageAge

The actual age of this TLV.

- FwdDelay

The delay before the port enters the forwarding status.

- HelloTime

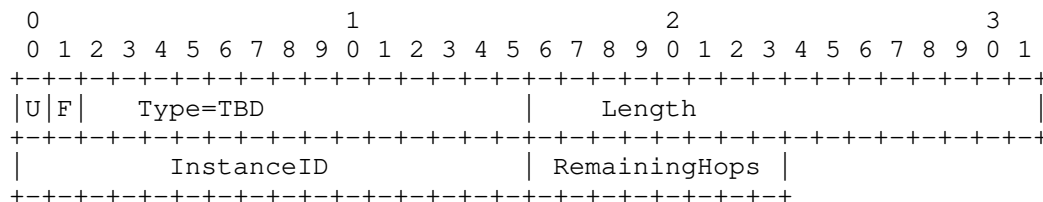
The interval between two continuous configuration BPDUs.

- RemainingHops

The remaining hops of this TLV

3.3.4. STP MSTI Root Time

This TLV is used to report the Value of MSTI Root Time to other members in the RG.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP MSTI Root Time"

- Length

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

- InstanceID

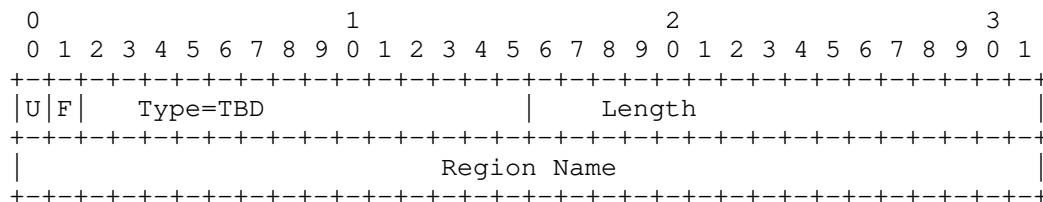
The instance identification number of the MSTI.

- remainingHops

The remaining hops of this TLV

3.3.5. STP Region Name

This TLV is used to report the Value of Region Name to other members in the RG.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Region Name"

- Length

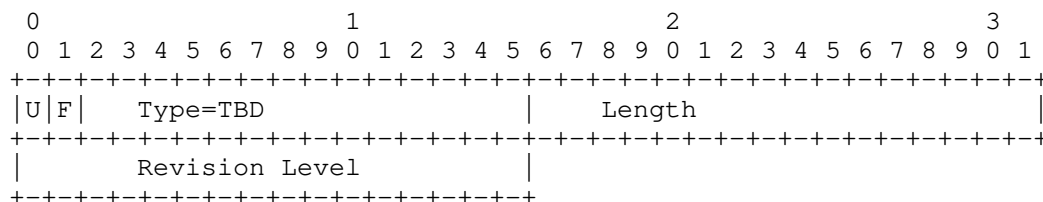
Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.

- Region Name

The Name of the MST Region.

3.3.6. STP Revision Level

This TLV is used to report the Value of Revision Level to other members in the RG.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Revision Level"

- Length

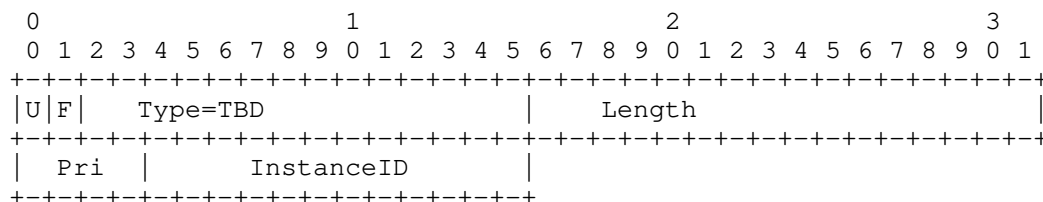
Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.

- Revision Level

The Revision Level as specified in [802.1q] section 3.21;

3.3.7. STP Instance Priority

This TLV is used to report the Value of Instance Priority to other members in the RG.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Instance Priority"

- Length

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

- Pri

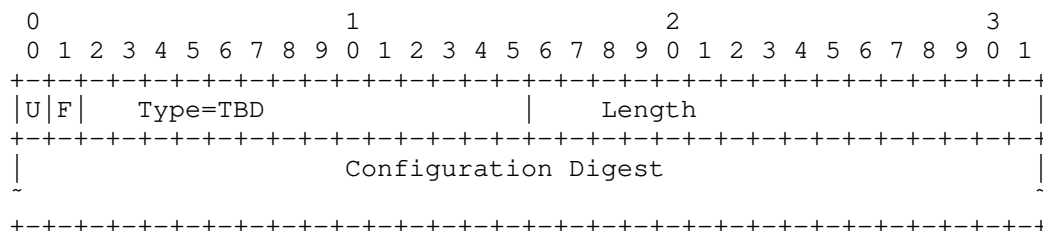
The Instance Priority

- InstanceID

The instance identification number of the MSTI.

3.3.8. STP Configuration Digest

This TLV is used to report the Value of STP VLAN Instance Mapping to other members in the RG.



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP Configuration Digest"

- Length

Length of the STP Configuration Digest which is 16 octets.

- Configuration Digest

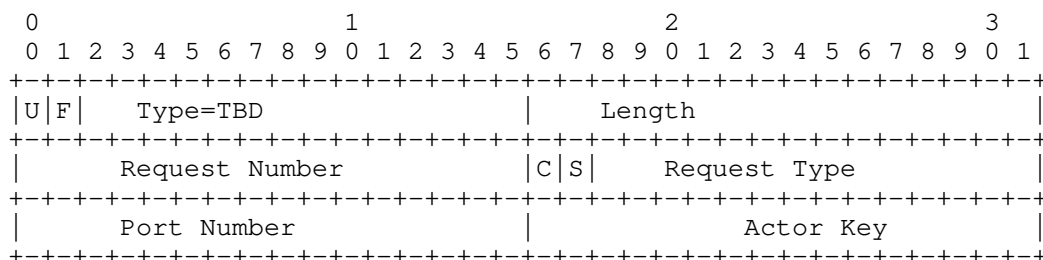
As specified in [802.1q] section 13.7.

3.4. STP Synchronization Request TLV

The STP 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

The format of the TLV is as follows:



- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP 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 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 STP 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 Port Data
0x3FFF	Request All Data

- Port Number

2 octets. When Request Type field is set to 'Request Port Data', this field encodes the STP Port Number for the requested port. When the value of this field is 0, it denotes that all ports, whose STP Key is specified in the "Actor Key" field, are being requested.

- Actor Key

2 octets. STP Actor key for the corresponding port. When the value of this field is 0 (and the Port Number field is 0 as well), it denotes that information for all ports in the system is being requested.

3.5. STP Synchronization Data TLV

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

This TLV has the following format:

- U and F Bits

Both are set to 0.

- Type

set to TBD for "STP 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 "STP 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

4. Security Considerations

This document raises no new security issues.

5. IANA Considerations

The types used by the application TLVs defined in Section 3 should be assigned.

6. References

6.1. Normative References

[ICCP] L. Martini, S. Salam, et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11.txt, work in progress.

6.2. Informative References

[802.1q] "IEEE Standard for Local and Metropolitan Area Networks---

Virtual Bridged Local Area Networks.". IEEE Std 802.1 Q-2005,
May 19, 2006.

Author's Addresses

Mingui Zhang
Huawei

Email: zhangmingui@huawei.com

Huafeng Wen
Huawei

Email: wenhuafeng@huawei.com