

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 14, 2012

J. Dong
H. Wang
Huawei Technologies
June 12, 2012

Pseudowire Redundancy on S-PE
draft-dong-pwe3-redundancy-spe-02

Abstract

This document describes Multi-Segment Pseudowire (MS-PW) protection scenarios in which the pseudowire redundancy is provided on the Switching-PE (S-PE). Signaling of preferential forwarding defined in [I-D.ietf-pwe3-redundancy-bit] is reused.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on December 14, 2012.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. PW Redundancy on S-PE	3
3. S-PE Operations	4
4. VCCV Considerations	6
5. IANA Considerations	6
6. Security Considerations	6
7. Acknowledgements	6
8. References	6
8.1. Normative References	6
8.2. Informative References	7
Authors' Addresses	7

1. Introduction

[I-D.ietf-pwe3-redundancy] and [I-D.ietf-pwe3-redundancy-bit] describe Pseudowire (PW) redundancy mechanism for scenarios where a set of redundant PWs terminate on either provider edge (PE) nodes in single-segment pseudowire (SS-PW) [RFC3985] applications, or on terminating provider edge (T-PE) nodes in multi-segment pseudowire (MS-PW) [RFC5659] applications. This document describes the scenarios where PW redundancy is provided on S-PEs of MS-PW. Signaling of preferential forwarding defined in [I-D.ietf-pwe3-redundancy-bit] is reused for these scenarios, and operations on S-PEs are specified.

2. PW Redundancy on S-PE

In some MS-PW deployment scenarios, PW redundancy may need to be provided on S-PE. This section gives some examples of PW redundancy on S-PE.

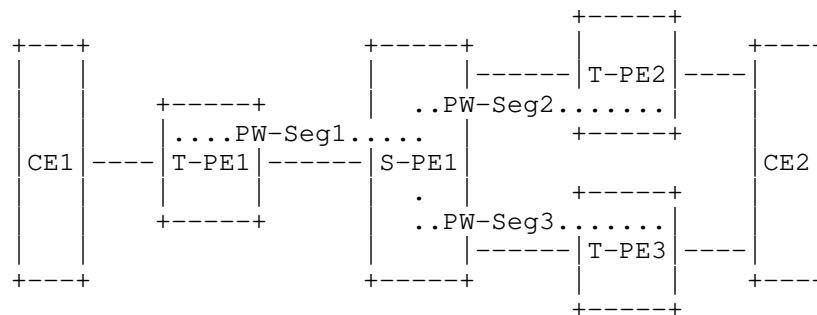


Figure 1. MS-PW Redundancy on S-PE

As illustrated in Figure 1, CE1 is connected to T-PE1 while CE2 is dual-homed to T-PE2 and T-PE3. T-PE1 is connected to S-PE1 only, and S-PE1 is connected to T-PE2 and T-PE3. The MS-PW is switched on S-PE1, and PW-Seg2 and PW-Seg3 provides resiliency on S-PE1 for failure of T-PE2 or T-PE3 or the connected ACs. PW-Seg2 is selected as primary PW segment, and PW-Seg3 is secondary PW segment.

MS-PW redundancy on S-PE is beneficial for scenario in Figure 1 since on T-PE1 side it may be impossible to provide PW redundancy, especially when the PW-Seg1 between T-PE1 and S-PE1 is statically configured. And with PW redundancy on S-PE, the number of PW segments needed between T-PE1 and S-PE1 is only half of the number of PW segments needed for end-to-end MS-PW redundancy. Also PW redundancy on S-PE could provide faster protection switching than end-to-end protection switching of MS-PW.

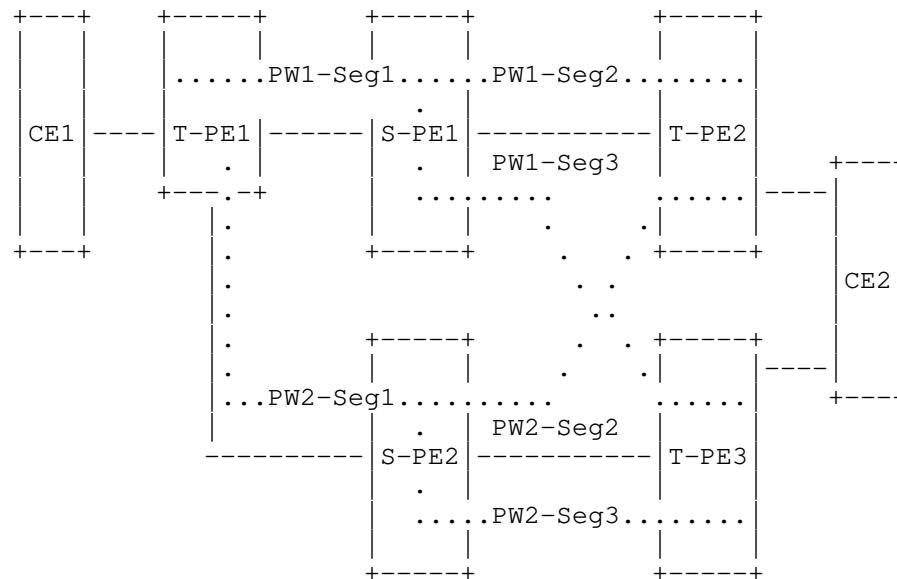


Figure 2. MS-PW Redundancy on S-PE with S-PE protection

As illustrated in Figure 2, CE1 is connected to T-PE1 while CE2 is dual-homed to T-PE2 and T-PE3. T-PE1 is connected to S-PE1 and S-PE2, both S-PE1 and S-PE2 are connected to T-PE2 and T-PE3. There are two MS-PWs which are switched at S-PE1 and S-PE2 respectively to provide S-PE node protection. For MS-PW1, the S-PE1 provides resiliency using PW1-Seg2 and PW1-Seg3. For MS-PW2, the S-PE2 provides resiliency using PW2-Seg2 and PW2-Seg3. MS-PW1 is the primary PW and PW1-Seg2 is the primary PW segment.

MS-PW redundancy on S-PE is beneficial for scenario in Figure 2 since it reduces the number of end-to-end MS-PWs required for both T-PE and S-PE protection. Also PW redundancy on S-PE could provide faster protection switching than end-to-end protection switching of MS-PW.

3. S-PE Operations

When S-PE redundancy is provisioned, it is necessary that S-PE could perform protection switching according to the status change of PW segments and announce appropriate PW status to adjacent PEs. Signaling of preferential forwarding defined in [I-D.ietf-pwe3-redundancy-bit] is reused for these scenarios, and operation on S-PE is specified as below.

For scenario of Figure 1, assume the AC from CE2 to T-PE2 is active. if S-PE1 knows PW-Seg1 is in "PW forwarding" State, it would

advertise "Preferential Forwarding" status bit of "Active" on both PW-Seg2 and PW-Seg3. T-PE2 advertises the preferential status "Active" and T-PE3 advertises the preferential status "Standby", by matching the local and remote preferential forwarding status, PW-Seg2 would be used for traffic forwarding.

On failure of the AC between CE2 and T-PE2, the forwarding state of AC on T-PE3 is changed to Active. T-PE3 would then advertise the preferential status "Active" to S-PE1, and T-PE2 would advertise the preferential status "Standby". S-PE1 would perform the switchover according to the updated local and remote preferential forwarding status, and select PW-Seg3 to forward traffic. If S-PE selects a new Active PW segment successfully, it SHOULD NOT advertise any change of the PW status to T-PE1. Hence T-PE1 would not be aware of the failure on the remote side.

For scenario of Figure 2, assume the AC from CE2 to T-PE2 is active. T-PE1 would advertise preferential status "Active" on PW1-Seg1 and "Standby" on PW2-Seg1. According to the received preferential status, S-PE1 SHOULD advertise preferential status "Active" on both PW1-Seg2 and PW1-Seg3, and S-PE2 SHOULD advertise preferential status "Standby" on both PW2-Seg2 and PW2-Seg3. T-PE2 advertises preferential status "Active" on both PW1-Seg2 and PW2-Seg2, and T-PE3 advertises preferential status "Standby" on both PW1-Seg3 and PW2-Seg3. By matching the local and remote preferential forwarding status, PW1-Seg2 would be used for traffic forwarding. Since S-PE1 connects to the primary PW segment PW1-Seg2, it would advertise preferential status "Active" to T-PE1. S-PE2 would advertise preferential status "Standby" to T-PE1 since it does not connect to the primary PW segment.

On failure of the AC between CE2 and T-PE2, the forwarding state of AC on T-PE3 is changed to Active. T-PE3 would then advertise the preferential status "Active" on both PW1-Seg3 and PW2-Seg3, and T-PE2 would advertise the preferential status "Standby" on both PW1-Seg2 and PW2-Seg2. S-PE1 would perform the switchover according to the updated local and remote preferential forwarding status, and select PW1-Seg3 to forward traffic. Since S-PE1 selects a new Active PW segment successfully, it SHOULD NOT advertise any change of the PW status to T-PE1, and T-PE would not be aware of the failure on the remote side.

When the S-PE1 fails, T-PE1 would advertise the preferential status "Active" to S-PE2. On receiving the change of preferential status, S-PE2 SHOULD advertise the preferential status "Active" on both PW2-Seg2 and PW2-Seg3. Then by matching the local and remote preferential forwarding status, PW2-Seg2 would be selected as primary PW segment, and traffic would be forwarded on MS-PW2.

4. VCCV Considerations

PW VCCV [RFC5085] CC type 1 "PW ACH" can be used with S-PE redundancy mechanism smoothly. If VCCV CC type 3 "TTL Expiry" is to be used, the hop counts from T-PE1 to the remote T-PE needs be obtained in advance. This can be achieved either by control plane SP-PE TLVs or through data plane tracing of the MS-PW.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

This document has the same security properties as in the PWE3 control protocol [RFC4447] and [I-D.ietf-pwe3-redundancy-bit].

7. Acknowledgements

The authors would like to thank Mach Chen and Lizhong Jin for their comments and suggestions.

8. References

8.1. Normative References

- [I-D.ietf-pwe3-redundancy]
Muley, P., Aissaoui, M., and M. Bocci, "Pseudowire Redundancy", draft-ietf-pwe3-redundancy-08 (work in progress), May 2012.
- [I-D.ietf-pwe3-redundancy-bit]
Muley, P. and M. Aissaoui, "Pseudowire Preferential Forwarding Status Bit", draft-ietf-pwe3-redundancy-bit-07 (work in progress), May 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC5659] Bocci, M. and S. Bryant, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659,

October 2009.

8.2. Informative References

- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC6073] Martini, L., Metz, C., Nadeau, T., Bocci, M., and M. Aissaoui, "Segmented Pseudowire", RFC 6073, January 2011.

Authors' Addresses

Jie Dong
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: rainsword.wang@huawei.com

Operations and Management Area Working Group
Internet-Draft
Intended status: Informational
Expires: January 16, 2014

P. Fan
L. Li
China Mobile
July 15, 2013

Requirements for IP/MPLS network transmission interruption duration
draft-fan-opsawg-transmission-interruption-03

Abstract

The transmission performance of IP/MPLS network affects upper layer services and networks, but there is no consensus in the industry on transmission interruption for IP/MPLS network up to now. This memo studies requirements for the interruption duration criteria in several service scenarios.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 16, 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. Services and Performance Criteria	3
2.1. Softswitch	3
2.2. SS7 transport	5
2.3. LTE Backhaul	6
2.4. Ethernet VPN	6
2.5. IPTV	7
3. Other considerations	7
4. Security Considerations	7
5. IANA Considerations	7
6. Appendix: Impact Analysis on Transmission Quality of IP Carried Softswitch Voice	7
7. Acknowledgements	10
8. Informative References	10
Authors' Addresses	11

1. Introduction

Today's IP/MPLS network is widely used as a bearer network to carry diversified packet switched services. The transmission qualities of these services are closely related to the performance of bearer layers, as network failure, delay, congestion and other abnormalities will inevitably bring about service interruption and user perception degradation. However, there is no consensus in the industry on transmission interruption for IP/MPLS network up to now. This memo studies relationships between service performance and transmission interruption duration in several scenarios, and is intended to reach a list of requirements for these interruption duration criteria.

For a long time the industry has been aspiring for the so-called golden standard for network resilience, that is the 50-millisecond recovery threshold. [HeavyReading] gives us a basic introduction to the origin of this fast protection legacy which can date back to 1980s. The 50ms threshold was established informally in the early 1980s, and then formally through standardization of [G.841] recommendation on SDH network protection architects. The specific requirement shows a maximum threshold for detecting and restoring a fault of 60ms, which adds up fault detection duration of less than 10ms and protection switching time of less than 50ms. The report also mentions original concerns that the threshold results from. The voice channel banks deployed in early 1980s had limited fault tolerance. Failures that lasted longer than 200ms would generate a Carrier Group Alarm (CGA) which caused the channel bank to terminate all connections over that given TDM line. So an outage budget was developed by carriers and the 50ms standard was employed to protect voice services. However newer channel banks at that time had started

to implement a CGA timer of 2s, so the 50ms protection was adopted to protect a small and diminishing fraction of digital network.

Historically this 50ms fast protection speed has been achieved by SDH network. Using various fast convergence technics, IP/MPLS is also able to react within 50ms. As for network applications that are carried by optical or packet core, changes have been made through the past decades, accompanied by the continuing questions about needs for 50ms protection. Here we list three basic considerations about services and their requirement for IP/MPLS: for services like TDM over IP/MPLS, the traditional 50ms guarantee should be kept and met; for current IP services (e.g. voice, internet), experiences or experiments are to be provided for guidance; for services in future, we are supposed to propose requirement early and give consideration to IP/MPLS.

2. Services and Performance Criteria

Services delivered by IP/MPLS network have different transmission quality requirements, thus introduce different performance criteria for the bearing IP/MPLS network. We believe there are two principles that need to be considered during network and service design, configuration and operation. The IP/MPLS bearer should satisfy quality requirements of upper level services and applications, while services and applications should also take into account the intrinsic IP capabilities. In this section we will describe concerns on IP/MPLS and service mutual adaptation from aspects of several kinds of service scenarios.

2.1. Softswitch

From the softswitch point of view, the IP carrying nature imposes certain influence to the service quality. Especially when speech is delivered by IP, the communication quality of voice is impaired, and in turn makes higher requirements for the transmission performance of IP. The following table gives a list of criteria regarding transmission quality of a typical GSM network as well as impacting factors brought by IP bearer.

	Criteria of GSM Transmission Quality	Impacting Factors Brought by IP Bearer
	Call loss of wireless channel	None
Call Loss	Call loss between switches (typical value: <=1%)	Failure of Nc/Mc interface carried by IP

	Call loss between switch and BSC (typical value: $\leq 0.5\%$)	None
Call Cut-off	Call cut-off rate (typical value: $< 1\%$)	Failure of Nc/Mc interface carried by IP
Connection Delay	Service providing delay	None
	Calling party connection delay (typical value: $\leq 4s$)	IP carried signaling delay
	Called party connection delay (typical value: $\leq 4s$)	None

If voice is carried by IP, communication quality criteria of call loss, call cut-off and connection delay are likely to be influenced. This subsection focuses on the three criteria and their impacting factors to give requirements for softswitch and IP bearer networks, with detailed analysis described in the appendix. Note that the current discussion on softswitch is focused on quality of transmission while not on quality of voice. In another word, the scope of discussion is limited to network related QoS aspect, while subjective QoE criteria such as PESQ (Perceptual Evaluation of Speech Quality) and MOS (Mean Opinion Score) are left to later revisions.

Call loss related requirement: The duration of SCTP interface association timer should be shorter than that of the state machine message timer of upper layer protocols, and this duration is further recommended to be no longer than 6 seconds in order to maintain detection sensitivity; the interruption duration of IP bearer network should be as short as possible to avoid call loss, and this duration is further recommended to be no longer than 5 seconds.

Call cut-off related requirement: The SCTP association should be guaranteed during IP layer interruption to avoid interface breakoff alert. The requirements are the same as those related to call loss.

Connection delay related requirement: The IP convergence time should be no longer than 3 seconds to ensure that connection delay is shorter than 4 seconds.

The overall requirement for IP/MPLS interruption duration is no longer than 3 seconds.

2.2. SS7 transport

The Signaling System No. 7 (SS7/C7) network is one of the examples of the principle that services should take into account the ability of IP. The bearer of SS7 protocol stack has been experiencing evolution from TDM to IP. Traditionally the user parts of SS7 (including MAP, CAP, BSSAP+, ISUP, etc.) are carried by MTP layers, but the bearer has gradually been evolved into a packetized form with SIGTRAN (including M2PA, M2UA, M3UA, etc.) using SCTP associations over IP. The change requires transport layer to take mechanisms to meet demand of SCN signaling, and more importantly it requires protocols to make adaption to the "best effort" fact of IP.

The SIGTRAN uses an architecture that can be described as standard IP plus unified transport plus diversified adaption units. It introduces SCTP to realize reliable signaling transport over IP. The SCTP itself provides reliable transmission mechanisms, such as path selection and monitoring, validation and acknowledgment mechanisms, and retransmission timing management.

The unreliable nature of IP makes it necessary for the upper-level protocols to be more tolerable to the possible instability of bearer. Once a service request from a UE is accepted, the system allocates resources and establishes paths for the user. A breakoff caused by IP will result in signaling disconnection or rerouting. Signaling transmission path may also be switched back after IP layer restores. Frequent switchovers and disconnections lead to unnecessary system cost and service interruption, so parameters should be configured a little bit "insensitive" to try to sustain connections on control plane.

One of the examples of parameter configuration is the timer value. The following gives two cases about SCTP on transport layer and M2PA on adaption layer. The values should not be set very small to prevent unnecessary disconnection caused by IP instability. However, because upper services of SS7 may also have timeout rules, values should not be set very large too to avoid violating the rules.

1) SCTP

SCTP uses RTO to manage timeout duration for retransmission in case of feedback missing. The RTO is given an initial, a max and a min value, and is calculated instantaneously with a set of management rules. Many other parameters are used for fault detection in SCTP. Association.Max.Retrans is used to indicate the upper limit of number of possible retransmission without considering endpoint down. Path.Max.Retrans is a similar value to detect path failure. The parameters together characterize the ability of SCTP to tolerate

bearer downwards and provide reliable SS7 transport upwards. The typical values of the parameters are RTO.Initial = 0.5 sec, RTO.MIN = 0.5 sec, RTO.MAX = 1.5 sec, Path.Max.Retrans = 5, Assoc.Max.Retrans = 10.

2) M2PA

Although protocols like H.248 and BICC can be carried directly upon SCTP, the user part protocols of SS7 usually have to be carried by SCTP/IP with the help of different adaption layers. In this case, the attributes of adaption layers, e.g. M2PA used between STPs, are more important to SS7. M2PA uses a T7 timer to indicate the maximum delay of acknowledgement and start T7 at the time of data transmission. If no message is acknowledged after the maximum waiting time, T7 expires and M2PA sends a message of out of service to the peer end. Because propagation delays in IP networks are more variable than in traditional SS7 networks, the value of T7 should be set considering IP propagation delays, as well as acknowledgement time, SCTP slow-start algorithms, upper service timers and other factors. Typical value of T7 is 7~10 sec.

Parameter configuration induced tolerance to bearer may have some influence on service, but it avoids service cut-off or severe user perception degradation. For services like SMS or route lookup, possible latency may be introduced, but operations can still be completed after short delay. Because SMS has no strict requirement for instantaneity, impact on service is limited. If route lookup takes more time due to IP interruption and convergence, user may experience longer setup delay when dialing. For service of location update, even if operation fails because bearer is interrupted for too long, UE has the mechanism to initiate request again.

2.3. LTE Backhaul

To be further analyzed.

2.4. Ethernet VPN

Ethernet VPNs (e.g. VPLS) are used to provide transparent Ethernet type layer 2 connections for customers. Ethernet frames are treated as service payload and encapsulated and transported in providers MPLS network. The interruption criteria of IP/MPLS bearer should guarantee continuity of Ethernet service, and IP/MPLS failover is not supposed to generate outage of Ethernet service.

[Y.1731] and [IEEE802.1ag] describe in detail OAM functions and mechanisms for Ethernet, with specific recommendation on connectivity fault management. Ethernet uses continuity check function to detect

loss of continuity between any pair of MEPs in a MEG, and this function is realized by sending CCMs (connectivity check messages) between peer MEPs. When a MEP does not receive CCM from a peer MEP within a certain interval, it detects loss of continuity to that peer MEP. The threshold interval is specified as 3.5 times the CCM transmission period, which corresponds to a loss of three consecutive CCMs from the peer MEP, and the CCM transmission period is recommended to be the default value of 1 second. So the interruption duration of IP/MPLS for Ethernet VPN services should be less than 3 seconds.

2.5. IPTV

To be further analyzed.

3. Other considerations

So far this document has focused on use cases and their requirement for IP/MPLS, and other practical issues are not included in this version. For example, an IP/MPLS packet core is expected to carry a variety of services, so the requirement for IP/MPLS may have to include additional concerns on this multi-service co-existence scenario. A simple and straight-forward way may be to satisfy the most critical need for protection time required by the services. Another issue is related to service awareness. Whether service type is or can be known by IP/MPLS would influence the ability of IP/MPLS to provide reliability guarantee accordingly. It seems to be easier to perform service identification on edge devices than network core. We believe these kinds of issues need to be taken into account, and currently we will just leave them to be updated in future revisions.

4. Security Considerations

TBD

5. IANA Considerations

This memo includes no request to IANA.

6. Appendix: Impact Analysis on Transmission Quality of IP Carried Softswitch Voice

This section describes impact on transmission quality of softswitch voice when carried by IP and requirements for IP bearer convergence time.

1) Call Loss

Call loss is used to describe the circumstance where a phone call fails to establish after initiated by a subscriber due to network faults. In the practical network, the call loss rate is mainly associated by the factors as follows:

1. Interfaces, including Nc, Mc and interface between MSS and SG.
2. State machine message timer. If a timeout takes place, the state machine releases signaling messages, producing a call loss. Typical value of BICC timer is 10~15 seconds and value of DTAP timer about 15 seconds.
3. Interface association timer. Associations breaks off at the expiration of timer.
4. Bearer network convergence time.

If the configured timer duration of a state machine is shorter than the timer duration of interface association, then although interface association may not be broken off, call loss is still possible to occur due to message timer expiration. If the association timer duration is shorter than IP routing convergence time, the association is considered broken off by SCTP, hence message loss at interface between MSS and SG as well as interface Nc results in massive call loss, and new calling request cannot be satisfied because of interface Mc breakoff. In this case, the call loss rate can be calculated as

$$\text{Call Loss Rate} = (\text{IP Convergence Time} + \text{Association Restoration Time}) * \text{CAPS} / \text{BHCA}.$$

However, if the association timer duration is longer than IP routing convergence time, then the association is considered normal by SCTP, and data will be retransmitted. Although this may cause buffer overflow leading to call loss, the call loss rate is possible to achieve approximately zero if buffer is big enough.

From the analysis above and practical operation experience, the requirements for softswitch and IP bearer are as follows: the duration of SCTP interface association timer should be shorter than that of the state machine message timer, and this duration is further recommended to be no longer than 6 seconds in order to maintain detection sensitivity; the interruption duration of IP bearer network should be as short as possible to avoid call loss during the IP layer interruption period, and this duration is further recommended to be no longer than 5 seconds.

2) Call Cut-off

Call cut-off is referred to the abnormal release during a phone call due to reasons other than intentional release by any of the parties involved in the call. The call cut-off rate is related with:

1. Interfaces, including Nc and interface between MSS and SG.
2. Interface association timer.
3. Bearer network convergence time.

If the association timer duration is shorter than IP routing convergence time, established phone calls will be released once interruption of interface Nc or interface connecting MSS and SG is detected. In the case of association breakoff, call cut-off rate can be calculated as

$$\text{Call Cut-off Rate} = (\text{CAPS} * \text{Call Duration}) * \text{Busy Hour Association Breakoffs} / \text{BHCA}.$$

While if the association is not interrupted, the call cut-off rate can be approximately zero.

In conclusion, the SCTP association should be guaranteed during IP layer interruption to avoid interface breakoff alert. The requirements for softswitch and IP bearer are the same as those related to call loss.

3) Connection Delay

The connection delay from a call initiation by a calling party to PLMN should be no longer than 4 seconds. This delay is affected by factors below:

1. RRC connection setup delay (irrelevant to whether service is carried by IP or not).
2. Core network signaling interaction delay. The message number at interface Nc/Nb is 6, and is 8 (calling side) or 16 (called side, in case of IP-IP) at interface Mc. Each message is with a delay of no longer than 50 milliseconds. Calling message delay at interface Nc is no longer than 300 milliseconds. If long distance call is made through CMN, the message delay is to be increased by transmission delay of 5 msec/km and CMN process delay. So the message delay is likely to be 400 milliseconds.

3. IP bearer network QoS and load.

The connection delay is influenced by the delay criterion defined in the IP bearer network QoS, and is raised by delay, jitter, packet loss caused by network overload. In addition, if the configured timer duration of interface association is too long, the SCTP sensitivity to the retransmitted messages after packet loss will be decreased, which increases connection delay.

Connection delay is generally expressed as

$$\text{Connection Delay} = \text{IP convergence time} + \text{RRC connection setup delay} + \text{Signaling Interaction Delay},$$

and is no longer than 4 seconds. So the IP network in normal working state should be constrained within a certain range of load to ensure that delay is shorter than 50 milliseconds, while in interruption state the IP convergence time should be no longer than 3 seconds to ensure that connection delay is shorter than 4 seconds.

From the analysis of IP/MPLS performance according to the three criteria above, we suggest the transmission interruption duration of IP/MPLS network for softswitch service should be no longer than 3 seconds.

7. Acknowledgements

The authors would like to thank Chris Donley and Melinda Shore for their kind help in content enrichment, and Christopher Liljenstolpe, Andrew Malis and Adrian Farrel for their helpful comments on the document.

8. Informative References

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

[HeavyReading]

Bennett, G., "Resilience Reliability and OAM in Converged Network", Heavy Reading, Vol. 2, No. 6, February 2004.

[IEEE802.1ag]

IEEE Std 802.1ag-2007, ., "IEEE Standard for Local and metropolitan area networks, Virtual Bridged Local Area Networks, Amendment 5: Connectivity Fault Management", December 2007.

[Y.1731] ITU-T Recommendation Y.1731, ., "OAM Functions and Mechanisms for Ethernet based Networks", July 2011.

Authors' Addresses

Peng Fan
China Mobile
32 Xuanwumen West Street, Xicheng District
Beijing 100053
P.R. China

Email: fanpeng@chinamobile.com

Lianyuan Li
China Mobile
32 Xuanwumen West Street, Xicheng District
Beijing 100053
P.R. China

Email: lilianyuan@chinamobile.com

Pseudowire Emulation Edge to Edge
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2013

H. Hao
Y. Ma
ZTE Corporation
W. Cheng
China Mobile
D. Cohn
Orckit-Corrigent
July 11, 2012

ICCP extension for the MSP application
draft-hao-pwe3-iccp-extension-for-msp-03

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 2013.

Copyright Notice

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

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

Table of Contents

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

1. Introduction

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

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

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

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

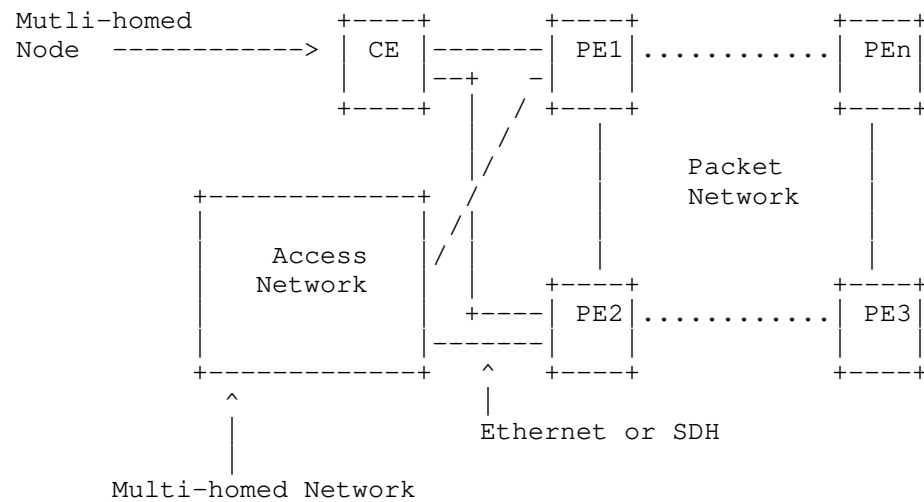


Figure 1: Attachment circuit multi-homed to Packet Network

1.1. Conventions used in this document

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

2. Terminology

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

3. ICCP extension requirements

3.1. Multi-chassis MSP Protection Model

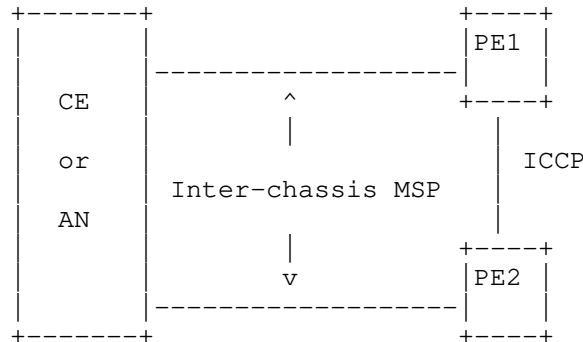


Figure 2: Generic Multi-chassis MSP Protection Model

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

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

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

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

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

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

The ICCP application described in this document is responsible for

the state synchronization between different chassis forming a RG.

3.2. ICCP aspects

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

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

4. ICCP TLV extensions for MSP

The following sections specify the format of MSP application TLVs.

4.1. MSP connect TLV

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

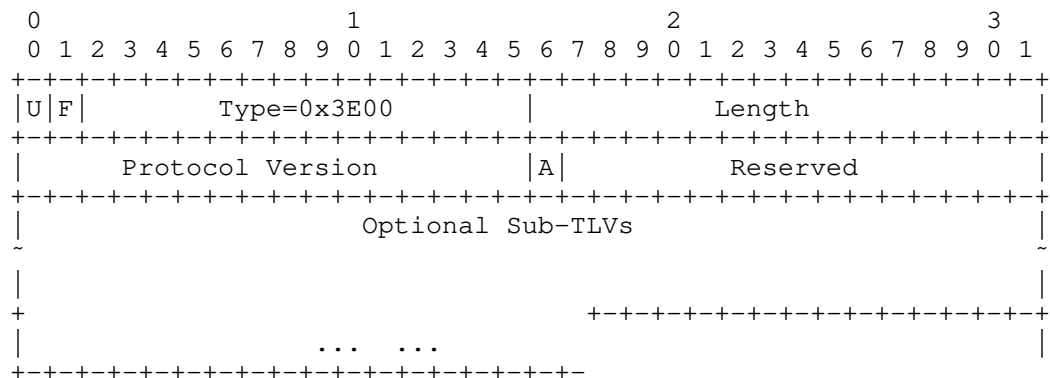


Figure 3: MSP connect TLV

- U and F Bits

Both are set to 0.

- Type

Set temporarily to 0x3E00 for "MSP connect TLV"

- Length

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

- Protocol Version

The version of this particular protocol for the purposes of ICCP. This is set to 0x0001.

- A Bit

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

- Reserved

Reserved for future use.

- Optional Sub-TLVs

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

4.2. MSP disconnect TLV

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

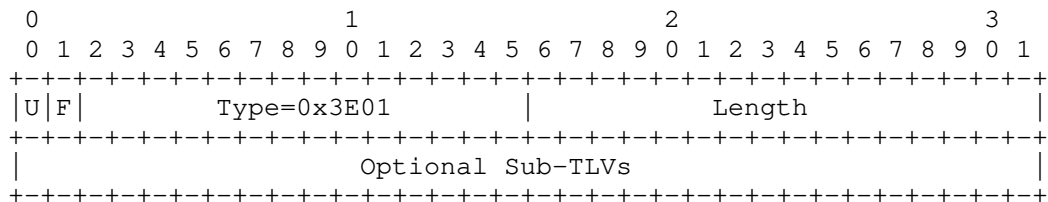


Figure 4: MSP disconnect TLV

- U and F Bits

Both are set to 0.

- Type

Set temporarily to 0x3E01 for "MSP disconnect TLV"

- Length

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

- Optional Sub-TLVs

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

4.2.1. MSP disconnect cause TLV

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

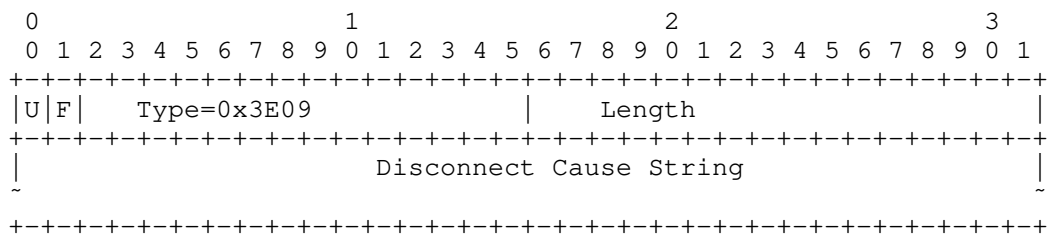


Figure 5: MSP disconnect TLV

- U and F Bits

Both are set to 0.

- Type

```
Set temporarily to 0x3E09 for "MSP disconnect cause TLV"
```

- Length

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

- Disconnect Cause String

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

4.3. MSP group config TLV

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

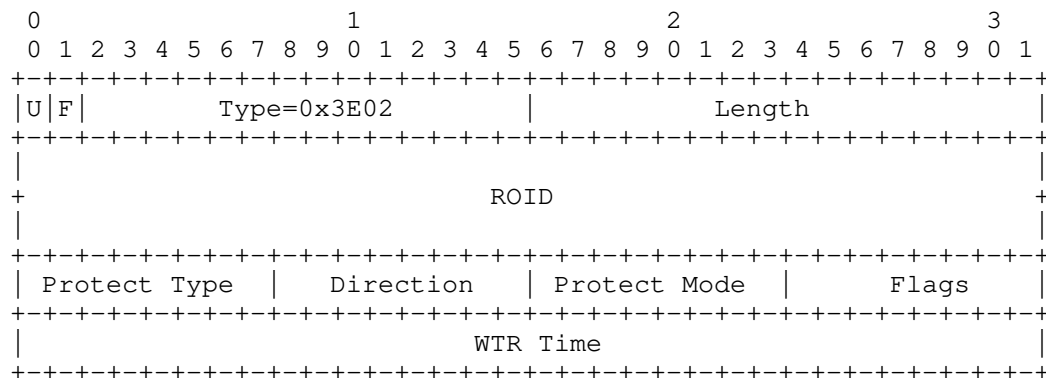


Figure 6: MSP group config TLV

- U and F Bits
Both are set to 0.

- Type
Set temporarily to 0x3E02 for "MSP group config TLV"

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

- ROID
Defined in the [I-D:ietf-pwe3-iccp]. Eight octets, uniquely identifies the Redundant Object.

- Protect Type
One octet encoding the protect type of the MSP protect group as follows:
0x00 1+1
0x01 1:1
0x02-0xFF reserved

- Direction
One octet encoding the architecture of the network as follows:
0x00 unidirectional
0x01 bidirectional

- Reversion Mode
One octet encoding the mode of operation as follows:
0x00 non-revertive operation
0x01 revertive operation

- Flags

One octet. Valid values are:

- i Synchronized (0x01)

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

- ii Purge Configuration (0x02)

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

- WTR Time

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

4.4. MSP port config TLV

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

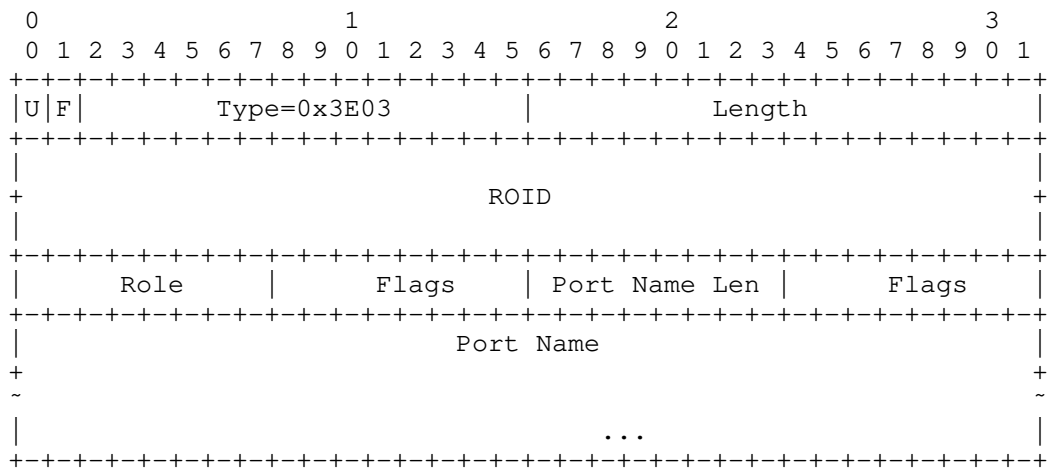


Figure 7: MSP port config TLV

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

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

4.5. MSP section state TLV

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

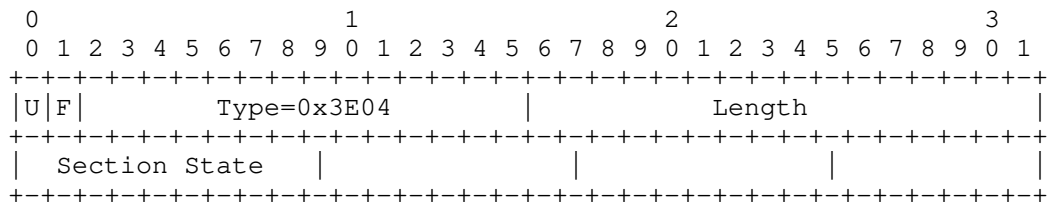


Figure 8: MSP section state TLV

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

fields.

- Section State

One octet encoding the section state as follows:

0x00 the signal is ok
 0x01 Signal fail high priority
 0x02 Signal fail low priority
 0x03 Signal degrade high priority
 0x04 Signal degrade low priority

4.6. MSP switch command TLV

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

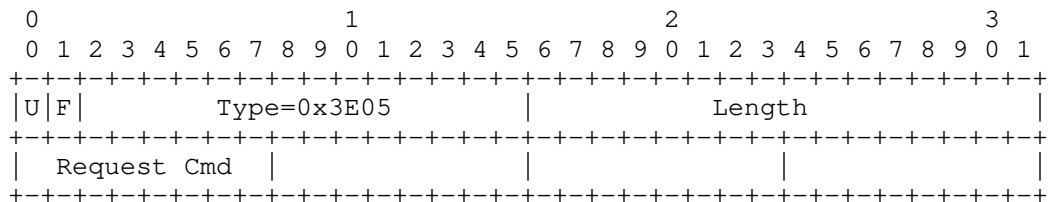


Figure 9: MSP switch command TLV

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Request Cmd

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

(1111) Clear
 (1101) Lockout of protection(LP)
 (1011) Forced Switch working-to-protection
 (1001) Forced Switch protection-to-working
 (0111) Manual switch working-to-protection
 (0101) Manual switch protection-to-working
 (0100) Exercise

4.7. MSP group state TLV

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

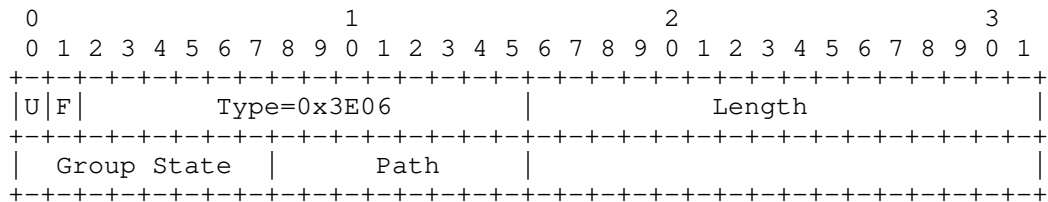


Figure 10: MSP group state TLV

- U and F Bits
Both are set to 0.

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

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

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

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

- Path

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

- 0x00 the active path is the working section
- 0x01 the active path is the protection section

4.8. MSP Synchronization Request TLV

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

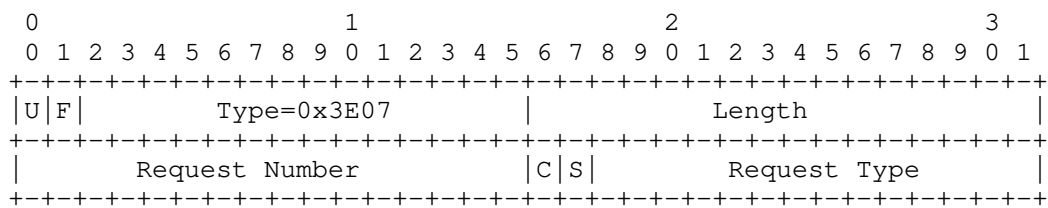


Figure 11: MSP Synchronization Request TLV

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Request Number

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

- C Bit

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

- S Bit

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

- Request Type

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

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

4.9. MSP Synchronization Data TLV

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

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

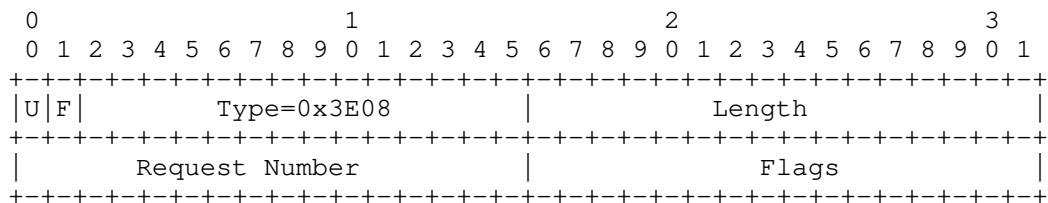


Figure 12: MSP group notify TLV

- U and F Bits
 Both are set to 0.

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

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

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

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

5. PE Node Failure

Section 9.2.3 of [I-D.ietf-pwe3-iccp] specifies the behavior in the event of PE node failure. Additionally, a signal fail request for the working section (SF-W) over the K1 byte will be received by the PE node which terminates the protection section and then it will follow normal MSP procedure for this condition.

6. Security Considerations

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

7. IANA Consideration

TBD.

8. References

8.1. Normative References

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

8.2. Informative References

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

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

Authors' Addresses

Hongjie Hao
ZTE Corporation

Email: hao.hongjie@zte.com.cn

Yuxia Ma
ZTE Corporation

Email: ma.yuxia@zte.com.cn

Weiqiang Cheng
China Mobile

Email: chengweiqiang@cmcc.com.cn

Daniel Cohn
Orckit-Corrigent

Email: daniel.cohn.ietf@gmail.com

Wanming Cao
ZTE Corporation

Email: cao.wanming@zte.com.cn

Jinghai Yu
ZTE Corporation

Email: yu.jinghai@zte.com.cn

TICTOC Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 9, 2012

S. Davari
A. Oren
Broadcom Corp.
M. Bhatia
P. Roberts
Alcatel-Lucent
L. Montini
Cisco Systems
October 7, 2011

Transporting PTP messages (1588) over MPLS Networks
draft-ietf-tictoc-1588overmpls-02

Abstract

This document defines the method for transporting PTP messages (PDUs) over an MPLS network. The method allows for the easy identification of these PDUs at the port level to allow for port level processing of these PDUs in both LERs and LSRs.

The basic idea is to transport PTP messages inside dedicated MPLS LSPs. These LSPs only carry PTP messages and possibly Control and Management packets, but they do not carry customer traffic.

Two methods for transporting 1588 over MPLS are defined. The first method is to transport PTP messages directly over the dedicated MPLS LSP via UDP/IP encapsulation, which is suitable for IP/MPLS networks. The second method is to transport PTP messages inside a PW via Ethernet encapsulation, which is more suitable for MPLS-TP networks.

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 9, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction	6
2. Terminology	7
3. Problem Statement	8
4. 1588 over MPLS Architecture	9
5. Dedicated LSPs for PTP messages	10
6. 1588 over MPLS Encapsulation	11
6.1. 1588 over LSP Encapsulation	11
6.2. 1588 over PW Encapsulation	11
7. 1588 Message Transport	14
8. Protection and Redundancy	16
9. ECMP	17
10. OAM, Control and Management	18
11. QoS Considerations	19
12. FCS Recalculation	20
13. UDP Checksum Correction	21
14. Routing extensions for 1588aware LSRs	22
14.1. 1588aware Link Capability for OSPF	22
14.2. 1588aware Link Capability for IS-IS	23
15. RSVP-TE Extensions for support of 1588	25
16. Behavior of LER/LSR	26
16.1. Behavior of 1588-aware LER	26
16.2. Behavior of 1588-aware LSR	26
16.3. Behavior of non-1588-aware LSR	26
17. Other considerations	28
18. Security Considerations	29
19. Acknowledgements	30
20. IANA Considerations	31

20.1. IANA Considerations for OSPF	31
20.2. IANA Considerations for IS-IS	31
20.3. IANA Considerations for RSVP	31
21. References	32
21.1. Normative References	32
21.2. Informative References	32
Authors' Addresses	34

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 [RFC2119].

When used in lower case, these words convey their typical use in common language, and are not to be interpreted as described in RFC2119 [RFC2119].

1. Introduction

The objective of Precision Time Protocol (PTP) is to synchronize independent clocks running on separate nodes of a distributed system. [IEEE] defines PTP messages for clock and time synchronization. The PTP messages include PTP PDUs over UDP/IP (Annex D and E of [IEEE]) and PTP PDUs over Ethernet (Annex F of [IEEE]). This document defines mapping and transport of the PTP messages defined in [IEEE] over MPLS networks.

PTP defines several clock types: ordinary clocks, boundary clocks, end-to-end transparent clocks, and peer-to-peer transparent clocks. One key attribute of all of these clocks is the recommendation for PTP messages processing to occur as close as possible to the actual transmission and reception at the physical port interface. This targets optimal time and/or frequency recovery by avoiding variable delay introduced by queues internal to the clocks. To facilitate the fast and efficient recognition of PTP messages at the port level when the PTP messages are carried over MPLS LSPs, this document defines the specific encapsulations that should be used. In addition, it can be expected that there will exist LSR/LEs where only a subset of the physical ports will have the port based PTP message processing capabilities. In order to ensure that the PTP carrying LSPs always enter and exit ports with this capability, routing extensions are defined to advertise this capability on a port basis and to allow for the establishment of LSPs that only transit such ports. While this path establishment restriction may be applied only at the LER ingress/egress ports, it becomes more important when using Transparent Clock capable LSRs in the path.

The port based PTP message processing involves PTP event message recognition. Once the PTP event messages are recognized they can be modified based on the reception or transmission timestamp. An alternative technique to actual packet modification could include the enforcement of a fixed delay time across the LSR to remove variability in the transit delay. This latter would be applicable in a LSR which does not contain a PTP transparent Clock function.

This document provides two methods for transporting PTP messages over MPLS. One is principally focused on an IP/MPLS environment and the second is focused on the MPLS-TP environment.

While the techniques included herein allow for the establishment of paths optimized to include PTP Timestamping capable links, the performance of the Slave clocks is outside the scope of this document.

2. Terminology

1588: The timing and synchronization as defined by IEEE 1588

PTP: The timing and synchronization protocol used by 1588

Master Clock: The source of 1588 timing to a set of slave clocks.

Master Port: A port on a ordinary or boundary clock that is in Master state. This is the source of timing toward slave ports.

Slave Clock: A receiver of 1588 timing from a master clock

Slave Port: A port on a boundary clock or ordinary clock that is receiving timing from a master clock.

Ordinary Clock: A device with a single PTP port.

Transparent Clock. A device that measures the time taken for a PTP event message to transit the device and then updates the correctionField of the message with this transit time.

Boundary Clock: A device with more than one PTP port. Generally boundary clocks will have one port in slave state to receive timing and then other ports in master state to re-distribute the timing.

PTP LSP: An LSP dedicated to carry PTP messages

PTP PW: A PW within a PTP LSP that is dedicated to carry PTP messages.

CW: Pseudowire Control Word

LAG: Link Aggregation

ECMP: Equal Cost Multipath

CF: Correction Field, a field inside certain PTP messages (message type 0-3) that holds the accumulative transit time inside intermediate switches

3. Problem Statement

When PTP messages are transported over MPLS networks, there is a need for PTP message processing at the physical port level. This requirement exists to minimum uncertainty in the transit delays. If PTP message processing occurs interior to the MPLS routers, then the variable delay introduced by queuing between the physical port and the PTP processing will add noise to the timing distribution. Port based processing applies at both the originating and terminating LERs and also at the intermediate LSRs if they support transparent clock functionality.

PTP messages over Ethernet or IP can always be tunneled over MPLS. However there is a requirement to limit the possible encapsulation options to simplify the PTP message processing required at the port level. This applies to all 1588 clock types implemented in MPLS routers. But this is particularly important in LSRs that provide transparent clock functionality.

When 1588-awareness is needed, PTP messages should not be transported over LSPs or PWs that are carrying customer traffic because LSRs perform Label switching based on the top label in the stack. To detect PTP messages inside such LSPs require special hardware to do deep packet inspection at line rate. Even if such hardware exists, the payload can't be deterministically identified by LSRs because the payload type is a context of the PW label and the PW label and its context are only known to the Edge routers (PEs); LSRs don't know what is a PW's payload (Ethernet, ATM, FR, CES, etc). Even if one restricts an LSP to only carry Ethernet PWs, the LSRs don't have the knowledge of whether PW Control Word (CW) is present or not and therefore can't deterministically identify the payload.

Therefore a generic method is defined in this document that does not require deep packet inspection at line rate, and can deterministically identify PTP messages. The defined method is applicable to both MPLS and MPLS-TP networks.

4. 1588 over MPLS Architecture

1588 communication flows map onto MPLS nodes as follows: 1588 messages are exchange between PTP ports on Ordinary and boundary clocks. Transparent clocks do not terminate the PTP messages but they do modify the contents of the PTP messages as they transit across the Transparent clock. SO Ordinary and boundary clocks would exist within LERs as they are the termination points for the PTP messages carried in MPLS. Transparent clocks would exist within LSRs as they do not terminate the PTP message exchange.

Perhaps a picture would be good here.

5. Dedicated LSPs for PTP messages

Many methods were considered for identifying the 1588 messages when they are encapsulated in MPLS such as by using GAL/ACH or a new reserved label. These methods were not attractive since they either required deep packet inspection and snooping at line rate or they required use of a scarce new reserved label. Also one of the goals was to reuse existing OAM and protection mechanisms.

The method defined in this document can be used by LER/LSRs to identify PTP messages in MPLS tunnels by using dedicated LSPs to carry PTP messages.

Compliant implementations MUST use dedicated LSPs to carry PTP messages over MPLS. These LSPs are herein referred to as "PTP LSPs" and the labels associated with these LSPs as "PTP labels". These LSPs could be P2P or P2MP LSPs. The PTP LSP between Master Clocks and Slave Clocks MAY be P2MP or P2P LSP while the PTP LSP between each Slave Clock and Master Clock SHOULD be P2P LSP. The PTP LSP between a Master Clock and a Slave Clock and the PTP LSP between the same Slave Clock and Master Clock MUST be co-routed. Alternatively, a single bidirectional co-routed LSP can be used. The PTP LSP MAY be MPLS LSP or MPLS-TP LSP. This co-routing is required to limit differences in the delays in the Master clock to Slave clock direction compared to the Slave clock to Master clock direction.

The PTP LSPs could be configured or signaled via RSVP-TE/GMPLS. New RSVP-TE/GMPLS TLVs and objects are defined in this document to indicate that these LSPs are PTP LSPs.

The PTP LSPs MAY carry essential MPLS/MPLS-TP control plane traffic such as BFD and LSP Ping but the LSP user plane traffic MUST be PTP only.

6. 1588 over MPLS Encapsulation

This document defines two methods for carrying PTP messages over MPLS. The first method is carrying IP encapsulated PTP messages over PTP LSPs and the second method is to carry PTP messages over dedicated Ethernet PWs (called PTP PWs) inside PTP LSPs.

6.1. 1588 over LSP Encapsulation

The simplest method of transporting PTP messages over MPLS is to encapsulate PTP PDUs in UDP/IP and then encapsulate them in PTP LSP. The 1588 over LSP format is shown in Figure 1.

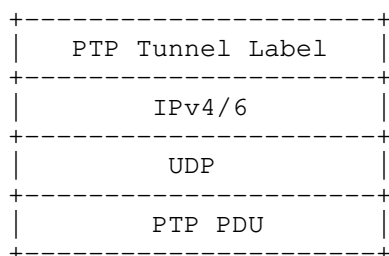


Figure 1 - 1588 over LSP Encapsulation

This encapsulation is very simple and is useful when the networks between 1588 Master Clock and Slave Clock are IP/MPLS networks.

In order for an LSR to process PTP messages, the PTP Label must be the top label of the label stack.

The UDP/IP encapsulation of PTP MUST follow Annex D and E of [IEEE].

6.2. 1588 over PW Encapsulation

Another method of transporting 1588 over MPLS networks is by encapsulating PTP PDUs in Ethernet and then transporting them over Ethernet PW (PTP PW) as defined in [RFC4448], which in turn is transported over PTP LSPs. Alternatively PTP PDUs MAY be encapsulated in UDP/IP/Ethernet and then transported over Ethernet PW.

Both Raw and Tagged modes for Ethernet PW are permitted. The 1588 over PW format is shown in Figure 2.

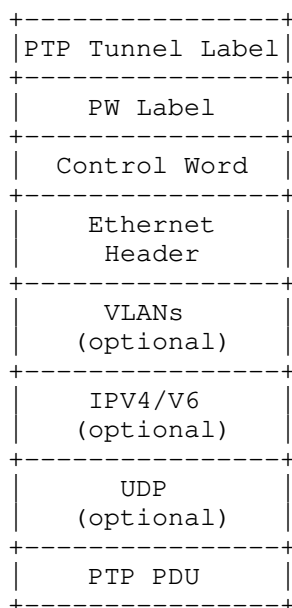


Figure 2 - 1588 over PW Encapsulation

The Control Word (CW) as specified in [RFC4448] SHOULD be used to ensure a more robust detection of PTP messages inside the MPLS packet. If CW is used, the use of Sequence number is optional.

The use of VLAN and UDP/IP are optional. Note that 1 or 2 VLANs MAY exist in the PW payload.

In order for an LSR to process PTP messages, the top label of the label stack (the Tunnel Label) MUST be from PTP label range. However in some applications the PW label may be the top label in the stack, such as cases where there is only one-hop between PEs or in case of PHP. In such cases, the PW label SHOULD be chosen from the PTP Label range.

In order to ensure congruency between the two directions of PTP message flow, ECMP should not be used for the PTP LSPs. Therefore, no Entropy label [I-D.ietf-pwe3-fat-pw] is necessary and it SHOULD NOT be present in the stack.

The Ethernet encapsulation of PTP MUST follow Annex F of [IEEE] and the UDP/IP encapsulation of PTP MUST follow Annex D and E of [IEEE].

For 1588 over MPLS encapsulations that are PW based, there are some cases in which the PTP LSP label may not be present:

- o When PHP is applied to the PTP LSP, and the packet is received without PTP LSP label at PW termination point .
- o When the PW is established between two routers directly connected to each other and no PTP LSP is needed.

In such cases it is required for a router to identify these packets as PTP packets. This would require the PW label to also be a label that is distributed specifically for carrying PTP traffic (aka PTP PW label). Therefore there is a need to add extension to LDP/BGP PW label distribution protocol to indicate that a PW label is a PTP PW labels.

7. 1588 Message Transport

1588 protocol comprises of the following message types:

- o Announce
- o SYNC
- o FOLLOW UP
- o DELAY_REQ (Delay Request)
- o DELAY_RESP (Delay Response)
- o PDELAY_REQ (Peer Delay Request)
- o PDELAY_RESP (Peer Delay Response)
- o PDELAY_RESP_FOLLOW_UP (Peer Delay Response Follow up)
- o Management
- o Signaling

A subset of PTP message types that require timestamp processing are called Event messages:

- o SYNC
- o DELAY_REQ (Delay Request)
- o PDELAY_REQ (Peer Delay Request)
- o PDELAY_RESP (Peer Delay Response)

SYNC and DELAY_REQ are exchanged between Master Clock and Slave Clock and MUST be transported over PTP LSPs. PDELAY_REQ and PDELAY_RESP are exchanged between adjacent PTP clocks (i.e. Master, Slave, Boundary, or Transparent) and MAY be transported over single hop PTP LSPs. If Two Step PTP clocks are present, then the FOLLOW_UP, DELAY_RESP, and PDELAY_RESP_FOLLOW_UP messages must also be transported over the PTP LSPs.

For a given instance of 1588 protocol, SYNC and DELAY_REQ MUST be transported over two PTP LSPs that are in opposite directions. These PTP LSPs, which are in opposite directions MUST be congruent and co-routed. Alternatively, a single bidirectional co-routed LSP can be used.

Except as indicated above for the two-step PTP clocks, Non-Event PTP message types don't need to be processed by intermediate routers. These message types MAY be carried in PTP Tunnel LSPs.

8. Protection and Redundancy

In order to ensure continuous uninterrupted operation of 1588 Slaves, usually as a general practice, Redundant Masters are tracked by each Slave. It is the responsibility of the network operator to ensure that physically disjoint PTP tunnels that don't share any link are used between the redundant Masters and a Slave.

When redundant Masters are tracked by a Slave, any prolonged PTP LSP or PTP PW outage will trigger the Slave Clock to switch to the Redundant Master Clock. However LSP/PW protection such as Linear Protection Switching (1:1,1+1), Ring protection switching or MPLS Fast Reroute (FRR) SHOULD still be used to provide resiliency to individual network segment failures..

Note that any protection or reroute mechanism that adds additional label to the label stack, such as Facility Backup Fast Reroute, MUST ensure that the pushed label is a PTP Label to ensure recognition of the MPLS frame as containing PTP messages as it transits the backup path..

9. ECMP

To ensure the optimal operation of 1588 Slave clocks and avoid errors introduced by forward and reverse path delay asymmetry, the physical path for PTP messages from Master Clock to Slave Clock and vice versa must be the same for all PTP messages listed in section 7 and must not change even in the presence of ECMP in the MPLS network.

To ensure the forward and reverse paths are the same PTP LSPs and PWs MUST NOT be subject to ECMP.

10. OAM, Control and Management

In order to manage PTP LSPs and PTP PWs, they MAY carry OAM, Control and Management messages. These control and management messages can be differentiated from PTP messages via already defined IETF methods.

In particular BFD [RFC5880], [RFC5884] and LSP-Ping [RFC4389] MAY run over PTP LSPs via UDP/IP encapsulation or via GAL/G-ACH. These Management protocols are easily identified by the UDP Destination Port number or by GAL/ACH respectively.

Also BFD, LSP-Ping and other Management messages MAY run over PTP PW via one of the defined VCCVs (Type 1, 2 or 3) [RFC5085]. In this case G-ACH, Router Alert Label (RAL), or PW label (TTL=1) are used to identify such management messages.

11. QoS Considerations

In network deployments where not every LSR/LER is PTP-aware, then it is important to reduce the impact of the non-PTP-aware LSR/LERs on the timing recovery in the slave clock. The PTP messages are time critical and must be treated with the highest priority. Therefore 1588 over MPLS messages must be treated with the highest priority in the routers. This can be achieved by proper setup of PTP tunnels. It is recommended that the PTP LSPs are setup and marked properly to indicate EF-PHB for the CoS and Green for drop eligibility.

In network deployments where every LSR/LER supports PTP LSPs, then it MAY NOT be required to apply the same level of prioritization as specified above.

12. FCS Recalculation

Ethernet FCS of the outer encapsulation MUST be recalculated at every LSR that performs the Transparent Clock processing and FCS retention for the payload Ethernet described in [RFC4720] MUST NOT be used.

13. UDP Checksum Correction

For UDP/IP encapsulation mode of 1588 over MPLS, the UDP checksum is optional when used for IPv4 encapsulation and mandatory in case of IPv6. When IPv4/v6 UDP checksum is used each 1588-aware LSR must either incrementally update the UDP checksum after the CF update or should verify the UDP checksum on reception from upstream and recalculate the checksum completely on transmission after CF update to downstream node.

14. Routing extensions for 1588aware LSRs

MPLS-TE routing relies on extensions to OSPF [RFC2328] [RFC5340] and IS-IS [ISO] [RFC1195] in order to advertise Traffic Engineering (TE) link information used for constraint-based routing.

Indeed, it is useful to advertise data plane TE router link capabilities, such as the capability for a router to be 1588-aware. This capability **MUST** then be taken into account during path computation to prefer or even require links that advertise themselves as 1588-aware. In this way the path can ensure the entry and exit points into the LERs and, if desired, the links into the LSRs are able to perform port based timestamping thus minimizing their impact on the performance of the slave clock.

For this purpose, the following sections specify extensions to OSPF and IS-IS in order to advertise 1588 aware capabilities of a link.

14.1. 1588aware Link Capability for OSPF

OSPF uses the Link TLV (Type 2) that is itself carried within either the Traffic Engineering LSA specified in [RFC3630] or the OSPFv3 Intra-Area-TE LSA (function code 10) defined in [RFC5329] to advertise the TE related information for the locally attached router links. For an LSA Type 10, one LSA can contain one Link TLV information for a single link. This extension defines a new 1588-aware capability sub-TLV that can be carried as part of the Link TLV.

The 1588-aware capability sub-TLV is **OPTIONAL** and **MUST NOT** appear more than once within the Link TLV. If a second instance of the 1588-aware capability sub-TLV is present, the receiving system **MUST** only process the first instance of the sub-TLV. It is defined as follows:

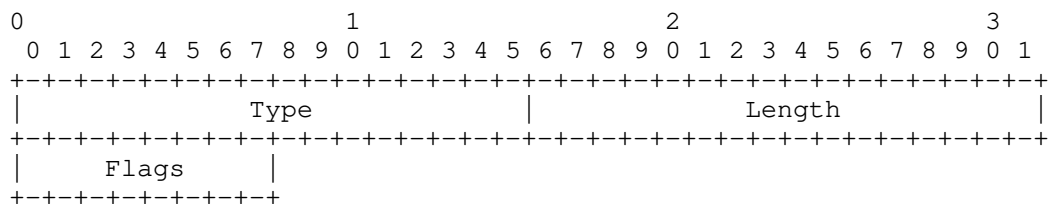


Figure 3: 1588-aware Capability TLV

Where:

Type, 16 bits: 1588-aware Capability TLV where the value is TBD

Length, 16 bits: Gives the length of the flags field in octets, and is currently set to 1

Flags, 8 bits: The bits are defined least-significant-bit (LSB) first, so bit 7 is the least significant bit of the flags octet.

```

  0 1 2 3 4 5 6 7
+---+---+---+---+
|   Reserved   |C|
+---+---+---+---+

```

Figure 4: Flags Format

Correction (C) field Update field, 1 bit: Setting the C bit to 1 indicates that the link is capable of recognizing the PTP event packets and can compensate for residence time by updating the PTP packet Correction Field. When this is set to 0, it means that this link cannot perform the residence time correction but is capable of performing MPLS frame forwarding of the frames with PTP labels using a method that support the end to end delivery of accurate timing. The exact method is not defined herein.

Reserved, 7 bits: Reserved for future use. The reserved bits must be ignored by the receiver.

The 1588-aware Capability sub-TLV is applicable to both OSPFv2 and OSPFv3.

14.2. 1588aware Link Capability for IS-IS

The IS-IS Traffic Engineering [RFC3784] defines the intra-area traffic engineering enhancements and uses the Extended IS Reachability TLV (Type 22) [RFC5305] to carry the per link TE-related information. This extension defines a new 1588-aware capability sub-TLV that can be carried as part of the Extended IS Reachability TLV.

The 1588-aware capability sub-TLV is OPTIONAL and MUST NOT appear more than once within the Extended IS Reachability TLV or the Multi-Topology (MT) Intermediate Systems TLV (type 222) specified in [RFC5120]. If a second instance of the 1588-aware capability sub-TLV is present, the receiving system MUST only process the first instance of the sub-TLV.

The format of the IS-IS 1588-aware sub-TLV is identical to the TLV format used by the Traffic Engineering Extensions to IS-IS [RFC3784]. That is, the TLV is comprised of 1 octet for the type, 1 octet

specifying the TLV length, and a value field. The Length field defines the length of the value portion in octets.

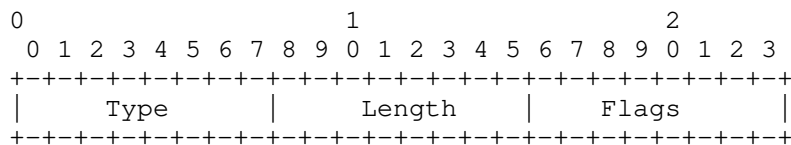


Figure 5: 1588-aware Capability sub-TLV

Where:

Type, 8 bits: 1588-aware Capability sub-TLV where the value is TBD

Length, 8 bits: Gives the length of the flags field in octets, and is currently set to 1

Flags, 8 bits: The bits are defined least-significant-bit (LSB) first, so bit 7 is the least significant bit of the flags octet.

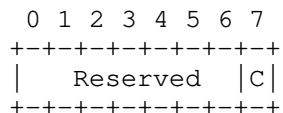


Figure 6: Flags Format

Correction (C) field Update field, 1 bit: Setting the C bit to 1 indicates that the link is capable of recognizing the PTP event packets and can compensate for residence time by updating the PTP packet Correction Field. When this is set to 0, it means that this link cannot perform the residence time correction but is capable of performing MPLS frame forwarding of the frames with PTP labels using a method that support the end to end delivery of accurate timing. The exact method is not defined herein.

Reserved, 7 bits: Reserved for future use. The reserved bits must be ignored by the receiver.

15. RSVP-TE Extensions for support of 1588

RSVP-TE signaling MAY be used to setup the PTP LSPs. A new RSVP object is defined to signal that this is a PTP LSP. The OFFSET to the start of the PTP message header MAY also be signaled. Implementations can trivially locate the correctionField (CF) location given this information. The OFFSET points to the start of the PTP header as a node may want to check the PTP messageType before it touches the correctionField (CF). The OFFSET is counted from TBD

The LSRs that receive and process the RSVP-TE/GMPLS messages MAY use the OFFSET to locate the start of the PTP message header.

Note that the new object/TLV Must be ignored by LSRs that are not compliant to this specification.

The new RSVP 1588_PTP_LSP object should be included in signaling PTP LSPs and is defined as follows:

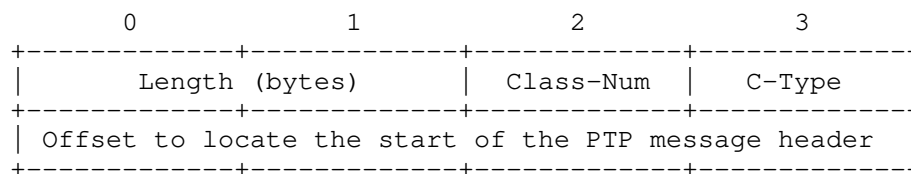


Figure 7: RSVP 1588_PTP_LSP object

The ingress LSR MUST include this object in the RSVP PATH Message. It is just a normal RSVP path that is exclusively set up for PTP messages

16. Behavior of LER/LSR

16.1. Behavior of 1588-aware LER

A 1588-aware LER advertises it's 1588-awareness via the OSPF procedure explained in earlier section of this specification. The 1588-aware LER then signals PTP LSPs by including the 1588_PTP_LSP object in the RSVP-TE signaling.

When a 1588 message is received from a non-MPLS interface, the LER MUST redirect them to a previously established PTP LSP. When a 1588 over MPLS message is received from an MPLS interface, the processing is similar to 1588-aware LSR processing.

16.2. Behavior of 1588-aware LSR

1588-aware LSRs are LSRs that understand the 1588_PTP_LSP RSVP object and can perform 1588 processing (e.g. Transparent Clock processing).

A 1588-aware LSR advertises it's 1588-awareness via the OSPF procedure explained in earlier section of this specification.

When a 1588-aware LSR distributes a label for PTP LSP, it maintains this information. When the 1588-aware LSR receives an MPLS packet, it performs a label lookup and if the label lookup indicates it is a PTP label then further parsing must be done to positively identify that the payload is 1588 and not OAM, BFD or control and management. Ruling out non-1588 messages can easily be done when parsing indicates the presence of GAL, ACH or VCCV (Type 1, 2, 3) or when the UDP port number does not match one of the 1588 UDP port numbers.

After a 1588 message is positively identified in a PTP LSP, the PTP message type indicates whether any timestamp processing is required. After 1588 processing the packet is forwarded as a normal MPLS packet to downstream node.

16.3. Behavior of non-1588-aware LSR

It is most beneficial that all LSRs in the path of a PTP LSP be 1588-aware LSRs. This would ensure the highest quality time and clock synchronization by 1588 Slave Clocks. However, this specification does not mandate that all LSRs in path of a PTP LSP be 1588-aware.

Non-1588-aware LSRs are LSRs that either don't have the capability to process 1588 packets (e.g. perform Transparent Clock processing) or don't understand the 1588_PTP_LSP RSVP object.

Non-1588-aware LSRs ignore the RSVP 1588_PTP_LSP object and just

switch the MPLS packets carrying 1588 messages as data packets and don't perform any timestamp related processing. However as explained in QoS section the 1588 over MPLS packets MUST be still be treated with the highest priority.

17. Other considerations

The use of Explicit Null (Label= 0 or 2) is acceptable as long as either the Explicit Null label is the bottom of stack label (applicable only to UDP/IP encapsulation) or the label below the Explicit Null label is a PTP label.

The use of Penultimate Hop Pop (PHP) is acceptable as long as either the PHP label is the bottom of stack label (applicable only to UDP/IP encapsulation) or the label below the PHP label is a PTP label.

18. Security Considerations

MPLS PW security considerations in general are discussed in [RFC3985] and [RFC4447], and those considerations also apply to this document.

An experimental security protocol is defined in [IEEE]. The PTP security extension and protocol provides group source authentication, message integrity, and replay attack protection for PTP messages.

19. Acknowledgements

The authors would like to thank Luca Martini, Ron Cohen, Yaakov Stein, Tal Mizrahi and other members of the TICTOC WG for reviewing and providing feedback on this draft.

20. IANA Considerations

20.1. IANA Considerations for OSPF

IANA has defined a sub-registry for the sub-TLVs carried in an OSPF TE Link TLV (type 2). IANA is requested to assign a new sub-TLV codepoint for the 1588aware capability sub-TLV carried within the Router Link TLV.

Value	Sub-TLV	References
-----	-----	-----
TBD	1588aware node sub-TLV	(this document)

20.2. IANA Considerations for IS-IS

IANA has defined a sub-registry for the sub-TLVs carried in the IS-IS Extended IS Reacability TLV. IANA is requested to assign a new sub-TLV code-point for the 1588aware capability sub-TLV carried within the Extended IS Reacability TLV.

Value	Sub-TLV	References
-----	-----	-----
TBD	1588aware node sub-TLV	(this document)

20.3. IANA Considerations for RSVP

IANA is requested to assign a new Class Number for 1588 PTP LSP object that is used to signal PTP LSPs.

1588 PTP LSP Object

Class-Num of type 11bbbbbb

Suggested value TBD

Defined CType: 1 (1588 PTP LSP)

21. References

21.1. Normative References

- [IEEE] IEEE 1588-2008, "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems".
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4389] Thaler, D., Talwar, M., and C. Patel, "Neighbor Discovery Proxies (ND Proxy)", RFC 4389, April 2006.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [RFC4720] Malis, A., Allan, D., and N. Del Regno, "Pseudowire Emulation Edge-to-Edge (PWE3) Frame Check Sequence Retention", RFC 4720, November 2006.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

21.2. Informative References

- [I-D.ietf-pwe3-fat-pw] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS Packet Switched Network", draft-ietf-pwe3-fat-pw-07 (work in progress), July 2011.

- [ISO] ISO/IEC 10589:1992, "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)".
- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3784] Smit, H. and T. Li, "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)", RFC 3784, June 2004.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2008.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

Authors' Addresses

Shahram Davari
Broadcom Corp.
San Jose, CA 95134
USA

Email: davari@broadcom.com

Amit Oren
Broadcom Corp.
San Jose, CA 95134
USA

Email: amito@broadcom.com

Manav Bhatia
Alcatel-Lucent
Bangalore,
India

Email: manav.bhatia@alcatel-lucent.com

Peter Roberts
Alcatel-Lucent
Kanata,
Canada

Email: peter.roberts@alcatel-lucent.com

Laurent Montini
Cisco Systems
San Jose CA
USA

Email: lmontini@cisco.com

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: October 30, 2012

Parag Jain, Ed.
Sami Boutros
Cisco Systems, Inc.

Sam Aldrin
Huawei Technologies

May 4, 2012

Definition of P2MP PW TLV for LSP-Ping Mechanisms
draft-jain-pwe3-p2mp-pw-lsp-ping-00.txt

Abstract

LSP-Ping is a widely deployed Operation, Administration, and Maintenance (OAM) mechanism in MPLS networks. This document describes a mechanism to verify connectivity of Point-to-Multipoint (P2MP) Pseudowires (PW) using LSP Ping.

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), 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 December 28, 2011.

Copyright Notice

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

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

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Terminology	3
4. Identifying a P2MP PW	3
4.1. FEC 130 Pseudowire Sub-TLV	4
5. Operations	4
6. Echo Reply using Downstream Assigned Label	6
7. Controlling Echo Responses	6
8. Security Considerations	6
9. IANA Considerations	6
10. References	6
10.1. Normative References	6
10.2. Informative References	7
11. Acknowledgments	7

1. Introduction

A Point-to-Multipoint (P2MP) Pseudowire (PW) emulates the essential attributes of a unidirectional P2MP Telecommunications service such as P2MP ATM over PSN. Requirements for P2MP PW are described in [PPWREQ]. P2MP PWs are carried over P2MP MPLS LSP. The Procedure for P2MP PW signaling using LDP for single segment P2MP PWs are described in [PPWPWE3]. Many P2MP PWs can share the same P2MP MPLS LSP and this arrangement is called Aggregate P-tree. The aggregate P2MP trees require an upstream assigned label so that on the tail of the P2MP LSP, the traffic can be associated with a VPN or a VPLS instance. When a P2MP MPLS LSP carries only one VPN or VPLS service instance, the arrangement is called Inclusive P-Tree. For Inclusive P-Trees, P2MP MPLS LSP label itself can uniquely identify the VPN or VPLS service being carried over P2MP MPLS LSP. The P2MP MPLS LSP can also be used in Selective P-Tree arrangement for carrying multicast traffic. In a Selective P-Tree arrangement, traffic to each multicast group in a VPN or VPLS instance is carried by a separate

unique P-tree. In Aggregate Selective P-tree arrangement, traffic to a set of multicast groups from different VPN or VPLS instances is carried over a same shared P-tree.

The P2MP MPLS LSP are setup either using MLDP [MLDP] or P2MP RSVP-TE [RFC4875]. Mechanisms for fault detection and isolation for data plane failures for P2MP MPLS LSPs are specified in [PLSPING]. This document describes a mechanism to detect data plane failures for P2MP PW carried over P2MP MPLS LSPs.

This document defines a new FEC 130 Pseudowire sub-TLV for Target FEC Stack for P2MP PW. The FEC 130 Pseudowire sub-TLV is added in Target FEC Stack TLV by the originator of the echo request to inform the receiver at P2MP MPLS LSP tail, of the P2MP PW being tested.

Multi-segment Pseudowires support is out of scope of this document at present and may be included in future.

2. Conventions used in this document

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

The term "FEC-Type" is used to refer to a tuple consisting of <FEC Element Type, Address Family>.

3. Terminology

ATM: Asynchronous Transfer Mode

LSR: Label Switching Router

MPLS-OAM: MPLS Operations, Administration and Maintenance

P2MP-PW: Point-to-Multipoint PseudoWire

PW: PseudoWire

TLV: Type Length Value

4. Identifying a P2MP PW

This document introduces a new LSP Ping Target FEC Stack sub-TLV, FEC 130 Pseudowire sub-TLV, to identify the P2MP PW under test at the P2MP LSP Tail/Bud node.

4.1. FEC 130 Pseudowire Sub-TLV

The FEC 130 Pseudowire sub-TLV fields are taken from P2MP PW FEC Element (FEC Type 0x82) defined in [PPWPWE3]. The PW Type is a 15-bit number indicating the encapsulation type. It is carried right justified in the field below PW Type with the high-order bit set to zero. All the other fields are treated as opaque values and copied directly from P2MP PW FEC Element (FEC Type 0x82) format.

The FEC 130 Pseudowire sub-TLV has the format shown in Figure 1. This TLV will be included in the echo request sent over P2MP PW by the originator of request.

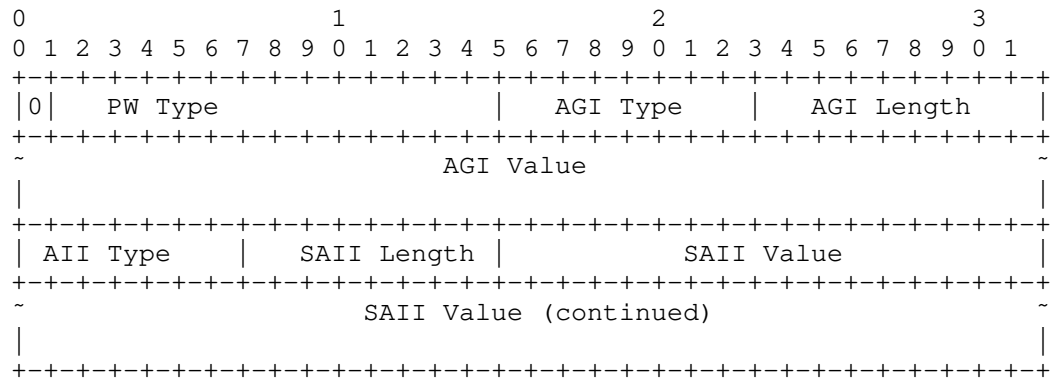


Figure 1: FEC 130 Pseudowire sub-TLV format

For Inclusive and Selective P2MP MPLS P-trees, the echo request will be sent using the P2MP MPLS LSP label.

For Aggregate Inclusive and Aggregate Selective P-trees, the echo request will be sent using a label stack of <P2MP MPLS P-tree label, upstream assigned P2MP PW label>. The P2MP MPLS P-tree label is the outer label and upstream assigned P2MP PW label is inner label.

5. Operations

In this section, we explain the operation of the LSP Ping over P2MP PW. Figure 2 shows a P2MP PW PW1 setup from T-PE1 to remote PEs (T-

PE2, T-PE3 and T-PE4). The transport LSP associated with the P2MP PW1 can be MLDP P2MP MPLS LSP or P2MP TE tunnel.

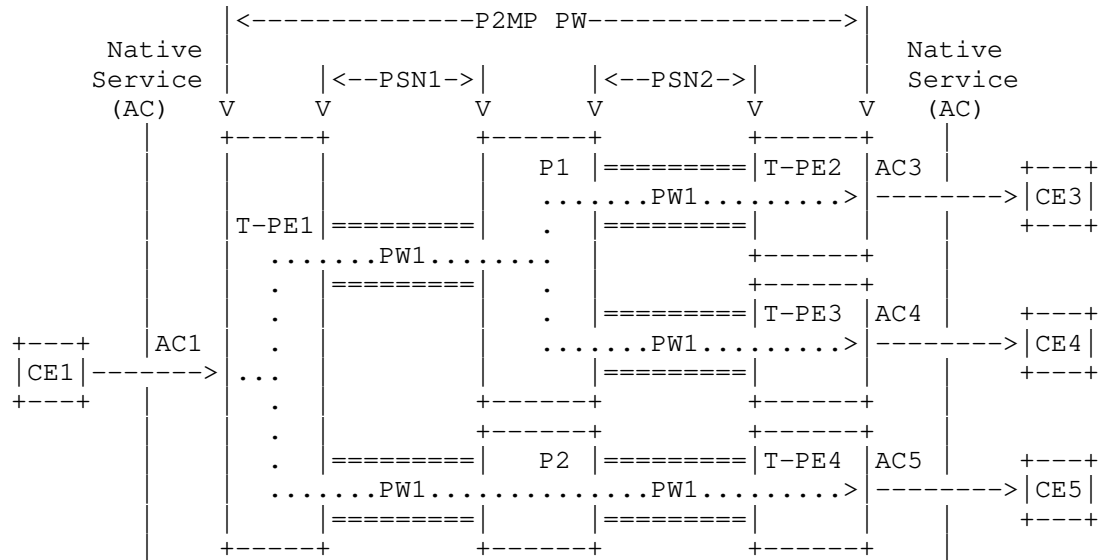


Figure 2: P2MP PW

When an operator wants to perform a connectivity check for the P2MP PW1, the operator initiates a LSP-Ping request with the Target FEC Stack TLV containing FEC 130 Pseudowire sub-TLV in the echo request packet. The echo request packet is sent over the P2MP MPLS LSP using the P2MP MPLS LSP label for Inclusive P-tree or with a label stack with Upstream assigned P2MP PW label as bottom label and P2MP MPLS LSP label as the top label. The intermediate P router will do swap and replication based on the MPLS LSP label. Once the packet reaches remote terminating PEs, the T-PEs will process the packet and perform checks for the FEC 130 Pseudowire sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC4379] and respond according to [RFC4379] processing rules.

6. Echo Reply using Downstream Assigned Label

Root of a P2MP PW may send an optional downstream assigned p2p MPLS label in the LDP Label Mapping message for the P2MP PW signaling. If the root of a P2MP PW expects leaf to send echo reply using the downstream assigned label signaled in the Label Mapping message of the P2MP PW message, the Reply Mode value of 4 "Reply via application level control channel" should be used in Reply Mode field described in Section 3 in [RFC4379] in echo request message for the P2MP PW.

7. Controlling Echo Responses

The procedures described in [PLSPING] for preventing congestion of Echo Responses (Echo Jitter TLV) and limiting the echo reply to a single egress node (Node Address P2MP Responder Identifier TLV) can be applied to P2MP PW LSP Ping.

8. Security Considerations

The proposal introduced in this document does not introduce any new security considerations beyond that already apply to [PLSPING].

9. IANA Considerations

This document defines a new sub-TLV type to be included in Target FEC Stack TLV (TLV Type 1) [RFC4379] in LSP Ping.

IANA is requested to assign a sub-TLV type value to the following sub-TLV from the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

FEC 130 Pseudowire sub-TLV (See Section 3). Suggested value 24.

10. References

10.1. Normative References

- [RFC4379] K. Kompella, G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [PPWPWE3] Martini, L. et. al, "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-pwe3-p2mp-pw-03.txt, Work in Progress, March 2011.

[PLSPPING] Saxena, S et. Al, "Detecting Data Plane Failures in Point-to-Multipoint Multiprotocol Label Switching (MPLS) - Extensions to LSP. draft-ietf-mpls-p2mp-lsp-ping-17, Work in Progress, June 2011

10.2. Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.
- [RFC5085] T. Nadeau, et. al, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires ", RFC 5085, December 2007.
- [MLDP] Minei, I., Kompella, K., Wijnands, I., and Thomas, B., "LDP Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-10.txt, Work in Progress, July 2010.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and Yasukawa, S., "Extensions to Resource Reservation Protocol" Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [PPWREQ] F. Jounay, et. al, "Requirements for Point to Multipoint Pseudowire", draft-ietf-pwe3-p2mp-pw-requirements-03.txt, Work in Progress, August 2010.

11. Acknowledgments

The authors would like to thank Shaleen Saxena, Michael Wildt, Tomofumi Hayashi, Danny Prairie for their valuable input and comments.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Parag Jain
Cisco Systems, Inc.,
2000 Innovation Drive,
Kanata, ON K2K3E8, Canada.
E-mail: paragj@cisco.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way,
San Jose, CA 95134, USA.
E-mail: sboutros@cisco.com

Sam Aldrin
Huawei Technologies, co.
2330 Central Express Way,
Santa Clara, CA 95051, USA.
E-mail: aldrin.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2013

L. Jin
R. Chen
ZTE
S. Boutros
Cisco Systems
S. Kini
Ericsson
July 6, 2012

Static pseudowire configuration checking using Generic Associated
Channel (G-ACh) Advertisement Protocol
draft-jc-pwe3-static-config-check-00.txt

Abstract

This draft defines a method to verify the configuration parameters of static pseudowires (PW). Since a static PW can be independently provisioned at each end of the PW there is a potential for a configuration parameter mismatch and this can result in the PW not being operational. The procedures in this draft intend to solve this problem and simplify the provisioning.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 7, 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. Terminology	3
3. GAP Extensions	4
3.1. Static PW Application Message	4
3.2. PE Procedure	7
3.2.1. Sending PW application Element TLV	7
3.2.2. Receiving PW application Element TLV	7
3.2.3. PW Configuration Verification Process	8
3.2.4. Remote Label Advertisement	8
4. Security Considerations	8
5. IANA Considerations	8
6. Acknowledgements	9
7. References	9
7.1. Normative references	9
7.2. Informative References	9
Authors' Addresses	10

1. Introduction

The manual configuration of static PW in MPLS and MPLS-TP network requires configuring different PW parameters at the two terminating PEs (Provider Edge). The PW parameters include PW-id, PW-Type, Control word setting, interface and VCCV parameters settings.

The PW provisioned parameters MUST be aligned, so as to make the PW operational. For dynamically signaled PW, the PW parameters are negotiated using the signaling protocol, and only when the PW parameters match at the terminating PE end points, the P2P (Point-to-Point) PW is made operational and can be used to forward data traffic.

In the absence of a signaling protocol, this draft defines a method to do static PW configuration verification, so as to ease the troubleshooting of end to end static PW provisioning in both MPLS and MPLS-TP networks. The mechanism to exchange the PW configuration parameters uses the Generic Associated Channel (G-ACH) Advertisement Protocol (GAP) defined in [I-D.ietf-mpls-gach-adv]. In this draft, the GAP functionality assumes that the PW's underlying PSN Tunnel with GAP enabled is operational.

In the following sections we will describe the extension to the GAP mechanism to do the PW configuration verification at the two terminating PEs for P2P PW. The P2MP (Point-to-Multipoint) PW configuration verification is for further study.

2. Terminology

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].

This document uses some terms and acronyms as follows:

MPLS: Multi Protocol Label Switching.

OAM: MPLS Operations, Administration and Maintenance.

PE: Provide Edge Node.

PW: PseudoWire.

TLV: Type, Length, and Value.

VPLS: Virtual Private LAN Services.

MS-PW: Multi-segment PseudoWire

3. GAP Extensions

3.1. Static PW Application Message

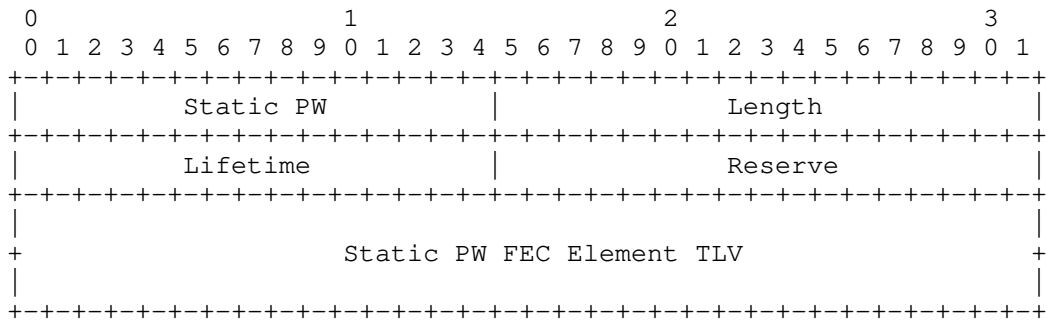


Figure 1

A new GAP application "Static PW" is defined in this draft. The Static PW Application ID is to be assigned by IANA, and suggested value is 0x0002.

Length: as per [I-D.ietf-mppls-gach-adv].

Lifetime: as per [I-D.ietf-mppls-gach-adv], and the default value is suggested to be 120 seconds.

Static PW FEC Element TLV for "Static PW" GAP application:

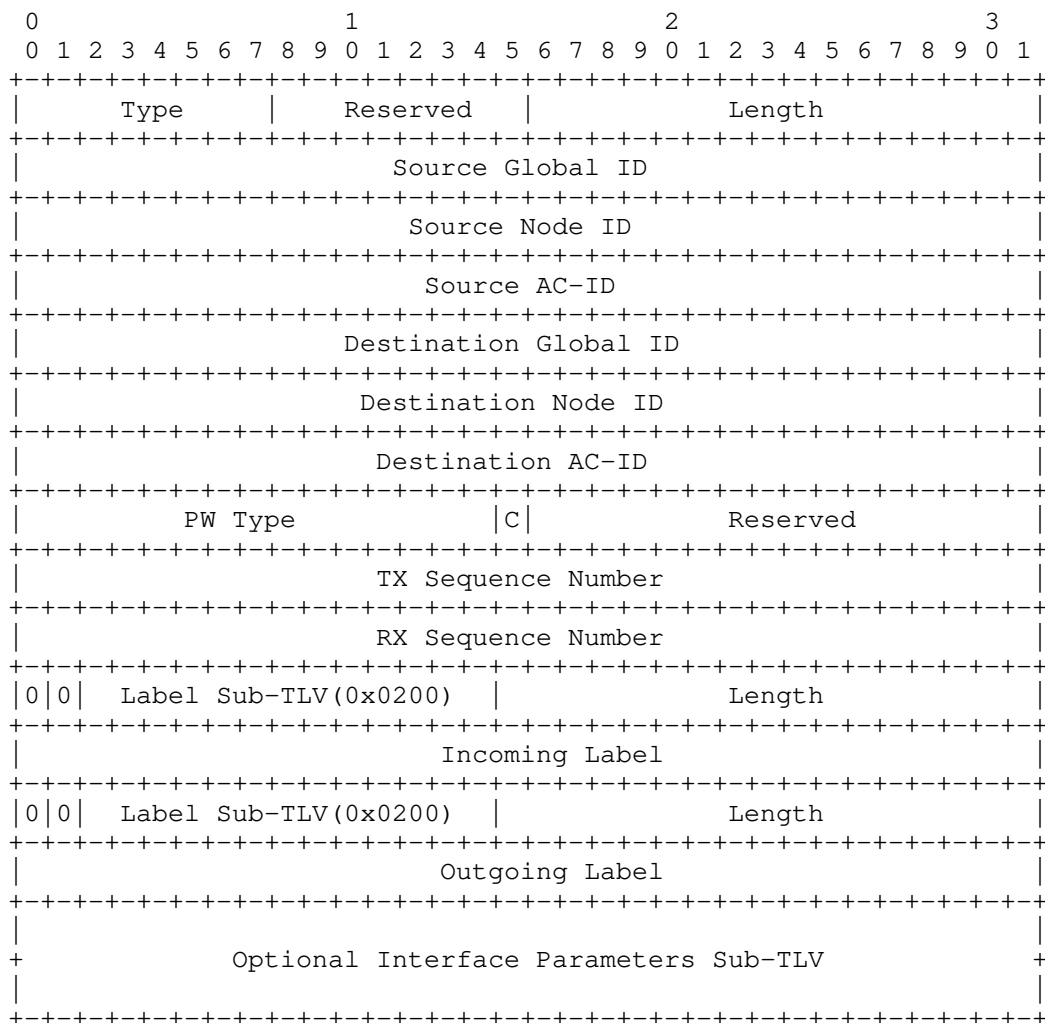


Figure 2

The Static PW FEC Element TLV type is to be assigned by IANA. The Length field specifies the length in octets of the Static PW FEC Element and all Optional Interface Parameters Sub-TLVs.

The Static PW FEC element TLV value MUST include the following:

- o The Global ID and Node ID fields MUST be set as per [RFC6370].

- o The AC-ID fields MUST be set as per [RFC5003].
- o PW-Type and control word bit (C) MUST be set as per [RFC4447].
- o TX Sequence Number: The transmitted message sequence number for the associated Static PW FEC Element TLV.
- o RX Sequence Number: The last received sequence number for the associated Static PW FEC Element TLV.
- o Two Generic Label TLVs as defined in [RFC5036] to encode static PW incoming and outgoing labels in the order shown above.
- o Optional Interface parameters Sub-TLV as defined in [RFC4447].

The GAP Suppress message defined in [I-D.ietf-mppls-gach-adv] only applies all TLVs for a given application. We define a new TLV, static PW suppress TLV, to suppress static PW FEC element transmission. Multiple static PW FEC element TLVs could be included in this TLV. The format would be as follows:

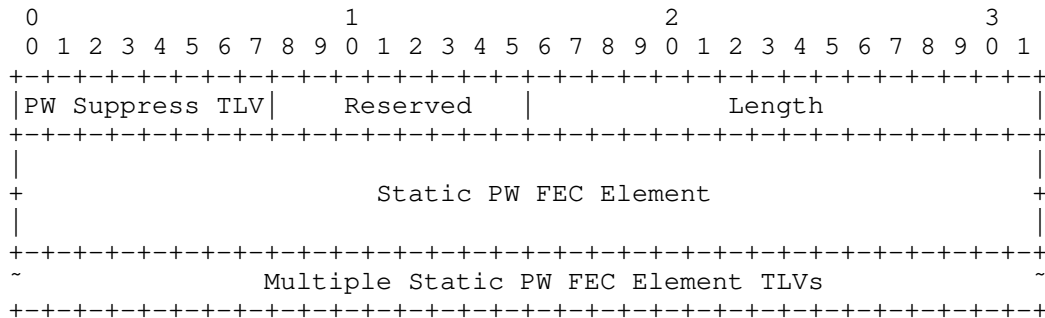


Figure 3

The type of static PW suppress TLV is to be assigned by IANA.

The static PW suppress TLV could be sent by a receiving PE to request a transmitting PE to stop sending GAP messages for the static PW FEC Element TLVs in the static PW suppress TLV.

The static PW application MUST follow all procedures defined in [I-D.ietf-mppls-gach-adv].

3.2. PE Procedure

The mechanism defined in this draft provides a verification tool for the P2P PW configuration information between two PEs. Upon the provisioning or re-provisioning of a PW at an endpoint PE, GAP messages carrying the static PW application TLV will be sent over the PW's corresponding PSN tunnel which the endpoints PEs of the P2P PW selects by local policy.

3.2.1. Sending PW application Element TLV

When a PW is configured at one endpoint PE, and the PW corresponding PSN Tunnel is operational and UP, the PE MUST send its local PW configuration information using the GAP over the PSN tunnel.

The transmitting PE MUST set the TX sequence number to a non-zero value in Static PW FEC Element TLV, and MUST increment the TX sequence number each time any local PW parameters change.

If the transmitting PE has previously received a GAP message with the static PW FEC Element, the transmitting PE MUST verify local PW parameters with the remote PE parameters as specified in section 4.2.3. The RX sequence number MUST be set to the previously received TX sequence number, otherwise set to zero.

3.2.2. Receiving PW application Element TLV

The receiving PE MUST update the remote PW parameters associated with a static PW FEC Element TLV, when the received TX sequence number in the GAP message is different from the last one received.

If the receiving PE has been provisioned locally with the PW parameters and has previously sent GAP message for the PW parameters, it MUST check if the RX sequence number in the received GAP message is equal to the TX sequence number it previously sent.

If the RX sequence number is equal, the receiving PE MUST send GAP message with static PW suppress TLV as a response to remote PE, and then verify local static PW parameters with the remote static PW FEC parameters as specified in section 3.2.3.

Otherwise, if the RX sequence number is not equal, the receiving PE MUST continue sending GAP message with static PW FEC element TLV, with the RX sequence number set to the last received TX sequence number from the remote PE.

If there is no local PW configuration associated with the static PW FEC Element TLV, the receiving PE MUST retain the remote static PW

FEC Element information.

Whenever PE receives the GAP message with static PW suppress TLV, it MUST stop sending GAP messages with the specified static PW FEC element TLVs included in the static suppress TLV.

The GAP message of static PW application SHOULD be sent at least three times within lifetime.

The mechanism described above applies as well for MS-PW.

3.2.3. PW Configuration Verification Process

Using source/destination Global-IDs, and source/destination node-ID and AC-IDs, to identify a locally provisioned static PW, once found, perform the following parameter verification checks:

1. Check the control word bit (C), and MUST do logical operation "AND". Only when both ends have the use of control word enabled, the result would be with control word presented on this PW.
2. Check PW type mismatch as defined in [RFC4447].
3. Check and negotiate interface parameters as defined in [RFC4447].
4. Check incoming and outgoing static PW labels. The local incoming label should be equal to remote outgoing label, and the local outgoing label should be equal to remote incoming label, otherwise checking failed.

3.2.4. Remote Label Advertisement

The mechanism described in this draft MAY also be used to communicate local static PW labels to allow for single side provisioning of labels. As such, only incoming label will be included in the GAP message and this label will be used by the remote PE as the output label for the PW.

4. Security Considerations

The mechanisms defined in this draft do not introduce any new threats more than what's described in [I-D.ietf-mpls-gach-adv].

5. IANA Considerations

IANA is requested to allocate a new "Static PW" Application ID in the

"G-Ach Advertisement Protocol Applications" registry.

Application ID	Description	Reference
(TBD)	Static PW Application	(this draft)

This document requests that IANA create a new registry, "GAP Static PW Application: TLV objects", with fields and initial value as follows:

Type Name	Type ID	Reference
Static PW FEC Element	0	(this draft)
Static PW suppress TLV	1	(this draft)

The range of the Type ID field is 0 - 255.

The allocation policy for this registry is IETF Review.

6. Acknowledgements

The authors would like to thank Stewart Bryant, Dan Frost for their review and contributions.

7. References

7.1. Normative references

[I-D.ietf-mppls-gach-adv]
Frost, D., Bryant, S., and M. Bocci, "MPLS Generic Associated Channel (G-ACh) Advertisement Protocol", draft-ietf-mppls-gach-adv-02 (work in progress), May 2012.

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

7.2. Informative References

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC5003] Metz, C., Martini, L., Balus, F., and J. Sugimoto, "Attachment Individual Identifier (AII) Types for Aggregation", RFC 5003, September 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP

Specification", RFC 5036, October 2007.

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.

Authors' Addresses

Lizhong Jin
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China

Email: lizhong.jin@zte.com.cn

Ran Chen
ZTE Corporation
No.19 East Huayuan Road
Beijing, 100191, China

Email: chen.ran@zte.com.cn

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way
San Jose, California 95134
USA

Email: sboutros@cisco.com

Sriganesh Kini
Ericsson
Ericsson
San Jose, CA 95134

Email: sriganesh.kini@ericsson.com

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: December 16, 2012

G. Manhoudt
AimValley

S. Roullot
Alcatel-Lucent

P. Roberts
Alcatel-Lucent

July 09, 2012

Transparent SDH/SONET over Packet
draft-manhoudt-pwe3-tsop-00

Abstract

This document describes the Transparent SDH/SONET over Packet (TSoP) mechanism to encapsulate Synchronous Digital Hierarchy (SDH) or Synchronous Optical NETwork (SONET) bit-streams in a packet format, suitable for Pseudowire (PW) transport over a packet switched network (PSN). The key property of the TSoP method is that it transports the SDH/SONET client signal in its entirety through the PW, i.e., no use is made of any specific characteristic of the SONET/SDH signal format, other than its bit rate. The TSoP transparency includes transporting the timing properties of the SDH/SONET client signal. This ensures a maximum of transparency and a minimum of complexity, both in implementation and during operation.

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

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

Table of Contents

1. Introduction	4
2. Terminology and Conventions	5
2.1. Conventions Used in This Document	5
2.2. Acronyms and Terms	5
3. Emulated STM-N Services	6
3.1 PSN-bound direction	8
3.1 CE-bound direction	9
4. TSoP Encapsulation Layer	11
4.1. TSoP Packet Format	11
4.2. PSN/PW Headers	11
4.2.1 Transport over an MPLS (-TP) PSN	11
4.2.2 Transport over an IPv4/IPv6 PSN	12
4.3. TSoP Encapsulation Headers	12
4.3.1. Location and Order of TSoP Encapsulation Headers	12
4.3.2. Usage and Structure of the TSoP Control Word	14
4.3.3. Usage of the RTP Header	15
5. TSoP Payload Field	17
6. TSoP Operation	17
6.1. Common Considerations	17
6.2. IWF Operation	17
6.2.1. PSN-Bound Direction	17
6.2.2. CE-Bound Direction	18
6.3. TSoP Defects	20
6.4. TSoP Performance Monitoring	21
7. Quality of Service (QoS) Issues	23
8. Congestion Control	23
9. Security Considerations	24
10. Applicability Statements	25
11. IANA Considerations	26
12. Acknowledgements	26
13. References	26
13.1. Normative References	26
13.2. Informative References	27
Appendix A. Parameters to be configured to set up a TSoP PW	29
Authors' Addresses	30

1. Introduction

This document describes the Transparent SDH/SONET over Packet (TSoP) method for encapsulating SDH or SONET signals with bit rates of 51.84 Mbit/s or $N * 155.52$ Mbit/s (where $N = 1, 4, 16$ or 64) for Pseudowire (PW) transport over a packet switched network (PSN), using circuit emulation techniques.

The selected approach for this encapsulation scheme avoids using any particular signal characteristics of the SDH/SONET signal, other than its bit rate. This approach closely follows the SAToP method described in [RFC4553] for PW transport of E1, DS1, E3 or DS3 over a PSN.

An alternative to the TSoP method for STM-N transport over PW is known as CEP (Circuit Emulation over Packet) and is described in [RFC4842]. The key difference between the CEP approach and the TSoP approach is that within CEP an incoming STM-N is terminated and demultiplexed to its constituent VCs (Virtual Containers). Subsequently, each VC is individually circuit emulated and encapsulated into a PW and transported over the PSN to potentially different destinations, where they are reassembled into (newly constructed) STM-N signals again. The TSoP approach, on the other hand, is to encapsulate the entire STM-N in a single circuit emulating Pseudowire and transport it to a single destination over the PSN. The essential difference between both methods is that CEP offers more routing flexibility and better bandwidth efficiency than TSoP at the cost of the loss of transparency (overhead, timing, scrambling) at the STM-N layer and at the cost of added complexity associated with the inclusion of what in essence is an SDH/SONET VC cross-connect function in the PEs.

Within the context of this document, there is no difference between SONET [GR-253] signals, often denoted as OC-M, and SDH [G.707] signals, usually denoted as STM-N. For ease of reading, this document will only refer to STM-N, but any statement about an STM-N signal should be understood to apply equally to the equivalent OC-M signal, unless it is specifically mentioned otherwise. The equivalency can be described by the following relations between N and M: If $N = 0$ then $M = 1$ and if $N \geq 1$ then $M = 3 * N$.

The TSoP solution presented in this document conforms to the PWE3 architecture described in [RFC3985] and satisfies the relevant general requirements put forward in [RFC3916].

As with all PWs, TSoP PWs may be manually configured or set up using the PWE3 control protocol [RFC4447].

2. Terminology and Conventions

2.1. 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].

2.2. Acronyms and Terms

The following acronyms used in this document are defined in [RFC3985], [RFC4197] and [RFC4553]:

AC	Attachment Circuit
ATM	Asynchronous Transfer Mode
CE	Customer Edge
CES	Circuit Emulation Service
IWF	Interworking Function
NSP	Native Service Processing
PE	Provider Edge
PREP	Pre-Processing
PSN	Packet Switched Network
PW	Pseudowire
SDH	Synchronous Digital Hierarchy
SONET	Synchronous Optical Network
TDM	Time Division Multiplexing

In addition, the following specific terms are used in this document:

LOF	Loss Of Frame - A condition of an STM-N signal in which the frame pattern cannot be detected. Criteria for raising and clearing a LOF condition can be found in [G.783].
LOS	Loss Of Signal - A condition of the STM-N attachment circuit in which the incoming signal has an insufficient energy level for reliable reception. Criteria for raising and clearing a LOS condition can be found in [G.783].
G-AIS	Generic Alarm Indication Signal - A specific bit pattern that replaces the normal STM-N signal in the case of certain failure scenarios. The G-AIS pattern [G.709] is constructed by continuously repeating the 2047 bit pseudo random bit sequence based on the generating polynomial $1 + x^9 + x^{11}$ according to [O.150].
NIM	Non-Intrusive Monitor - A circuit that monitors a signal in a certain direction of transmission, without changing the binary content of it. A NIM can be used for Fault Management

and Performance Monitoring purposes

- SF Signal Fail: A control signal, that exists internally in a system, to convey the failed state of an incoming signal, from a server layer process to the adjacent client layer process. See [G.783]
- LOPS Loss of Packet State - A defect that indicates that the PE at the receiving end of a TSoP carrying PW experiences an interruption in the stream of received TSoP packets. See [RFC5604]

3. Emulated STM-N Services

The TSoP emulated STM-N service over a Pseudowire makes use of a bi-directional point-to-point connection over the PSN between two TSoP-IWF blocks, located in the PE nodes that terminate the PW that interconnects them, as shown in figure 1. The TSoP-IWF blocks each consist of two half-functions, a PSN-bound IWF and a CE-bound IWF, one for each direction of transmission. As the name implies, the PSN-bound part of the TSoP IWF performs the conversion of an STM-N bitstream to a packet flow, suitable for transport over the PSN and the CE-bound part of the TSoP-IWF performs the inverse operation.

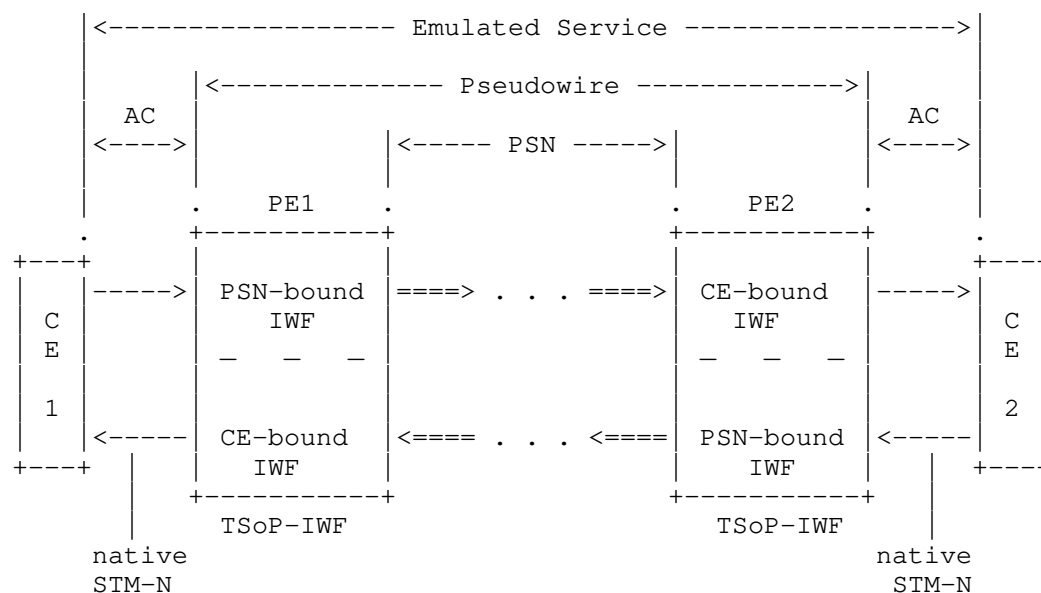


Figure 1. Overview of STM-N emulated service architecture

The following list provides the STM-N services, as specified in [G.707] and [GR-253], that can be supported by a TSoP PW:

1. STM-0 or OC-1 (51.84 Mbit/s)
2. STM-1 or OC-3 (155.52 Mbit/s)
3. STM-4 or OC-12 (622.08 Mbit/s)
4. STM-16 or OC-48 (2488.32 Mbit/s)
5. STM-64 or OC-192 (9953.28 Mbit/s)

The TSoP protocol used for emulation of STM-N services does not depend on the method in which the STM-N is delivered to the PE. For example, an STM-1 attachment circuit is treated in the same way regardless of whether it is a copper [G.703] or a fiber optic [G.707] link.

Also, in case the STM-N is carried in an OTN signal [G.709], the functionality in the TSoP-IWF operates in the same way, but a PWE3 Pre-processing (PREP) functional block will be present between the AC and the PE to perform the OTN (de)multiplexing functions.

The TSoP-IWF function in figure 1 is further broken down in functional blocks in figure 2. These individual functional blocks are described in the next two sections.

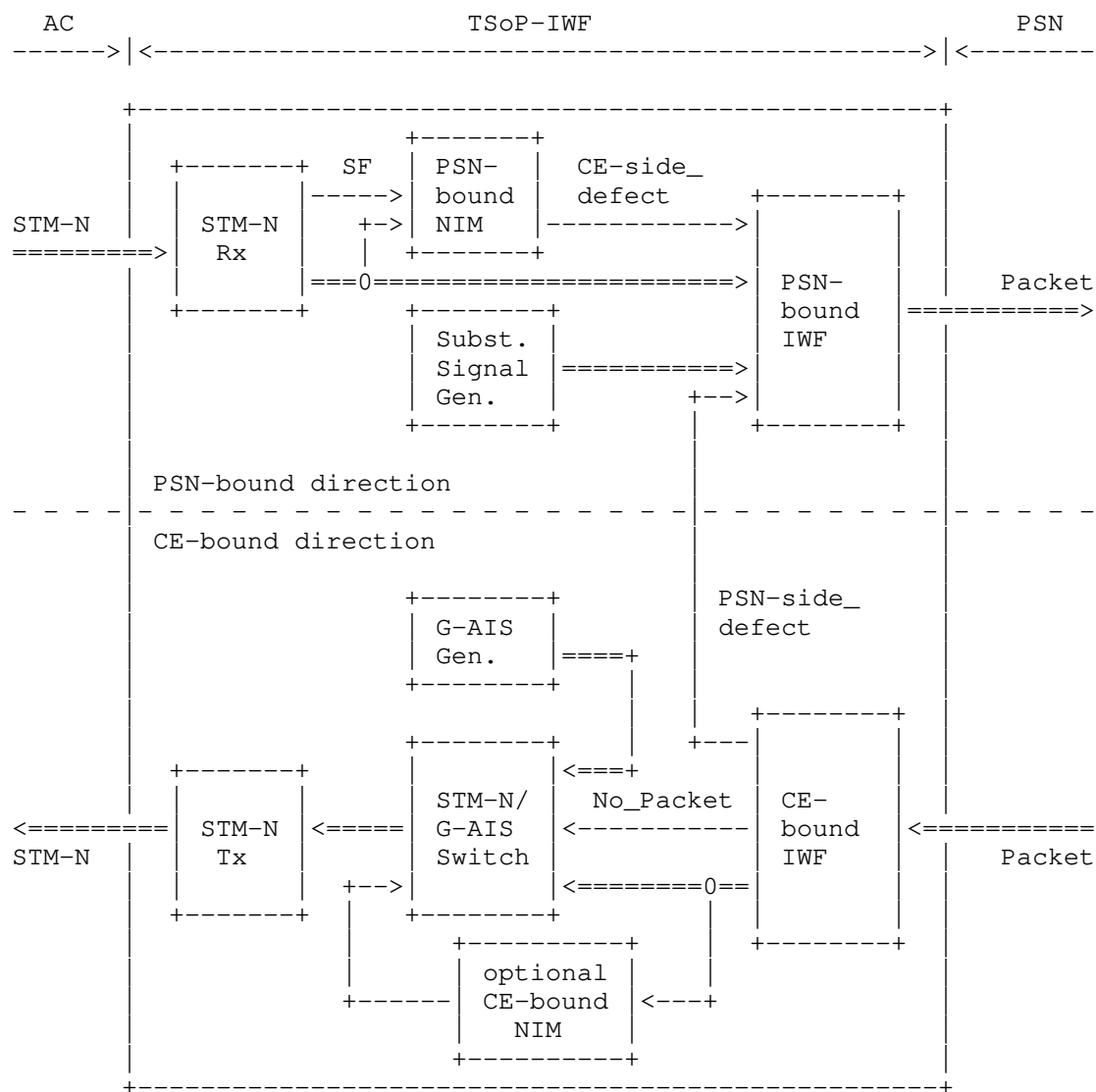


Figure 2. TSoP functional block diagram

3.1 PSN-bound direction

In the PSN-bound direction the STM-N signal is received from the CE via an AC by the STM-N Rx function. This function recovers the optical or electrical signal and converts it to a suitable internal format. In addition, it detects the LOS condition and it asserts the SF signal whenever this is the case. The STM-N Rx block is

equivalent to the OSn_TT_Sk & OSn/Rn_A_Sk (in the case of an optical STM-N) or the ESn_TT_Sk & ESn/Rn_A_Sk (in the case of an electrical STM-N interface) function pairs defined in [G.783].

The CE-bound IWF segments the STM-N ingress bitstream, which it receives from the STM-N Rx function, in blocks of equal length. Each block of bits is supplied with the appropriate TSoP Encapsulation Headers and then delivered to the PSN Multiplexing layer to add the required headers for transport over the PSN.

The PSN-bound NIM function controls the state of the CE-side_defect signal. It will assert this signal in case the SF signal is asserted or in case another defect is detected in the incoming STM-N signal. The inclusion of other defects than LOS in the CE-side_defect signal is OPTIONAL.

When the CE-side_defect signal is asserted, the PSN-bound IWF will set the corresponding flag (L-bit) in the overhead of the affected packets. Packets in which the L-bit is set MUST have a substitution payload (created by the Substitution Signal Generator function) of the same length as the regular TSoP payload. This substitution payload is RECOMMENDED to be the G-AIS pattern or a fixed "all ones" pattern.

Lastly, when the PSN-side_defect state is asserted, the PSN-bound IWF will set the corresponding flag (R-bit) in the overhead of all packets that are transmitted while this signal is in the asserted state.

3.1 CE-bound direction

In the CE-bound direction, the CE-bound IWF receives the PW packets from the PSN and strips off the PSN, PW, and TSoP encapsulation headers and writes the payload data in a buffer. The output data stream towards the CE is created by playing out this buffer with a suitable clock signal. The thus reconstructed STM-N signal is forwarded to the STM-N/G-AIS Switch function.

The No_Packet signal is asserted by the CE-bound IWF in case the internal packet buffer empties due to lack of input packets from the PSN or in case a packet is missing or invalid.

The PSN-side_defect signal is asserted by the CE-bound IWF in case the LOPS condition is detected by the CE-bound IWF (see section 6.2.2). The state of this signal controls the value of the R-bit in the overhead of the packets returned towards the far-end TSoP-IWF.

The G-AIS Generator generates a G-AIS signal at the nominal frequency

of the recovered STM-N signal, ± 20 ppm.

The STM-N/G-AIS Switch normally takes its input from the CE-bound IWF and forwards the recovered STM-N signal towards the STM-N Tx function, but during the time that the No_Packet signal is asserted, it will select the G-AIS Generator as its active input and forward a G-AIS signal towards the STM-N Tx function.

The CE-bound NIM function is an OPTIONAL function that can be used to detect additional defects in the recovered CE-bound STM-N signal. The presence of such defects (e.g. STM-N LOF) MAY be used as an additional reason for the STM-N/G-AIS Switch function to select the G-AIS signal as its active input.

Lastly, the STM-N Tx function converts the internal signal that is output by the STM-N/G-AIS Switch block into a regular STM-N signal towards the CE via the AC. The STM-N Tx block is equivalent to the OSn_TT_So & OSn_RSn_A_So (in the case of an optical STM-N) or the ESn_TT_So & ESn_RSn_A_So (in the case of an electrical STM-N interface) function pairs defined in [G.783].

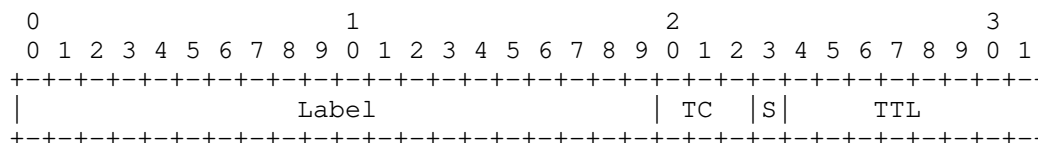


Figure 4. PW Label (S = 1)

4.2.2 Transport over an IPv4/IPv6 PSN

Both UDP and L2TPv3 [RFC3931] can provide the PW demultiplexing mechanisms for TSoP PWs over an IPv4/IPv6 PSN. The PW label (figure 4) provides the demultiplexing function for an IPv4/IPv6 PSN as described in [RFC3985]. The total length of a TSoP packet for a specific PW MUST NOT exceed path MTU between the pair of PEs terminating this PW. TSoP implementations using an IPv4 PSN MUST mark the IPv4 datagrams they generate as "Don't Fragment" [RFC791] (see also [RFC4623]).

4.3. TSoP Encapsulation Headers

4.3.1. Location and Order of TSoP Encapsulation Headers

The TSoP Encapsulation Headers MUST contain the TSoP Control Word (figure 7) and MUST contain a Minimum length RTP Header [RFC3550] (figure 8). The TSoP Encapsulation Headers must immediately follow the PSN/PW header, as shown in figure 3.

In case the TSoP packets are transmitted over a PSN based on UDP over IPv4/IPv6 technology, the TSoP Encapsulation Headers have the RTP Header first and then the TSoP control word immediately next, as shown in figure 6. In case the TSoP packets are transmitted over a PSN based on a technology other than UDP over IPv4/IPv6, the TSoP Encapsulation Headers have the TSoP control word first and then the RTP header immediately next, as shown in figure 5.

Note: This arrangement complies with the traditional usage of RTP for the IPv4/IPv6 PSN with UDP multiplexing while making TSoP PWs Equal Cost Multi-Path (ECMP)-safe for the MPLS PSN by providing for PW-IP packet discrimination (see [RFC3985]). Furthermore, it facilitates seamless stitching of L2TPv3-based and MPLS-based segments of TSoP PWs (see [RFC5254]).

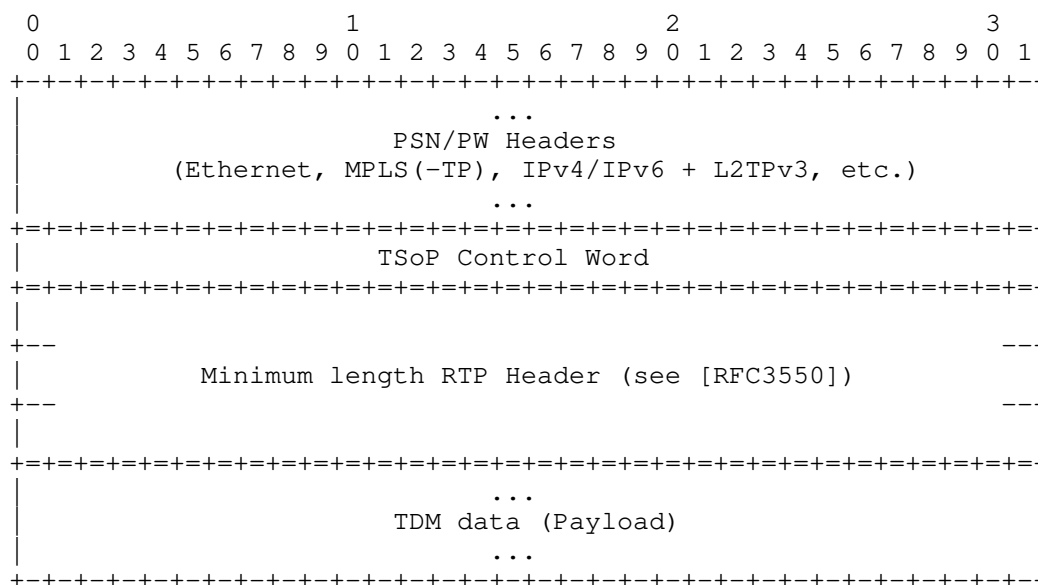


Figure 5. General TSoP Packet Format for all PSNs, other than an IPv4/IPv6 PSN with UDP PW Demultiplexing

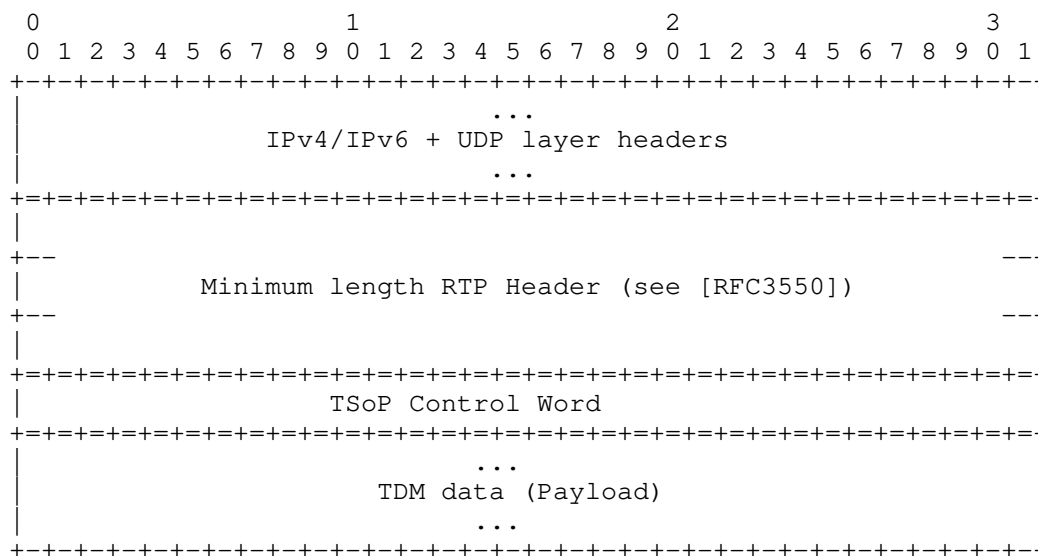


Figure 6. TSoP Packet Format for an IPv4/IPv6 PSN with UDP PW Demultiplexing

4.3.2. Usage and Structure of the TSoP Control Word

The purpose of the TSoP control word is to allow:

1. Detection of packet loss or misordering
2. Differentiation between PSN and attachment circuit problems as a cause for outage of the emulated service
3. Signaling of faults detected at the PW egress to the PW ingress

The structure of the TSoP Control Word is in accordance with the general PW Control Word format specified in [RFC4385]. The TSoP CW format is shown in Figure 7 below. This TSoP Control Word MUST be present in each TSoP PW packet.

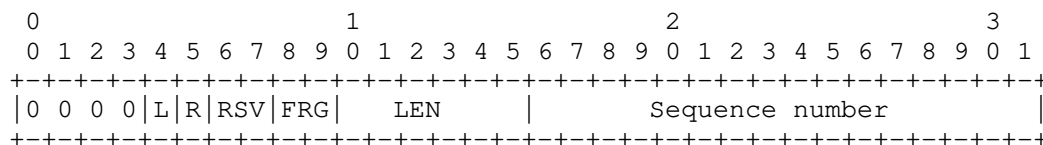


Figure 7. Structure of the TSoP Control Word

The use of Bits 0 to 3 is described in [RFC4385]. These bits MUST be set to zero unless they are being used to indicate the start of an Associated Channel Header (ACH). An ACH is needed if the state of the TSoP PW is being monitored using Virtual Circuit Connectivity Verification [RFC5085] or in case OAM functionality according to [RFC6371] is added.

L (bit 4) - If this bit is set, it indicates that the STM-N data ingressing in the PSN-bound IWF is currently experiencing a fault condition. Once set, if the fault is rectified, the L-bit MUST be cleared. For each frame that is transmitted with L-bit = 1, the PSN-bound IWF MUST insert such an amount of substitution data in the TSoP payload field that the TSoP frame length, as it is during normal operation, is maintained. The CE-bound IWF MUST play out an amount of G-AIS data corresponding to the original TSoP Payload Field for each received packet with the L-bit set.

Note: This document does not prescribe exactly which STM-N fault conditions are to be treated as invalidating the payload carried in the TSoP packets. An example of such a fault condition would be LOS.

R (bit 5) - If this bit is set by the PSN-bound IWF, it indicates that its local CE-bound IWF is in the LOPS state, i.e., it has lost a preconfigured number of consecutive packets. The R-bit MUST be cleared by the PSN-bound IWF once its local CE-bound IWF

has exited the LOPS state, i.e., has received a preconfigured number of consecutive packets. See also section 6.2.2.

RSV (bits 6 to 7) - This field MUST be set to 0 by the PSN-bound IWF and MUST be ignored by the CE-bound IWF. RSV is reserved.

FRG (bits 8 to 9) - This field MUST be set to 0 by the PSN-bound IWF and MUST be ignored by the CE-bound IWF. FRG is fragmentation; see [RFC4623].

LEN (bits 10 to 15) - This field MAY be used to carry the length of the TSoP packet (defined as the length of the TSoP Encapsulation Header + TSoP Payload Field) if it is less than 64 octets, and MUST be set to zero otherwise. When the LEN field is set to 0, the preconfigured size of the TSoP packet payload MUST be assumed to be as described in Section 5, and if the actual packet size is inconsistent with this length, the packet MUST be considered malformed.

Sequence number (bits 16 to 31) - This field is used to enable the common PW sequencing function as well as detection of lost packets. It MUST be generated in accordance with the rules defined in Section 5.1 of [RFC3550] for the RTP sequence number:

- o Its space is a 16-bit unsigned circular space
- o Its initial value SHOULD be random (unpredictable).

It MUST be incremented with each TSoP data packet sent in the specific PW.

4.3.3. Usage of the RTP Header

A minimum length RTP Header as specified in [RFC3550] MUST be included in the TSoP Encapsulation Header. The reason for mandating the insertion of an RTP Header by the PSN-bound IWF is that it is expected that in most cases the CE-bound IWF will need to use the contained timestamps to be able to recover a clock signal of sufficient quality. By avoiding to make the presence of RTP Headers subject to configuration, the design of the of the CE-bound IWF can be simplified and another potential source of errors during commissioning is eliminated.

The RTP Header fields in the list below (see also figure 8) MUST have the following specific values:

V (version) = 2
P (padding) = 0
X (header extension) = 0

CC (CSRC count) = 0
M (marker) = 0

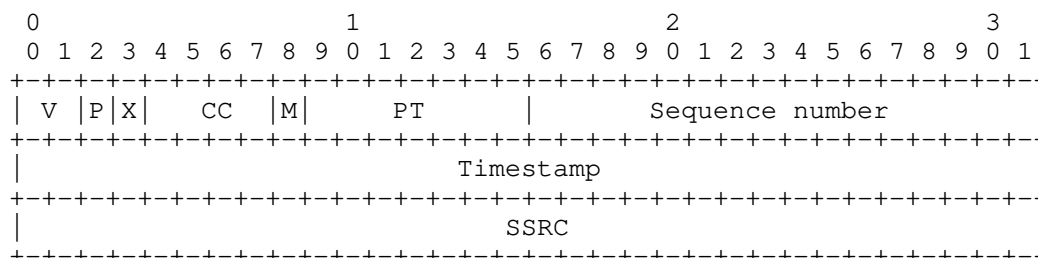


Figure 8. Structure of the RTP Header field

The PT (payload type) field is used as follows:

1. One PT value MUST be allocated from the range of dynamic values (see [RTP-TYPES]) for each direction of the PW. The same PT value MAY be reused for both directions of the PW and also reused between different PWs.
2. The PSN-bound IWF MUST set the PT field in the RTP header to the allocated value.
3. The CE-bound IWF MAY use the received value to detect malformed packets.

The sequence number MUST be the same as the sequence number in the TSoP control word.

The RTP timestamps are used for carrying timing information over the network. Their values MUST be generated in accordance with the rules established in [RFC3550].

A TSoP implementation MUST support RTP timestamping at the PW ingress with a nominal clock frequency of 25 MHz. This is also the default value. Other clock frequencies MAY be supported to generate the RTP Timestamps. Selection of the applicable clock frequency is done during commissioning of the PW that carries the emulated STM-N service.

The SSRC (synchronization source) value in the RTP header MAY be used for detection of misconnections, i.e., incorrect interconnection of attachment circuits. In case this option is not used, this field should contain an all zero pattern.

The usage of the options associated with the RTP Header (the timestamping clock frequency, selected PT and SSRC values) MUST be aligned between the two TSoP IWFs during Pseudowire commissioning.

5. TSoP Payload Field

In order to facilitate handling of packet loss in the PSN, all packets belonging to a given TSoP PW are REQUIRED to carry a fixed number of octets in its TSoP Payload Field.

The TSoP Payload Field length MUST be defined during PW commissioning, MUST be the same for both directions of the PW, and MUST remain unchanged for the lifetime of the PW.

All TSoP implementations MUST be capable of supporting the following TSoP Payload Field length:

- o STM-N (for $N = 0, 1, 4, 16$ and 64) - 810 octets

Notes:

1. Whatever the selected payload size, TSoP does not assume alignment to any underlying structure imposed by TDM framing (octet, frame, or multiframe alignment). The STM-N signal remains scrambled through the TSoP encapsulation and decapsulation processes.
2. With a payload size of 810 octets, the STM-N emulated service over the PSN will have a nominal packet rate of 8000 packets/s when $N = 0$ and a nominal packet rate of $24000 * N$ packets/s for $N \geq 1$.

TSoP uses the following ordering for packetization of the TDM data:

- o The order of the payload octets corresponds to their order on the attachment circuit.
- o Consecutive bits coming from the attachment circuit fill each payload octet starting from most significant bit to least significant.

6. TSoP Operation

6.1. Common Considerations

Edge-to-edge emulation of an STM-N service using TSoP is only possible when the two PW attachment circuits are of the same type, i.e., both are STM-N with equal N .

6.2. IWF Operation

6.2.1. PSN-Bound Direction

Once the PW is commissioned, the PSN-bound TSoP IWF operates as follows:

The ingressing STM-N bit-stream is segmented, such that each segment contains the configured number of payload octets per packet. This forms the TSoP Payload Field. The STM-N bit-stream **MUST NOT** be descrambled before segmentation and packetization for PW transport.

Subsequently, the TSoP Encapsulation Headers are prepended according to the rules in section 4.3.

Lastly, the PSN/PW Headers are added to the packetized service data, and, depending on the applicable PSN technology, a Frame Check Sum is added. The resulting packets are transmitted over the PSN.

6.2.2. CE-Bound Direction

Once the PW is commissioned, the CE-bound TSoP IWF operates as follows:

Each time a valid TSoP packet is received from the PSN, its sequence number is checked to determine its relative position in the stream of received packets. Packets that are received out-of-order **MAY** be reordered. Next, the data in the fixed length TSoP payload field of each packet is written into a (jitter) buffer in the order indicated by its sequence number. In case data is missing due to a lost packet or a packet that could not be re-ordered, an equivalent amount of dummy data (G-AIS pattern) is substituted.

Subsequently, the STM-N stream towards the CE is reconstructed by playing out the buffer content with a clock that is reconstructed to have the same average frequency as the STM-N clock at the PW ingress. In addition, this clock signal must have such properties that the following requirements can be met:

- o A reconstructed SDH-type STM-N signal delivered to an Attachment Circuit **MUST** meet [G.825] jitter and wander requirements, or,
- o A reconstructed SONET-type OC-M signal delivered to an Attachment Circuit **MUST** meet [GR-253] jitter and wander requirements.

The size of the buffer in the CE-bound TSoP IWF **SHOULD** be configurable to allow accommodation to the PSN specific packet delay variation.

The CE-bound TSoP IWF **SHOULD** use the sequence number in the TSoP

Control Word for detection of lost and misordered packets. The sequence numbers in the RTP Header MAY be used instead.

Note: A valid sequence number can be always found in bits 16 - 31 of the first 32-bit word immediately following the PW demultiplexing header regardless of the specific PSN type, multiplexing method, location of the RTP header, etc. This approach simplifies implementations supporting multiple encapsulation types as well as implementation of multi-segment (MS) PWs using different encapsulation types in different segments.

The CE-bound TSoP IWF MAY reorder misordered packets. Misordered packets that can not be reordered MUST be discarded and treated the same way as lost packets.

The payload of received TSoP packets marked with the L-bit set MUST be replaced by the equivalent number of bits from the G-AIS pattern. Likewise, the payload of each lost or malformed (see section 6.3) TSoP packet MUST be replaced with the equivalent number of bits from the G-AIS pattern.

Before a TSoP PW has been commissioned and after a PW has been decommissioned, the IWF MUST play out the G-AIS pattern to its STM-N attachment circuit.

Once a TSoP PW has been commissioned, the CE-bound IWF begins to receive TSoP packets and to store their payload in the buffer, but continues to play out the G-AIS pattern to its TDM attachment circuit. This intermediate state persists until a preconfigured degree of filling (for example half of the CE-bound IWF buffer) has been reached by writing consecutive TSoP packets or until a preconfigured intermediate state timer (started when the TSoP commissioning is complete) expires.

Each time an STM-N signal is replaced by a G-AIS signal at the same nominal bitrate, this signal may start at an arbitrary point in its repeating 2047-bit sequence. Once the starting point is selected, the G-AIS signal is sent uninterrupted until the condition that invoked it has been removed. The frequency of the clock that is used to generate this G-AIS signal MUST have an accuracy that is better than +/- 20 ppm relative to the nominal STM-N frequency.

Once the preconfigured amount of the STM-N data has been received, the CE-bound TSoP IWF enters its normal operational state where it continues to receive TSoP packets and to store their payload in the buffer while playing out the contents of the jitter buffer in accordance with the required clock. In this state, the CE-bound IWF performs clock recovery, MAY monitor PW defects, and MAY collect PW

performance monitoring data.

The CE-bound IWF enters the LOPS defect state in case it detects the loss of a preconfigured number of consecutive packets or if the intermediate state timer expires before the required amount of TDM data has been received. While in this state, the local PSN-bound TSoP IWF SHOULD mark every packet it transmits with the R-bit set. The CE-bound IWF leaves the LOPS defect state and transits to the normal state once a preconfigured number of consecutive valid TSoP packets have been received (successfully reordered packets contribute to the count of consecutive packets).

The RTP timestamps inserted in each TSoP packet at the PW ingress allow operation in differential mode provided that both PW ingress and PW egress IWFs have a local clock that is traceable to a common timing source.

The use of adaptive mode clocking mode, i.e., recovering the STM-N clock in the CE-bound IWF by essentially averaging the arrival times of the TSoP packets from the PSN without using RTP information, is not recommended for TSoP-based circuit emulation.

6.3. TSoP Defects

In addition to the LOPS state defined above, the CE-bound TSoP IWF MAY detect the following defects:

- o Stray packets
- o Malformed packets
- o Excessive packet loss rate
- o Buffer overrun
- o Buffer underrun
- o Remote packet loss

Corresponding to each defect is a defect state of the IWF, a detection criterion that triggers transition from the normal operation state to the appropriate defect state, and an alarm that MAY be reported to the management system and thereafter cleared. Alarms are only reported when the defect state persists for a preconfigured amount of time (typically 2.5 seconds) and MUST be cleared after the corresponding defect is undetected for a second preconfigured amount of time (typically 10 seconds). The trigger and release times for the various alarms may be independent.

Stray packets MAY be detected by the PSN and PW demultiplexing layers. The SSRC field in the RTP header MAY be used for this purpose as well. Stray packets MUST be discarded by the CE-bound IWF, and their detection MUST NOT affect mechanisms for detection of

packet loss.

Malformed packets are detected by mismatch between the expected packet size and the actual packet size inferred from the PSN and PW demultiplexing layers (taking the value of the L-bit into account). Differences between the received PT value and the PT value allocated for this direction of the PW MAY also be used for this purpose. Malformed, in-order packets MUST be discarded by the CE-bound IWF and replacement data generated as with lost packets.

Excessive packet loss rate is detected by computing the average packet loss rate over a configurable amount of time and comparing it with preconfigured raise and clear thresholds.

Buffer overrun is detected in normal operational state when the (jitter) buffer of the CE-bound IWF cannot accommodate newly arrived TSoP packets.

Buffer underrun can be detected in normal operational state when the (jitter) buffer of the CE-bound IWF has insufficient data to maintain playing out the STM-N signal towards the CE at the recovered clock rate. In this situation G-AIS MUST be substituted until the buffer fill has reached its preconfigured degree of filling again.

Remote packet loss is indicated by reception of packets with their R-bit set.

6.4. TSoP Performance Monitoring

Performance monitoring (PM) parameters are routinely collected for STM-N services and provide an important maintenance mechanism in SDH networks. However, STM-N level PM data provides the information over the performance of the end-to-end STM-N connection, which may extend well beyond the part in which it is carried over a TSoP Pseudowire.

It may be important to be able to measure the performance of a TSoP Pseudowire section, which forms a part of the STM-N end-to-end connection, in isolation. For that reason a set of packet level counters are specified that can be used to assess the performance of the TSoP Pseudowire section. Collection of the TSoP PW performance monitoring data is OPTIONAL and, if implemented, is only performed after the CE-bound IWF has exited its intermediate state.

The following counters are defined:

ENCAP_TXTOTAL_PKTS - The total number of TSoP packets that is transmitted towards the PSN by the PSN-bound IWF function. This includes packets with the L-bit set.

DECAP_RXTOTAL_PKTS - The total number of TSoP packets that is received from the PSN by the CE-bound IWF function. This includes malformed packets, out-of-order packets and packets with the L-bit set.

DECAP_REORDERED_PKTS - The number of out-of-order TSoP packets that is received from the PSN by the CE-bound IWF, based on the received sequence numbers, for which the ordering could be corrected by the CE-bound IWF.

DECAP_MISSING_PKTS - The number of TSoP packets that did not arrive at the CE-bound IWF from the PSN, based on the received sequence numbers.

DECAP_MALFORMED_PKTS - The number of TSoP packets that is received from the PSN by the CE-bound IWF function which contains one of the following RTP related errors: TSoP Payload Field length mismatch, PT-value mismatch (if checked) and/or SSRC mismatch (if checked).

DECAP_OUTOFORDER_PKTS - The number of out-of-order TSoP packets that is received from the PSN by the CE-bound IWF, based on the received sequence numbers, for which the ordering could not be corrected by the CE-bound IWF.

DECAP_OVERRUN_PKTS - The number of packets that is received from the PSN that is dropped by the CE-bound IWF due to the fact that the (jitter) buffer has insufficient capacity to store the TSoP Payload Field content.

DECAP_UNDERRUN_BITS - The number of bits that is not played out towards the CE by the CE-bound IWF because the (jitter) buffer is empty at the moment they need to be played out.

DECAP_PLAYEDOUT_PKTS - The number of packets that has been successfully played out towards the CE by the CE-bound IWF containing valid STM-N payload including the packets that have been received with the L-bit containing substituted data. Packets which are lost in transmission over the PSN or packets which discarded by the CE-bound IWF due to some error condition are not counted.

Note that packets with the L-bit set are considered normal data from the perspective of TSoP Pseudowire Performance Monitoring, since in such cases the location of the fault is before the signal ingresses the PSN-bound IWF, so outside the scope of the TSoP PW.

7. Quality of Service (QoS) Issues

TSoP SHOULD employ existing QoS capabilities of the underlying PSN.

If the PSN providing connectivity between PE devices is Diffserv-enabled and provides a PDB [RFC3086] that guarantees low jitter and low loss, the TSoP PW SHOULD use this PDB in compliance with the admission and allocation rules the PSN has put in place for that PDB (e.g., marking packets as directed by the PSN).

If the PSN is Intserv-enabled, then GS (Guaranteed Service) [RFC2212] with the appropriate bandwidth reservation SHOULD be used in order to provide a bandwidth guarantee equal or greater than that of the aggregate TDM traffic.

8. Congestion Control

As explained in [RFC3985], the PSN carrying the PW may be subject to congestion. TSoP PWs represent inelastic constant bit-rate (CBR) flows and cannot respond to congestion in a TCP-friendly manner prescribed by [RFC2914], although the percentage of total bandwidth they consume remains constant.

Unless appropriate precautions are taken, undiminished demand of bandwidth by TSoP PWs can contribute to network congestion that may impact network control protocols.

Whenever possible, TSoP PWs SHOULD be carried across traffic-engineered PSNs that provide either bandwidth reservation and admission control or forwarding prioritization and boundary traffic conditioning mechanisms. IntServ-enabled domains supporting Guaranteed Service (GS) [RFC2212] and DiffServ-enabled domains [RFC2475] supporting Expedited Forwarding (EF) [RFC3246] provide examples of such PSNs. Such mechanisms will negate, to some degree, the effect of the TSoP PWs on the neighboring streams. In order to facilitate boundary traffic conditioning of TSoP traffic over IP PSNs, the TSoP IP packets SHOULD NOT use the DiffServ Code Point (DSCP) value reserved for the Default Per-Hop Behavior (PHB) [RFC2474].

If TSoP PWs run over a PSN providing best-effort service, they SHOULD monitor packet loss in order to detect "severe congestion". If such a condition is detected, a TSoP PW SHOULD shut down bi-directionally for some period of time as described in Section 6.5 of [RFC3985].

Note that:

1. The TSoP IWF can inherently provide packet loss measurement since

the expected rate of arrival of TSoP packets is fixed and known

2. The results of the TSoP packet loss measurement may not be a reliable indication of presence or absence of severe congestion if the PSN provides enhanced delivery. For example:
 - a) If TSoP traffic takes precedence over non-TSoP traffic, severe congestion can develop without significant TSoP packet loss.
 - b) If non-TSoP traffic takes precedence over TSoP traffic, TSoP may experience substantial packet loss due to a short-term burst of high-priority traffic.
3. The STM-N services emulated by the TSoP PWs have high availability objectives (see [G.829]) that MUST be taken into account when deciding on temporary shutdown of TSoP PWs.

This specification does not define the exact criteria for detecting "severe congestion" using the TSoP packet loss rate or the specific methods for bi-directional shutdown the TSoP PWs (when such severe congestion has been detected) and their subsequent re-start after a suitable delay. This is left for further study. However, the following considerations may be used as guidelines for implementing the TSoP severe congestion shutdown mechanism:

1. If the TSoP PW has been set up using either PWE3 control protocol [RFC4447] or L2TPv3 [RFC3931], the regular PW teardown procedures of these protocols SHOULD be used.
2. If one of the TSoP PW end points stops transmission of packets for a sufficiently long period, its peer (observing 100% packet loss) will necessarily detect "severe congestion" and also stop transmission, thus achieving bi-directional PW shutdown.

9. Security Considerations

TSoP does not enhance or detract from the security performance of the underlying PSN; rather, it relies upon the PSN mechanisms for encryption, integrity, and authentication whenever required.

TSoP PWs share susceptibility to a number of Pseudowire layer attacks and will use whatever mechanisms for confidentiality, integrity, and authentication are developed for general PWs. These methods are beyond the scope of this document.

Although TSoP PWs MUST employ an RTP header to achieve an explicit transfer of timing information, SRTP (see [RFC3711]) mechanisms are NOT RECOMMENDED as a substitute for PW layer security.

Misconnection detection capabilities of TSoP increase its resilience to misconfiguration.

Random initialization of sequence numbers, in both the control word and the optional RTP header, makes known-plaintext attacks on encrypted TSoP PWs more difficult. Encryption of PWs is beyond the scope of this document.

10. Applicability Statements

TSoP is an encapsulation layer intended for carrying SDH STM-N circuits over the PSN in a structure-agnostic and fully transparent fashion.

TSoP fully complies with the principle of minimal intervention, minimizing overhead and computational power required for encapsulation.

TSoP provides sequencing and synchronization functions needed for emulation of STM-N bit-streams, including detection of lost or misordered packets and perform the appropriate compensation. Furthermore, explicit timing information is provided by the presence of an RTP timestamp in each TSoP packet.

STM-N bit-streams carried over TSoP PWs may experience delays exceeding those typical of native SDH networks. These delays include the TSoP packetization delay, edge-to-edge delay of the underlying PSN, and the delay added by the jitter buffer. It is recommended to estimate both delay and delay variation prior to setup of a TSoP PW.

TSoP carries STM-N streams over PSN in their entirety, including any control plane data contained within the data. Consequently, the emulated STM-N services are sensitive to the PSN packet loss. Appropriate generation of replacement data can be used to prevent shutting down the CE STM-N interface due to occasional packet loss. Other effects of packet loss on this interface (e.g., errored blocks) cannot be prevented.

TSoP provides for effective fault isolation by forwarding the local attachment circuit failure indications to the remote attachment circuit.

TSoP provides for a carrier-independent ability to detect misconnections and malformed packets via the PT and SSRC fields in the RTP Header. This feature increases resilience of the emulated service to misconfiguration.

Being a constant bit rate (CBR) service, TSoP cannot provide TCP

friendly behavior under network congestion.

Faithfulness of a TSoP PW may be increased by exploiting QoS features of the underlying PSN.

TSoP does not provide any mechanisms for protection against PSN outages, and hence its resilience to such outages is limited. However, lost-packet replacement and packet reordering mechanisms increase resilience of the emulated service to fast PSN rerouting events.

11. IANA Considerations

IANA is requested to assign a new MPLS Pseudowire (PW) type for the following TSoP encapsulated services:

PW type	Description	Reference
-----	-----	-----
0x0020	STM-0 or OC-1	RFC XXXX
0x0021	STM-1 or OC-3	RFC XXXX
0x0022	STM-4 or OC-12	RFC XXXX
0x0023	STM-16 or OC-48	RFC XXXX
0x0024	STM-64 or OC-192	RFC XXXX

The above value is suggested as the next available value and has been reserved for this purpose by IANA.

RFC Editor: Please replace RFC XXXX above with the RFC number of this document and remove this note.

12. Acknowledgements

The authors of this document are much indebted to the authors of [RFC4553]. This latter RFC has been used as a template and example for the current document. Moreover, many paragraphs and sentences have been copied from this RFC without alteration or with only slight modification into the current document.

Furthermore, we thank Zhu Bao, Jeff Towne, Willem van den Bosch and Matthew Bocci for their valuable feedback.

13. References

13.1. Normative References

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

- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.
- [G.707] ITU-T Recommendation G.707/Y.1322 (01/2007) - Network node interface for the synchronous digital hierarchy (SDH)
- [G.783] ITU-T Recommendation G.783 (03/2006) - Characteristics of synchronous digital hierarchy (SDH) equipment functional blocks
- [O.150] ITU-T Recommendation O.150 (05/1996) - General requirements for instrumentation for performance measurement on digital transmission equipment
- [G.825] ITU-T Recommendation G.825 (03/2000) - The control of jitter and wander within digital networks which are based on the synchronous digital hierarchy (SDH)
- [GR-253] Telcordia GR-253-CORE - Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria (September 2000)

13.2. Informative References

- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3916] Xiao, X., Ed., McPherson, D., Ed., and P. Pate, Ed., "Requirements for Pseudo-Wire Emulation Edge-to-Edge (PWE3)", RFC 3916, September 2004.
- [RFC3931] Lau, J., Ed., Townsley, M., Ed., and I. Goyret, Ed., "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.
- [RFC3985] Bryant, S., Ed., and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4447] Martini, L., Ed., Rosen, E., El-Aawar, N., Smith, T., and

- G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4553] Vainshtein, A., Ed., and YJ. Stein, Ed., "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", RFC 4553, June 2006.
- [RFC4623] Malis, A. and M. Townsley, "Pseudowire Emulation Edge-to-Edge (PWE3) Fragmentation and Reassembly", RFC 4623, August 2006.
- [RFC4842] Malis, A., Pate, P., Cohen, R., Ed., and D. Zelig, "Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH) Circuit Emulation over Packet (CEP)", RFC 4842, April 2007.
- [RFC5085] Nadeau, T., Ed., and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5254] Bitar, N., Ed., Bocci, M., Ed., and L. Martini, Ed., "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3)", RFC 5254, October 2008.
- [RFC5604] Nicklass, O., "Managed Objects for Time Division Multiplexing (TDM) over Packet Switched Networks (PSNs)", RFC 5604, July 2009.
- [RFC6371] Busi, I., Ed., and D. Allan, Ed., "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, September 2011.
- [G.709] ITU-T Recommendation G.709/Y.1331 (12/2009) - Interfaces for the Optical Transport Network (OTN)
- [G.829] ITU-T Recommendation G.829 (12/2002) - Error performance events for SDH multiplex and regenerator sections
- [802.1Q] IEEE Std. 802.1Q-2011, "Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks", 31 August 2011
- [MEF 8] Metro Ethernet Forum - Implementation Agreement for the Emulation of PDH Circuits over Metro Ethernet Networks (October 2004)
- [RTP-TYPE] RTP PARAMETERS, <<http://www.iana.org/assignments/rtp-parameters>>

Appendix A. Parameters to be configured to set up a TSoP PW

The following parameters of the TSoP IWF MUST be agreed upon between the peer IWFs during the PW setup. Such an agreement can be reached via manual configuration or via one of the PW set-up protocols:

1. Type of attachment circuit, i.e., the value of N of the STM-N signal, which corresponds to a bit rate as mentioned in section 3.
2. Payload size, i.e., the (constant) number of octets that is transmitted in the TSoP Payload Field of each TSoP packet. The default value is 810 octets.
3. Timestamping clock frequency: 25 MHz (default) or an alternative value.
4. The configurability of the following parameters (see section 6.2.2) governing the behavior of the CE-bound IWF buffer is optional:
 - a) The maximum amount of payload data that may be stored in the CE-bound IWF payload buffer
 - b) The desired degree of filling of the CE-bound IWF buffer in steady state (typically 50% of the configured buffer size)
 - c) The "intermediate state" timer, i.e., the maximum amount of time that the CE-bound IWF waits before after the first TSoP packet has been received, before it starts to play out data from the buffer towards the CE
5. The content of the following RTP header fields must be provided by the user:
 - a) The 7-bit RTP Payload Type (PT) value; any value can be assigned to be used with TSoP PWs. Default is an all zero pattern.
 - b) The 32-bit Synchronization Source (SSRC) value. Default is an all zero pattern.
6. The order of the RTP Header and TSoP-CW Header must be defined. This may be derived from the applied PSN transport technology, see section 4.3
7. The number of TSoP packets that must be missed consecutively before the CE-bound IWF enters the LOPS defect state (default 10) and the number of TSoP packets that must be received consecutively

to clear the LOPS defect state (default 2). See section 4.3.2 and [RFC5604]

8. To support the optional excessive packet loss event by the CE-bound IWF, the following parameters must be configured:
 - a) The length of the observation period for detecting excessive packet loss. Default value is 10 s.
 - b) The minimum number of lost packets that is to be detected in the observation interval before an excessive packet loss alarm is raised. Default value is 30% of the expected packets.
 - c) The maximum number of lost packets that is to be detected in the observation interval to clear an excessive packet loss alarm. Default value is 1% of the expected packets.

Authors' Addresses

Gert Manhoudt
AimValley B.V
Utrechtseweg 38
1213 TV Hilversum
The Netherlands
E-mail: gmanhoudt@aimvalley.nl

Stephan Roullot
Alcatel-Lucent Centre de Villarceaux
Route de Villejust
91620 Nozay
France
E-mail: stephan.roullot@alcatel-lucent.com

Peter Roberts
Alcatel-Lucent
600 March Road
Kanata, Ontario, K2K 2E6
Canada
E-mail: peter.roberts@alcatel-lucent.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 11, 2013

J. Medved
A. McLachlan
D. Meyer
Cisco Systems
July 10, 2012

MPLS-TP Pseudowire Configuration using OpenFlow 1.3
draft-medved-pwe3-of-config-01

Abstract

This document describes a method by which MPLS-TP Pseudowires (PW) can be configured in an LER using OpenFlow 1.3. In addition to the configuration of PWs this document also specifies how to enact OAM for these PWs using standard IETF conventions defined by the GAL label method. The primary goal of this document is to provide a simple and yet flexible method for configuring PWs using standardized tools from the emerging SDN toolkit.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 11, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
1.2. Terminology	3
1.3. Overview	3
2. The Reference Topology	5
2.1. The MPLS-TP Node	7
2.1.1. The Virtual OF Switch	9
2.1.2. The OAM Engine	10
3. PW Configuration	11
3.1. Configuration Messages	11
3.1.1. The Flow Modification Message	11
3.1.2. The Group Modification Message	13
3.2. PW Head-End Node Configuration	13
3.2.1. 'Modify Group Entry' Message Details	14
3.2.2. 'Modify Flow Entry' Message Details	15
3.3. PW Tail-End Node Configuration	15
3.3.1. 'Modify Flow Entry' Message Details	16
4. PW OAM Considerations	17
4.1. OAM Overview	17
4.2. PW OAM Engine Configuration	17
4.3. OAM and S-Bit considerations	18
5. IANA Considerations	18
6. Security Considerations	18
7. Acknowledgements	19
8. Normative References	19
Authors' Addresses	19

1. Introduction

1.1. Requirements Language

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

1.2. Terminology

This document uses the following terminology:

Term	Definition
AC:	Attachment Circuit
ACH:	Associated Channel Header
BFD:	Bidirectional Forwarding Detection
CE:	Customer Edge
G-ACh:	Generic Associated Channel
GAL:	G-ACh Label
iPPRoc:	input Packet Processing function
LER:	Label Edge Router
LSP:	Label Switch Path
LSR:	Label Switch Router
MPLS:	Multiprotocol Label Switching
MPLS-TP:	MPLS Transport Profile
MPLS-TP P:	MPLS-TP Provider LSR
MPLS-TP PE:	MPLS-TP Provider Edge LSR
PDU:	Protocol Data Unit
PG:	Port Group
PSN:	Packet Switching Network
PW:	Pseudowire
OAM:	Operations, Administration, and Maintenance
OAM Engine:	Operations, Administration, and Maintenance Engine
OF:	OpenFlow
oPPRoc:	output Packet Processing function
SDN:	Software Defined Networks
VP:	Virtual Port

1.3. Overview

MPLS-TP provides a relatively light weight layer 2 transport technology by leveraging elements of existing transport platforms and a subset of the more recent MPLS protocol standards. PWs are configured as bi-directional paths over the MPLS-TP network, usually by an external management platform. At present no open standards

exist to provision these PWs, and therefore there is a reliance on vendor specific provisioning platforms. It should be noted that there exists alternative methods for the static provisioning of PWs, including via SNMP ([RFC5601]).

This document describes a mechanism that uses the emerging OpenFlow standard ([OF-1.3.0]) to provision PWs and PW OAM at a Label Edge Router (LER) in a MPLS-TP environment. The method described here uses standard MPLS-TP control planes. In particular, this document does not specify new control planes for either MPLS-TP or for PW setup (e.g., T-LDP as specified in [RFC6373]). Naturally the implementation of OpenFlow will be required on the TP switch, as would an OAM Engine, the functions of which are described in this document. In addition, an OpenFlow Controller will be required for the provisioning functions.

Because OpenFlow is an open standard, it enables Service Providers to adopt a more consolidated approach to provisioning. An OpenFlow Controller can be common to a number of different elements in the network, as being driven by current industry Software Defined Networks (SDN) developments.

This document uses the reference MPLS-TP architecture defined in [RFC5921], which is shown in the following figure:

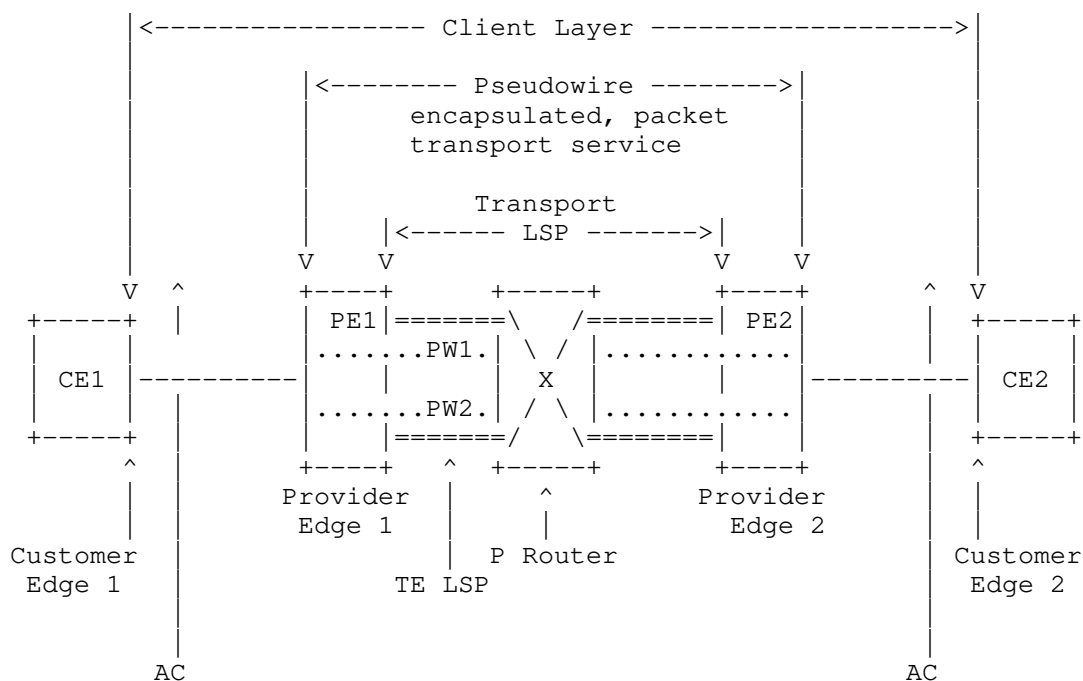


Figure 1: MPLS-TP Architecture (Single Segment PW) from RFC5921

A Pseudowire (PW) is configured between an ingress attachment circuit on a head-end switch (Provider Edge 1, PE1) and an egress attachment circuit on a tail-end node (Provider Edge 2, PE2). For a complete service, a PW must be configured in each direction.

2. The Reference Topology

Relevant components from the above architecture diagram for LERs are shown in the reference topology in Figure 2. For clarity, only one of the two PWs that constitute a complete service is shown in the reference topology and discussed in subsequent sections: the PW from Provider Edge 1 (Head-End Node) to Provider Edge 2 (Tail-End Node). A PW in the opposite direction (from PE2 to PE1) would be configured similarly.

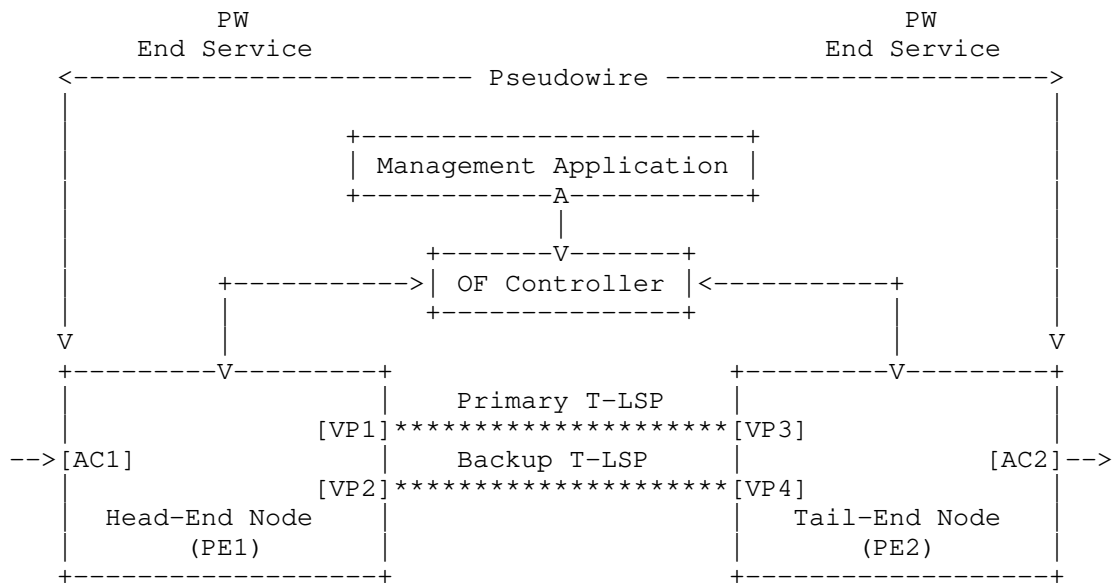


Figure 2: Reference topology

Pseudowires are configured from a network / provisioning Management Application which communicates with MPLS-TP nodes through an OpenFlow Controller (OF Controller). The Management Application configures an end-to-end Pseudowire ([PW1]) between Attachment Circuit [AC1] on the headend node and Attachment Circuit [AC2] on the tail-end node. At the head-end, all traffic coming from an input port (Attachment Circuit [AC1]) is switched onto the PW, and at the tail-end all traffic coming from the Pseudowire is switched to an output port (Attachment Circuit [AC2]). The Pseudowire is assigned a PW Label [PWL1]. The Management Application configures both packet forwarding and OAM function related to Pseudowires.

In the reference topology, the head-end and tail-end nodes are connected via a pair of transport LSPs - a primary Transport LSP (T-LSP) and a backup Transport LSP. Configuration and setup of transport the LSPs is outside the scope of this document. Tunnels are represented as Virtual Ports; in OpenFlow parlance, a function that performs functionality outside the OpenFlow specification is called a "virtual port" (see Section B.9.4 of [OF-1.3.0]). Here [VP1] and [VP2] are virtual ports on the head-end node, and [VP3] and [VP4] are virtual ports on the tail-end node.

2.1. The MPLS-TP Node

A reference MPLS-TP Node has three major programmable components: a Virtual OF Switch, an OAM Engine, and Packet Processing functions attached to input and output ports of the Virtual OF Switch. The MPLS-TP node shown in the following figure.

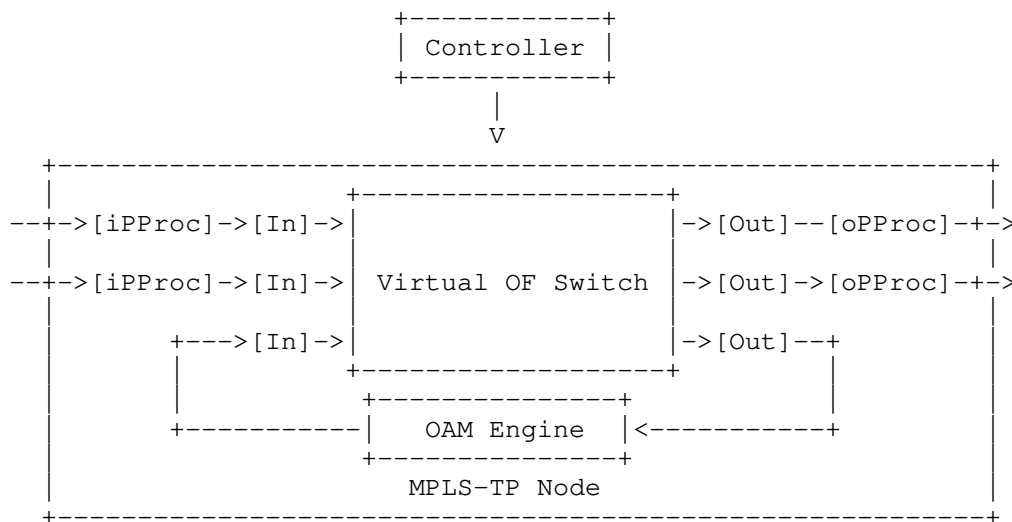


Figure 3: MPLS-TP Node Components

The Virtual OF Switch performs packet switching. It is controlled by an external controller, which in turn is driven by an MPLS-TP Management Application.

Packets arrive at the Virtual OF Switch through a virtual port consisting of an Attachment Circuits followed by an input Packet Processing function (iPProc). The the iPProc encapsulates arriving packets on the Attachment Circuit in the outer transport (Ethernet) header. This is required because OpenFlow can only push MPLS labels onto the top of a label stack encapsulated in an existing Ethernet header. Furthermore, since OpenFlow cannot push an Ethernet header (see Section 5.12 of [OF-1.3.0]), the encapsulation must be done in a virtual port in order to construct a valid packet. Note also that while pushing and popping MPLS labels is optional functionality in the OpenFlow specification, it is required to support the functionality described in this document.

When a packet arrives on an Attachment Circuit, say [AC1], the iPProc function receives the packet and encapsulates it in what will become the outer transport header. The packet is then handed to the Virtual

OF Switch which can now push the PW label onto the packet. Note that for purposes of this revision of this document it is assumed that the controller (and hence the Virtual OF Switch) receives PW label out-of-band. The packet is then output to a virtual port in which the oPProc pushes the appropriate Transport label and switches the packet onto the Transport LSP. This sequence is depicted in Figure 4. Finally, while it is in principle possible for the OF Switch to also push the Transport label, the design decision taken here is push the Transport label in a virtual port in order to keep the OF Switch as simple as possible; in this case by limiting it to one table.

A packet initially arrives at the AC with the following headers and payload. This packet becomes the PW payload.

```
<Payload>          \  
<Payload-SA>       - PW payload  
<Payload-DA>       /
```

Next, the iPProc adds the Transport header.

```
<Payload>          \  
<Payload-SA>       - PW payload  
<Payload-DA>       /  
<T-SA, T-DA>      - Transport header
```

The Virtual OF switch then pushes on PW label with S=1.

```
<Payload>          \  
<Payload-SA>       - PW payload  
<Payload-DA>       /  
<PW Label, S=1>   - PW label  
<T-SA, T-DA>      - Transport header
```

Finally, the oPProc pushes the Transport label and switches the packet onto the Transport LSP.

```
<Payload>          \  
<Payload-SA>       - PW payload  
<Payload-DA>       /  
<PW Label, S=1>   - PW label  
<T Label, S=0>    - Transport label  
<T-SA, T-DA>      - Transport header
```

Figure 4: Input Encapsulation Sequence

When receiving a packet on a Transport LSP, the iPProc pops the

Transport label from packets exiting the Transport LSP (again, the choice to pop the Transport label in the iPProc is to limit the OF switch one table). The PW Label is subsequently popped by the Virtual OF switch. Finally, the oPProc receives the packet sent to the virtual port, strips the transport header from the outgoing packet and delivers it to the Attachment Circuit.

The OAM Engine generates and receives&processes OAM packets. It can perform OAM functions for both Pseudowires and Transport LSPs. At the head-end node, the OAM Engine is connected to a Virtual OF Switch input port. OAM packets are switched through the Virtual OF Switch either onto either Pseudowires or onto the transport LSPs. At the tail-end node, the OAM Engine is connected to a Virtual OF Switch output port. OAM packets are switched either from transport LSPs or Pseudowires to the OAM Engine. The tail-end node OAM Engine detects failure conditions. The head-end OAM Engine performs corrective actions.

2.1.1. The Virtual OF Switch

The Virtual OF switch is comprised of a single flow table (Flow Table 1) and a single group table. The Virtual OF Switch is shown in the following figure.

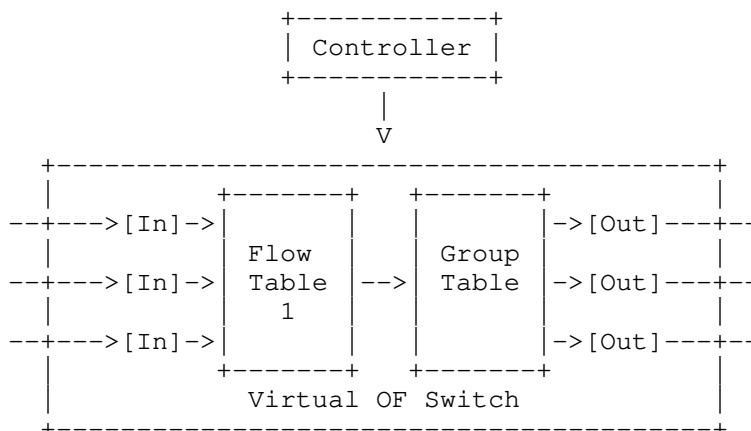


Figure 5: The Virtual OF Switch

Although the Virtual OF switch is the same at both the head-end and tail-end nodes, the group table is only used at the head-end node, where a Fast Failover group is set up for each pair of ports that correspond to the primary-backup Transport LSP pair. At the tail-end node, only flows are set up.

Transport LSPs are identified as virtual ports to the OF controller. For the Reference Topology in Figure 2, the primary transport LSP is identified to OpenFlow as Virtual Port [VP1] and the backup transport LSP is identified as Virtual Port [VP2]. At the head-end switch in the Reference Topology a Fast Failover group [PG1] is set up for ports [VP1] and [VP2].

The Virtual OF Switch MUST support the following OFPXMCT_OPENFLOW_BASIC match fields ([OF-1.3.0], Section A.2.3.7):

- o Match on Switch Input Port (OFPXMT_OFB_IN_PORT)
- o Match on MPLS Label (OFPXMT_OFB_MPLS_LABEL)
- o Match on MPLS BoS bit (OFPXMT_OFB_MPLS_BOS)
- o Match on Ethertype (OXM_OF_ETH_TYPE)

The Virtual OF Switch must support the following OF actions ([OF-1.3.0], Section A.2.5):

- o Output to switch port (OFPAT_OUTPUT)
- o Push a new MPLS tag (OFPAT_PUSH_MPLS)
- o Pop the outer MPLS tag (OFPAT_POP_MPLS)
- o Apply group (OFPAT_GROUP)

2.1.2. The OAM Engine

The liveness of Transport LSPs and PWs are monitored by OAM. The desired model is to have an OAM engine residing locally on the switch. At the ingress to a PW the OAM Engine will inject OAM packets into the PW data path. At the egress of a PW the switch will detect OAM packets (performing a flow-match operation) and punt OAM packets to the OAM Engine, which will evaluate them, and if need be perform corrective action and/or produce notifications.

The OAM function for Transport LSPs is out of scope for this revision of this document. However, PW-OAM mechanisms described in this document could also be applicable to Transport LSP-OAM.

In addition to generating and processing OAM packets, the OAM Engine participates in the liveness monitoring function (see Section 6.6 of [OF-1.3.0]) for virtual port Fast Failover groups at the head-end switch. When the OAM Engine detects a PW failure, it triggers the Virtual OF Switch to move traffic from the primary transport LSP's

virtual port [VP1] to the backup transport LSP's virtual port [VP2]. The OAM Engine's liveness monitoring function is described in more detail in [OF-1.3.0].

Note that in addition to the OAM Engine, the Virtual OF Switch MAY use other liveness monitoring mechanisms for the virtual port Fast Failover groups, which are out of scope of this document.

3. PW Configuration

3.1. Configuration Messages

The Controller uses OpenFlow protocol messages defined in [OF-1.3.0] to configure transport pseudo-wires. The Flow Modification message and the Group Modification message types are used. To configure a PW at the head-end node, the Controller uses a sequence of a Group Modification message followed by a Flow Modification message. At the tail-end node, the Controller uses Flow Modification messages only.

The message formats in this document are specified using Routing Backus-Naur Format (RBNF) encoding as specified in [RFC5511].

3.1.1. The Flow Modification Message

The Flow Modification message - 'Modify Flow Entry' - is defined in [OF-1.3.0], Section A.3.4.1 as follows:

```
<ofp-flow-mod> ::= <ofp-header>
                   <COOKIE>
                   <COOKIE_MASK>
                   <TABLE_ID>
                   <COMMAND>
                   <IDLE_TIMEOUT>
                   <HARD_TIMEOUT>
                   <PRIORITY>
                   <BUFFER_ID>
                   <OUT_PORT>
                   <OUT_GROUP>
                   <FLAGS>
                   <ofp-match>
                   <instructions>

<ofp-header> ::= <VERSION>
                 <OFP_MSG_TYPE>
                 <LENGTH>
                 <XID>
```

```
<ofp-match> ::= <MATCH_TYPE>
               <MATCH_LENGTH>
               <oxm_fields>

<oxm-fields> ::= <oxm-tlv> [<oxm-fields>]

<oxm-tlv> ::= <OXM_CLASS>
              <OXM_FIELD>
              <OXM_HASHMASK>
              <OXM_LENGTH>
              <PAYLOAD>

<instructions> ::= <instruction> [<instructions>]

<instruction> ::= ( <ofp-instruction-actions> |
                   <ofp-instruction-write-metadata> |
                   <ofp-instruction-goto-table> |
                   <ofp-instruction-meter> )

<ofp-instruction-actions> ::= <TYPE>
                              <LEN>
                              <PAD>
                              <actions>

<actions> ::= <ofp-action> [<actions>]

<ofp-action> ::= (<ofp-action-output> |
                  <ofp-action-group> |
                  <ofp-action-push-mpls> |
                  <ofp-action-pop-mpls> | ...)

<ofp-action-group> ::= <TYPE>
                      <LEN>
                      <GROUP_ID>

<ofp-action-output> ::= <TYPE>
                      <LEN>
                      <PORT>
                      <MAX_LEN>

<ofp-action-push-mpls> ::= <TYPE>
                          <LEN>
                          <ETHERTYPE>
                          <MPLS_HEADER>

<ofp-action-pop-mpls> ::= <TYPE>
                        <LEN>
```

<ETHERTYPE>

Figure 6: The 'Modify Flow Entry' message

Note that not all action types defined in [OF-1.3.0] for <ofp-action> are listed in Figure 6.

3.1.2. The Group Modification Message

The Group Modification message - 'Modify Group Entry' - is defined in [OF-1.3.0], Section A.3.4.2 as follows:

```

<ofp-group-mod> ::= <ofp-header>
                    <COMMAND>
                    <GROUP_MSG_TYPE>
                    <GROUP_ID>
                    <buckets>

<ofp-header> ::= <VERSION>
                 <OFF_MSG_TYPE>
                 <LENGTH>
                 <XID>

<buckets> ::= <ofp-bucket> [<buckets>]

<ofp-bucket> ::= <LEN>
                 <WEIGHT>
                 <WATCH_PORT>
                 <WATCH_GROUP>
                 <actions>

<actions> ::= <ofp-action> [<actions>]

<ofp-action> ::= (<ofp-action-output> | <ofp-action-group> | ...)

<ofp-action-output> ::= <TYPE>
                       <LEN>
                       <PORT>
                       <MAX_LEN>

```

Figure 7: The 'Modify Group Entry' message

3.2. PW Head-End Node Configuration

Consider the reference topology in Figure 2. The Management Application will configure a cross-connect between the Attachment Circuit [AC1] and the virtual port pair {[VP1], [VP2]} joined in the Fast Failover group [PG1]. The cross-connect determines that traffic

from Port AC will be switched to Group PG1. The internal mechanism in Group PG1 (outside the scope of this document) will determine whether traffic will go out on Port [VP1] (the primary transport LSP) or on Port [VP2] (the backup transport LSP).

The Management Application uses the following message sequence to create the cross-connect:

1. The 'Modify Group Entry' message, defined in Figure 7 creates or modifies an entry in the Group Table. Each entry in the Group Table corresponds to a pair of virtual ports that correspond to a pair of primary / backup transport LSPs. This entry states that as long as Port [VP1] is alive, traffic coming to group [PG1] will go out on [VP1]. If Port [VP1] is not alive AND Port [VP2] is alive, then traffic coming to group [PG1] will go out on Port [VP2]. If neither of the ports are alive, traffic will be dropped. Note that it's up to the switch to determine that a given port is alive - and it can use any mechanism that it wants to do that.
2. The 'Modify Flow Entry' message, defined in Figure 6, adds an entry to Flow Table 1 for a flow that matches traffic from Input Port [AC1]. The actions for the flow are 1.) Push the PW MPLS header on the packet, and 2.) Forward the packet to Group [PG1], which was setup in Step 1).

The following sections describe in details each message.

3.2.1. 'Modify Group Entry' Message Details

The fields in the 'Modify Group Entry' message are set as follows:>

'Modify Group Entry' message: <COMMAND> is set to 'OFPGC_ADD' or 'OFPGC_MODIFY', <GROUP_MSG_TYPE> is set to 'OFPGT_FF' (fast failover group) and <GROUP_ID> is set to the identifier of the Fast Failover group that was setup for the primary and secondary transport LSPs - [PG1].

OpenFlow Header (ofp-header): <VERSION> is set to 4, <OFF_MSG_TYPE> is set to 'OFPT_GROUP_MOD'.

Buckets: there are two action buckets - Bucket1 and Bucket2:

Bucket1 is associated with the virtual port corresponding to the primary transport LSP, and its fields are set as follows.
<WEIGHT> is set to 1, <WATCH_PORT> is set to [VP1],
<WATCH_GROUP> is set to 'OFPG_ANY'. <action-list> contains a single item - an <ofp-action-output> to Virtual Port [VP1].

Bucket2 is associated with the virtual port corresponding to the backup transport LSP, and its fields are set as follows.
<WEIGHT> is set to 10, <WATCH_PORT> is set to [VP2],
<WATCH_GROUP> is set to 'OFFPG_ANY'. <action-list> contains a single item - an <ofp-action-output> to Virtual Port [VP2].

3.2.2. 'Modify Flow Entry' Message Details

The fields in the Modify Flow Entry' message are set as follows:

'Modify Flow Entry' message: The controller MUST set the value of <TABLE_ID> to '1', the value of <COMMAND> 'OFFPC_MODIFY_STRICT', the value of <BUFFER_ID> to 'OFF_NO_BUFFER', the value of <OUT_PORT> to 'OFFP_ANY', and the value of <OUT_GROUP> to 'OFFPG_ANY'. The Controller SHOULD set the values of all other atomic fields to appropriate values as required by the operation of the configuration protocol. It is recommended that the 'IDLE_TIMEOUT' and 'HARD_TIMEOUT' fields are set to 0 for persistent PW configurations

OpenFlow Header (ofp-header): <VERSION> is set to 4, <OFF_MSG_TYPE> is set to 'OFFPT_GROUP_MOD'.

Ofp-Match: The Controller MUST set the <MATCH_TYPE> field to 'OFFPMT_OXM' and include a single <oxm-tlv> with the <OXM_CLASS> field set to 'OFFPXM_OPENFLOW_BASIC', the <OXM_FIELD> field set to 'OFFPXM_OFB_IN_PORT', the <OXM_HASHMASK> field set to '0', and the <PAYLOAD> field set to [AC1].

Instructions: The Controller MUST set the type field <TYPE> to 'OFFPIT_APPLY_ACTIONS' and include the following action list:

Push MPLS Header: the value of <TYPE> set to 'OFFPAT_PUSH_LABEL'; the value of <ETHERTYPE> set to MPLS Unicast; the value of <MPLS_HEADER> set as follows: Label=[PWL1], TTL=1, TC=???, S=1.

Group: the value of <TYPE> set to 'OFFPAT_GROUP' and the value of <GROUP_ID> set to [PG1].

3.3. PW Tail-End Node Configuration

Consider the reference topology in Figure 2. The Management Application will configure two cross-connects: one cross-connect between the primary Transport LSP's virtual port ([VP3]) and the Attachment Circuit [AC2], and one between the primary transport LSP's virtual port ([VP3]) and the Attachment Circuit [AC2]. Under normal circumstances, traffic will arrive at the primary Transport LSP's Virtual Port [VP3]. When the primary Transport LSP is not available

and the backup Transport LSP is ok, traffic will arrive at the backup Transport LSP's Virtual Port [VP4].

Note that the cross-connect between the backup Transport LSP's Virtual Port [VP4] and the Attachment Circuit [AC2] can be programmed along with the cross-connect between the primary Transport LSP's Virtual Port [VP3] and the Attachment Circuit [AC2], or at the time when the primary Transport LSP's Virtual Port [VP3] goes down.

The cross-connects between the primary Transport LSP's Virtual Port [VP3] and the Attachment Circuit [AC2], and between the backup transport LSP's Virtual Port [VP4] and the Attachment Circuit [AC2] are programmed by sending 'Modify Flow Entry' messages to the switch. Programming details are described in the following section.

3.3.1. 'Modify Flow Entry' Message Details

The fields in the Modify Flow Entry' messages are set as follows:

'Modify Flow Entry' message: The controller MUST set the value of <TABLE_ID> to '1', the value of <COMMAND> 'OFPFC_MODIFY_STRICT', the value of <BUFFER_ID> to 'OFP_NO_BUFFER', the value of <OUT_PORT> to 'OFPP_ANY', and the value of <OUT_GROUP> to 'OFPG_ANY'. The Controller SHOULD set the values of all other atomic fields to appropriate values as required by the operation of the configuration protocol. It is recommended that the 'IDLE_TIMEOUT' and 'HARD_TIMEOUT' fields are set to 0 for persistent PW configurations

OpenFlow Header (ofp-header): <VERSION> is set to 4, <OFP_MSG_TYPE> is set to 'OFPT_GROUP_MOD'.

Ofp-Match: The Controller MUST set the <MATCH_TYPE> field to 'OFPMT_OXM' and include the following <oxm-tlv> match list:

Match Input Port: the value of the <OXM_CLASS> field set to 'OFPXMC_OPENFLOW_BASIC'; the value of the <OXM_FIELD> field set to 'OFPXMT_OFB_IN_PORT'; the value of the <OXM_HASHMASK> field set to '0' and the value <PAYLOAD> field set to [VP3] (for the primary Transport LSP) or [VP4] (for the backup Transport LSP).

Match MPLS Label: the value of the <OXM_CLASS> field set to 'OFPXMC_OPENFLOW_BASIC'; the value of the <OXM_FIELD> field set to 'OFPXMT_OFB_MPLS_LABEL'; the value of the <OXM_HASHMASK> field set to '0' and the value <PAYLOAD> field set to [PWL1].

Match BoS bit: the value of the <OXM_CLASS> field set to 'OFPXMC_OPENFLOW_BASIC'; the value of the <OXM_FIELD> field set to 'OFPXMT_OFP_MPLS_BOS'; the value of the <OXM_HASHMASK> field set to '0' and the value <PAYLOAD> field set to 1.

Instructions: The Controller MUST set the type field <TYPE> to 'OFPIT_APPLY_ACTIONS' and include the following action list:

Pop MPLS Header: the value of <TYPE> set to 'OFPAT_POP_MPLS' and the value of <ETHERTYPE> set to MPLS Unicast.

Output: the value of <TYPE> set to 'OFPAT_OUTPUT' and the value of <PORT> set to [AC2].

4. PW OAM Considerations

4.1. OAM Overview

OAM for MPLS-TP is an important consideration and needs to be addressed in a scalable manner and needs to function with the same performance available today. Centralization of control or digestion of OAM messages, where they are redirected back to a central controller will introduce delay. Therefore the goal is to drive OAM setup for PWs using messages from the Controller to the Switch. The functions of the OAM, for example OAM packet generation, error detection, action/notification will therefore still reside locally on the switch.

[RFC6423] is used as a reference for unified OAM for MPLS-TP, in particular Section 3, which includes provision for GAL with PW in MPLS-TP. [RFC5860] lists OAM requirements for MPLS Transport networks.

4.2. PW OAM Engine Configuration

If OAM is required on a particular PW it requires only a small number of configuration actions on the OAM Engine and on the Virtual OF Switch.

- o The head-end OAM Engine is programmed to generate OAM packets for the PW. The header for these packets will be composed of at least 2 labels, the first being the PW label for which the OAM is being generated, and the following label being the GAL label (13). The contents of the GACH packet are out of scope for this draft and will be down to the individual OAM implementation.

- o At the head-end Virtual OF switch, a flow is programmed for each Pseudowire to switch the OAM packets generated by the OAM Engine to their respective destination PW. As the PW label is being placed on the OAM packet, we can easily match on this and forward the OAM packet down the correct PW to ensure it follows the same data path.
- o At the tail-end Virtual OF switch, a single flow is programmed to switch all received OAM packets to the OAM Engine. The match criteria on the flow is the MPLS BoS bit set to 0, which means that a GAL label is present on the packet. This flow MUST have higher priority than any flow matching on a PW label.
- o The tail-end OAM Engine is programmed to receive OAM packets, check that a valid PW label is present on the packet and to detect failures.

OAM mechanisms that can be implemented by the OAM Engine are out of scope for this revision of the document. For example, the OAM Engine can also implement BFD (Echo) Mode [RFC5880] where echo packets are returned via the remote forwarding plane, which can be done using an OF match rule.

4.3. OAM and S-Bit considerations

In order to ensure that the switch can identify the last label in the stack the S bit needs to be set on the label which will be at the bottom of the stack. The OAM Engine will be required to set the S bit on the GAL label (13) to ensure that the subsequent G-ACh packet is treated correctly. By using OF actions to move all OAM Engine packets into the PW we ensure that not only all types of OAM are supported transparently, but also that the S bit is correctly set.

5. IANA Considerations

This document does not introduce any IANA requirements.

6. Security Considerations

Procedures described in this document do not change the OpenFlow protocol security model described in [OF-1.3.0], Section 6.3.

A secure communications channel SHOULD be set up between the controller and the MPLS-TP node.

7. Acknowledgements

The authors would like to thank Frank Brockners, Dan Frost, Giles Heron, Zoltan Lajos Kis, Andy Malis, Yakov Stein, Joe Tardo, Sasha Vainshtein and Dave Ward for their review and insightful comments.

8. Normative References

- [OF-1.3.0] Open Networking Foundation, "OpenFlow Switch Specification, Version 1.3.0 (Wire Protocol 0x04)", April 16, 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.
- [RFC5601] Nadeau, T. and D. Zelig, "Pseudowire (PW) Management Information Base (MIB)", RFC 5601, July 2009.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6373] Andersson, L., Berger, L., Fang, L., Bitar, N., and E. Gray, "MPLS Transport Profile (MPLS-TP) Control Plane Framework", RFC 6373, September 2011.
- [RFC6423] Li, H., Martini, L., He, J., and F. Huang, "Using the Generic Associated Channel Label for Pseudowire in the MPLS Transport Profile (MPLS-TP)", RFC 6423, November 2011.

Authors' Addresses

Jan Medved
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: jmedved@cisco.com

Andrew McLachlan
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: amclachl@cisco.com

David Meyer
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: dmn@cisco.com

PWE3
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2013

T. Nadeau
Juniper Networks
C. Pignataro
Cisco Systems, Inc.
YJ. Stein
RAD Data Communications
July 4, 2012

Virtual Circuit Connectivity Verification version 2 (VCCV2)
draft-nadeau-pwe3-vccv2-00.txt

Abstract

This document describes VCCV2, a new version of Virtual Circuit Connectivity Verification (VCCV), the pseudowire OAM mechanism. This new version is backwards compatible with VCCV for MPLS PWs for modes that the versions share, although the Router Alert (RA) CV type is not supported by VCCV2. Furthermore, this document collects the complete description of VCCV2 into a single specification.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 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. Overview of the PW OAM Channel	3
3. Abbreviations	4
4. The Protocol and its Options	4
5. Security Considerations	4
6. IANA Considerations	4
7. References	5
7.1. Normative References	5
7.2. Informative References	5
Authors' Addresses	6

1. Introduction

Virtual Circuit Connectivity Verification (VCCV), the pseudowire OAM mechanism is described in [RFC5085], [RFC5885], and [I-D.ietf-pwe3-vccv-for-gal]. This mechanism has been widely implemented and deployed, but it has been reported [I-D.ietf-pwe3-vccv-impl-survey-results] that the large number of VCCV options has led to interoperability issues.

[RFC5085] together with [I-D.ietf-pwe3-vccv-for-gal] define four Control Channel (CC) types for MPLS PWs:

- Type 1 using the control word (CW),
- Type 2 using the Router Alert label (label=1) above the PW label,
- Type 3 using TTL expiry,
- Type 4 using G-ACh Label (label=13) [RFC5586] below the PW label.

In order to simplify implementations and operations, we herein obsolete Type 2, and provide guidance as to when to use the remaining three types.

[RFC5085] together with [RFC5885] define four Connectivity Verification (CV) types for MPLS PWs:

- ICMP ping,
- LSP ping,
- BFD with UDP/IP encapsulation,
- raw BFD (without IP encapsulation),

and BFD has several options of its own (see [RFC5880]). The description of what and how to implement these is spread over several documents, and we herein attempt to summarize the entire functionality set in one place.

This document only describes OAM for PWs over MPLS. Functionality for L2TPv2-based PWs remains as presently specified.

The present version of this document is a skeleton only, intended to initiate discussion. Once the principles are agreed upon, the authors will flesh out the rest.

2. Overview of the PW OAM Channel

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].

VCCV and VCCV2 are fault OAM mechanisms to verify liveliness and to further diagnose the pseudowire forwarding path. This section will provide an overview of the requirements and architecture of PW OAM.

3. Abbreviations

AC Attachment Circuit [RFC3985]
CC Control Channel (used as CC Type)
CE Customer Edge
CV Connectivity Verification (used as CV Type)
CW Control Word [RFC3985]
GACH Generic Associated Channel [RFC5586]
GAL GACH Channel Label [RFC5586]
MPLS-TP MPLS-Transport Profile
OAM Operations, Administration and Maintenance
PE Provider Edge
PSN Packet Switched Network [RFC3985]
PW Pseudowire [RFC3985]
PW-ACH PW Associated Channel Header [RFC4385]
VCCV Virtual Circuit Connectivity Verification

4. The Protocol and its Options

This section will detail all the CC and CV options, the signaling needed to choose each of them, the bit-masks and codings. The description will be concise, yet readable.

In particular, CC Type 2 is obsoleted. Subsections will discuss Types 1, 3, and 4.

In addition, the text will provide guidance for selection of CC types, as follows: When the PW employs a CW then CC Type 1 SHOULD be used. TDM PWs always use the CW, and thus SHOULD always use Type 1. Legacy (ATM, port mode frame relay, and HDLC PWs) without CWs SHOULD use Type 3. [RFC5994] states that Ethernet PWs over MPLS-TP MUST use the CW, and thus they SHOULD use Type 1, but MAY use Type 4.

Discussion is needed as to whether all CV types are required. Subsections will detail the use of the different CV types.

5. Security Considerations

Are there significant threats on PWs based on VCCV?

6. IANA Considerations

It is not clear what needs to be put here. Will CC Type 2 be removed?

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.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5885] Nadeau, T. and C. Pignataro, "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.
- [RFC5994] Bryant, S., Morrow, M., Swallow, G., Cherukuri, R., Nadeau, T., Harrison, N., and B. Niven-Jenkins, "Application of Ethernet Pseudowires to MPLS Transport Networks", RFC 5994, October 2010.
- [I-D.ietf-pwe3-vccv-for-gal] Nadeau, T. and L. Martini, "A Unified Control Channel for Pseudowires", draft-ietf-pwe3-vccv-for-gal-01 (work in progress), May 2012.

7.2. Informative References

- [I-D.ietf-pwe3-vccv-impl-survey-results] Regno, N., "The Pseudowire (PW) & Virtual Circuit Connectivity Verification (VCCV) Implementation Survey Results", draft-ietf-pwe3-vccv-impl-survey-results-00 (work in progress), April 2012.

Authors' Addresses

Thomas D. Nadeau
Juniper Networks

Email: tnadeau@juniper.net

Carlos Pignataro
Cisco Systems, Inc.
7200-12 Kit Creek Road
PO Box 14987
Research Triangle Park, NC 27709
USA

Email: cpignata@cisco.com

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

Email: yaakov_s@rad.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: December 27, 2012

Y. Shen, Ed.
Juniper Networks
R. Aggarwal
Arktan, Inc
W. Henderickx
Alcatel-Lucent
June 25, 2012

PW Endpoint Fast Failure Protection
draft-shen-pwe3-endpoint-fast-protection-02

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on December 27, 2012.

Copyright Notice

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

document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Specification of Requirements	4
3. Reference Models and Failure Cases	4
3.1. Single-Segment PW	5
3.2. Multi-Segment PW	7
4. Theory of Operation	8
4.1. Local Repair and Protector	8
4.2. Context Identifier	11
4.2.1. Uses of Context Identifier	11
4.2.2. Advertisement and Path Computation	12
4.3. Protection Models	14
4.4. Transport Tunnel	17
4.5. Bypass Tunnel	18
4.6. PW Forwarding State on Protector	18
4.6.1. Co-located Protector	19
4.6.2. Centralized Protector	20
4.7. PW Label Distribution from Primary PE to Protector	22
4.7.1. Protection FEC Element Encoding for PWid	24
4.7.2. Protection FEC Element Encoding for Generalized PWid	25
4.8. PW Label Distribution from Backup PE to Protector	26
4.9. Revertive Behavior	27
5. IANA Considerations	28
6. Security Considerations	28
7. Acknowledgements	28
8. References	28
8.1. Normative References	28
8.2. Informative References	30
Authors' Addresses	30

1. Introduction

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

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

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

This document specifies a fast protection mechanism for PWs based on local repair technique. It can protect PWs against the following types of failures.

a. Egress AC failure.

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

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

The mechanism is applicable to LDP signaled PWs.

2. Specification of Requirements

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

3. Reference Models and Failure Cases

This document refers to the following topologies to describe PW endpoint failures and protection procedures. These topologies are commonly seen in an environment with multi-homed CEs and redundant PWs for global repair. In this document the fast protection mechanism also use them for the local repair purposes. This SHALL enable local repair and global repair to work in tandem to achieve broader scope of protection with better performance.

3.1. Single-Segment PW

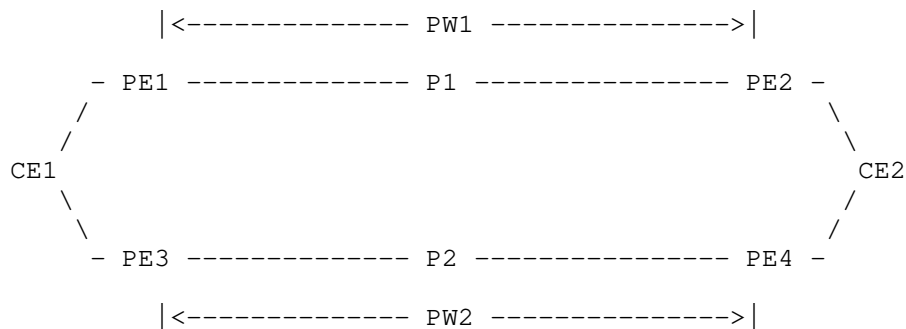


Figure 1

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

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

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

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

In this document, the following primary and backup roles are assigned

for the traffic going from CE1 to CE2:

Primary PW: PW1

Primary PE: PE2

Primary AC: CE2-PE2

Backup PW: PW2

Backup PE: PE4

Backup AC: CE2-PE4

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

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

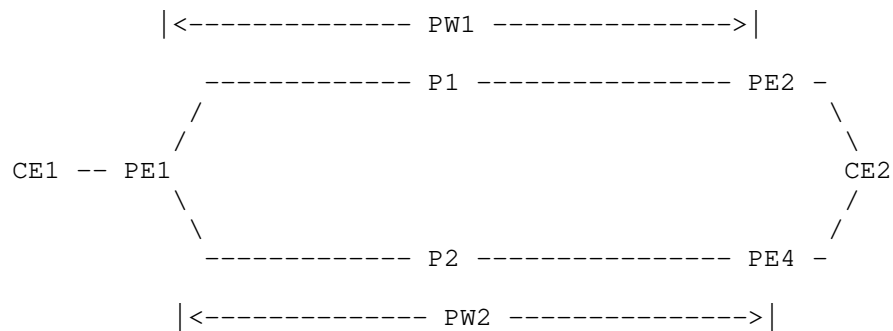


Figure 2

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

3.2. Multi-Segment PW

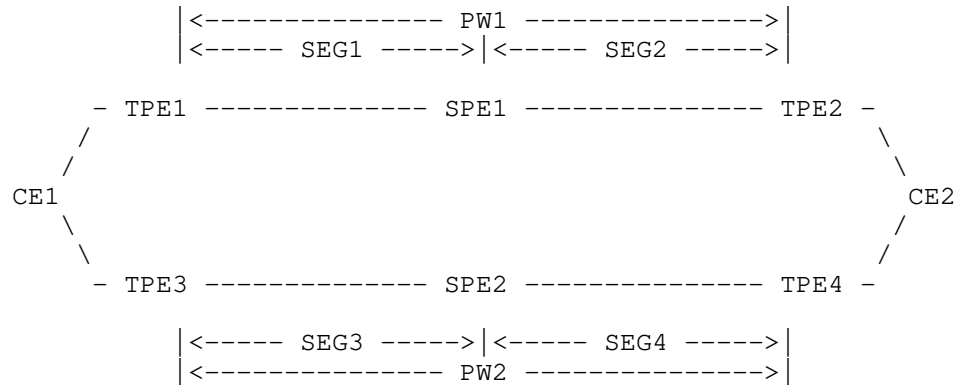


Figure 3

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

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

Primary PW: PW1

Primary T-PE: TPE2

Primary S-PE: SPE1

Primary AC: CE2-TPE2

Backup PW: PW2

Backup T-PE: TPE4

Backup S-PE: SPE2

Backup AC: CE2-TPE4

In this case, an egress AC failure refers to the failure of the primary AC, i.e. the AC CE2-TPE2. An egress node failure refers to

the failure of the primary T-PE, i.e. TPE2. In addition, an switching node failure refers to the failure of the primary S-PE, i.e. SPE1.

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

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

4. Theory of Operation

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

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

The mechanism is only relevant when the target CE is multi-homed to a primary PE and a backup PE, and when there is a backup PW in the network. In S-PE node protection, it is also assumed that there is a backup S-PE on the backup PW.

4.1. Local Repair and Protector

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

Different routers may serve as PLRs and protectors in different scenarios.

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

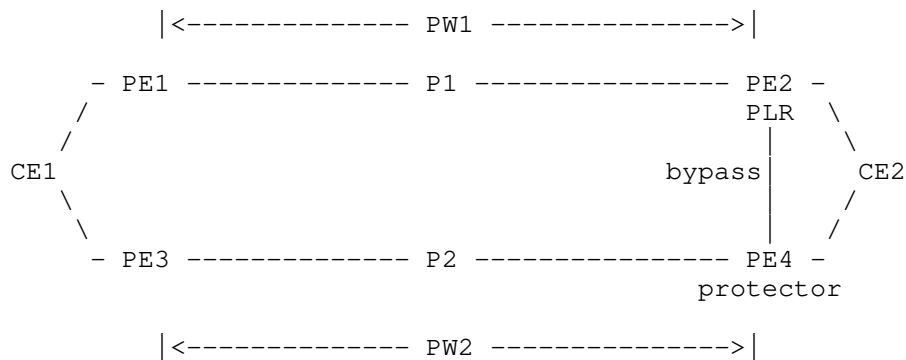


Figure 4

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

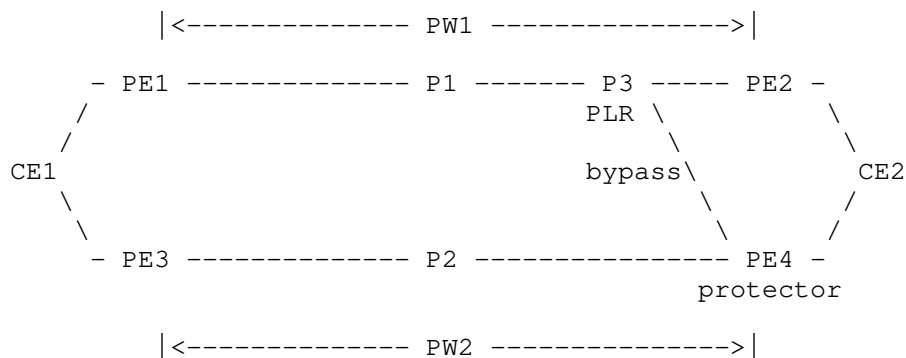


Figure 5

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

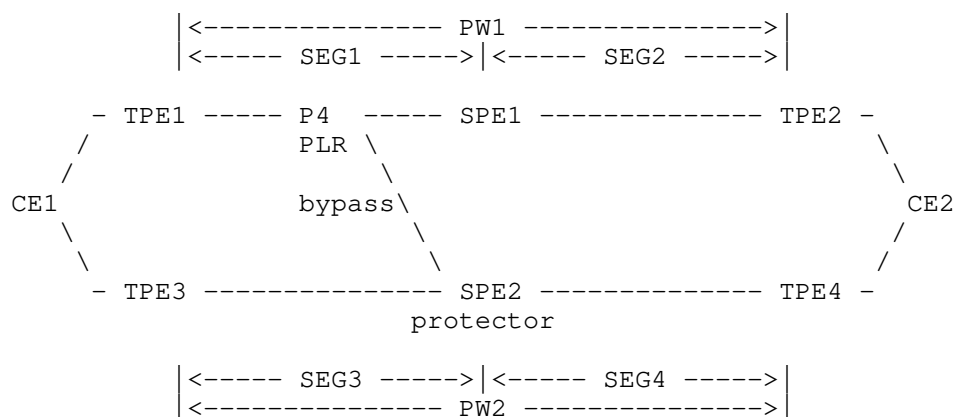


Figure 6

When a PLR forwards traffic through a bypass tunnel to a protector, it MUST keep the original PW label intact. In particular, it SHOULD NOT forward the traffic based on the PW label or modify the PW label. Such forwarding state on the PLR has the advantages that it represents simple forwarding operations and it is easy to set up. The PLR does not need to learn PW labels or install bypass routes on a per PW label basis.

This also means that the protector MUST forward the traffic based on a PW label that is assigned by the primary PE, and ensure that the traffic can eventually reach the target CE. From the protector's perspective, this PW label is an upstream assigned label (RFC 5331). This is accomplished by learning the PW label from the primary PE, installing the proper forwarding state for the PW label in the label space associated with the primary PE, and performing PW label lookup in this label space.

A protector MAY be a backup (S-)PE as illustrated in the above examples, or a dedicated router that assumes such a role. In the later case, the protector is not necessarily the backup (S-)PE of a given primary PW. During a local repair, the PLR still forwards traffic to the protector through a bypass tunnel, and the protector MUST then forward the traffic to the backup (S-)PE, which finally forwards the traffic to the target CE via a backup AC or a backup PW segment. More detail will be provided in Section 4.3.

A protector MAY protect primary PWs for one or multiple primary PEs. The protector MUST maintain a separate label space for each primary PE. Likewise, the primary PWs hosted by a primary PE MAY be protected by multiple protectors, each for a subset of the PWs. In any case, a primary PW is associated with one and only one pair of

{primary PE, protector}.

4.2. Context Identifier

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

4.2.1. Uses of Context Identifier

A context identifier serves two purposes.

- o It identifies a primary PE and an associated protector. In other words, it identifies a primary PE on a per protector basis. A given primary PE may be protected by multiple protectors, each for a subset of the primary PWs hosted by the primary PE. Therefore, a distinct context identifier MUST be assigned to the primary PE for each protector.

For a primary PW, its transport tunnel MUST be destined for the context identifier of its {primary PE, protector}, rather than an IP address of the primary PE. This not only enables the transport tunnel to follow a path to the primary PE, but also indicates the protector to the PLR(s).

- o It indicates the primary PE's label space to a protector. The protector may protect primary PWs for multiple primary PEs. It MUST maintain a separate label space for each primary PE. PW labels assigned by a given primary PE MUST be associated with the label space indicated by the context identifier of the {primary PE, protector}. The association is accomplished as below.

When the primary PE advertises the label of a primary PW to the protector, it MUST attach the information of the context identifier (Section 4.7). Upon receiving the advertisement, the protector MUST install the PW label in the label space corresponding to the context identifier.

A bypass tunnel MUST be destined for the context identifier, rather than an IP address of the protector. Therefore, the bypass tunnel (either MPLS tunnel label or IP tunnel destination address) is equivalent to the context identifier. All packets received on the bypass tunnel MUST be forwarded in the label space indicated by the bypass tunnel.

4.2.2. Advertisement and Path Computation

Using a context identifier as destination for both transport and bypass tunnels imposes the following requirements on path computation for these tunnels.

- o On the ingress PE, path computation for a transport tunnel MUST choose the primary PE as the endpoint.
- o On a PLR, path computation for a bypass tunnel MUST avoid the primary PE and choose the protector as the endpoint. The path MUST NOT traverse the primary PE.

In order to satisfy these requirements, a context identifier SHOULD be advertised by IGP and/or IGP-TE in the routing domain and/or the TE domain, depending on the tunnel technologies of the network.

- o If RSVP-TE is used to establish both transport and bypass tunnels, the context identifier MUST be advertised by IGP-TE.
- o If IP or LDP is used to establish both transport and bypass tunnels, the context identifier MUST be advertised by IGP.
- o If IP or LDP is used for transport tunnels while RSVP-TE is used for bypass tunnels, or vice versa, the context identifier MUST be advertised by both IGP and IGP-TE.

In any case, it is recommended that the context identifier SHOULD be advertised as a proxy node that is dual-attached to the primary PE and the protector via unnumbered point-to-point interfaces, as shown in Figure 7. This schema ensures that the CSPF (constrained shortest path first), LFA (loop free alternate; RFC 5286) and MRT (maximally redundant trees; [IP-LDP-FRR-MRT]) algorithms can compute the expected paths for the transport tunnel and bypass tunnel, whether the tunnels are MPLS tunnels or IP tunnels.

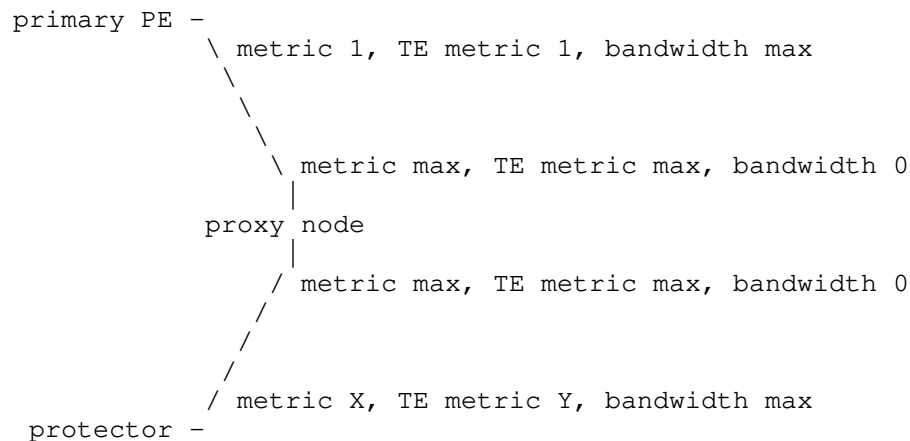


Figure 7

- o The primary PE advertises an unnumbered link to the proxy node, with metric 1, TE metric 1, and maximum bandwidth.
- o The protector advertises an unnumbered link to the proxy node, with metric X, TE metric Y, and maximum bandwidth. X SHOULD be carefully chosen so that the path from any given source node (ingress PE or PLR) via the protector to the proxy node will have a higher metric than the corresponding path from the source node via the primary PE to the proxy node. The same requirement applies to Y as well for TE paths.
- o The primary PE advertises the proxy node with two unnumbered links to the primary PE and the protector, respectively. The router ID of the proxy node is the context identifier. Both unnumbered links are advertised with maximum metric, maximum TE metric, and zero bandwidth. This ensures that the proxy node does not serve as a transit node for any paths.

In the case of ISIS [ISO10589], the system ID is derived from the context identifier with Binary Coded Decimal (BCD) encoding. The resulting system-ID MUST be unique. The LSP (Link State Packet) MUST include an Area Address TLV, and MAY include a Dynamic Hostname TLV. The area addresses MUST be a subset of or preferably identical to those advertised by the primary PE at the corresponding level. The hostname MAY be derived from the context identifier and the primary PE's hostname. The Overload bit MUST be set to 1. The Attached and the Partition Repair bits MUST be set to 0.

In the case of OSPF (RFC 2328), the Advertising Router and Link

State ID of the router LSA (Link State Advertisement) MUST both be the context identifier. All options bits in the router LSA MUST be set to zero.

With this schema, the proxy node is reachable via both the primary PE and the protector in the routing domain and the TE domain. For any given ingress PE or PLR, the path via the primary PE to the proxy node is considered to have a higher preference than the path via the protector, due to the lower metric. Therefore, in path computation for a transport tunnel, a path via the primary PE SHOULD always be selected. However, in path computation for a bypass tunnel, where the primary PE must be avoided, a path via the protector SHOULD be selected.

4.3. Protection Models

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

1. Co-located protector

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

In egress AC protection and egress PE node protection, when a protector receives traffic from the PLR, it forwards the traffic to the CE via the backup AC. This is shown in Figure 8, where PE2 is the PLR for egress AC failure, P3 is the PLR for PE2 failure, and PE4 (the backup PE) is the protector.

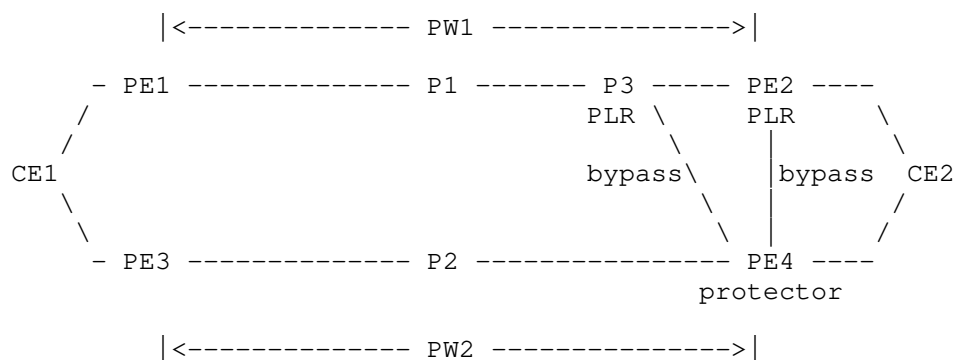


Figure 8

In S-PE node protection, when a protector receives traffic from the PLR, it MUST forward the traffic via the next segment of the backup PW. The T-PE of the backup PW MUST forward the traffic to the CE via a backup AC. This is shown in Figure 9, where P4 is the PLR for SPE1 failure, and SPE2 (the backup S-PE) is the protector for SPE1 (the primary S-PE).

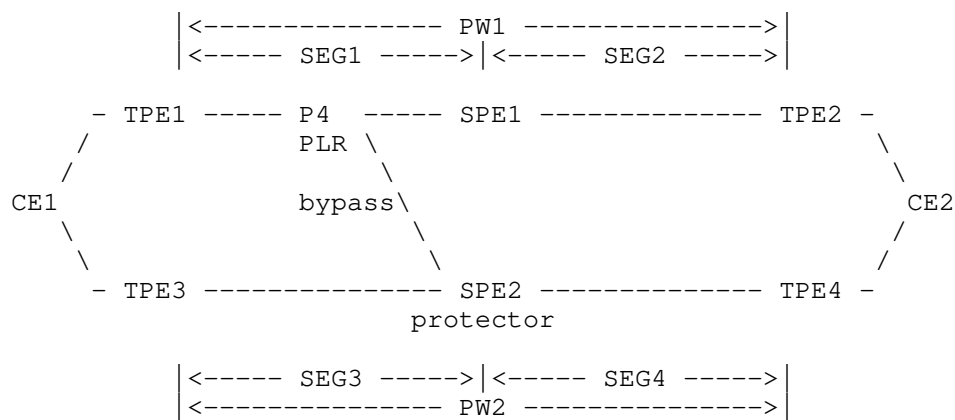


Figure 9

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

2. Centralized protector

In this model, the protector is a dedicated P router or PE router that protects all the primary PWs for one or multiple primary PEs. In egress AC protection and egress PE node protection, the protector MAY or MAY NOT be a backup PE with a direct connection to the target CE. In S-PE node protection, it MAY or MAY NOT be a backup S-PE of the backup PW.

In egress AC protection and egress PE node protection, when the protector receives traffic from the PLR, if the protector has a direct connection (i.e. backup AC) to the CE, it MUST forward the traffic to the CE via the backup AC, which is similar to Figure 8. Otherwise, it MUST forward the traffic to a backup PE, which MUST then forward the traffic to the CE via a backup AC. This is shown in Figure 10, where the protector receives traffic from P3 or PE2 (the PLRs) and forwards the traffic to PE4 (the backup PE). The protector may be protecting other PWs as well, which are not shown in this figure.

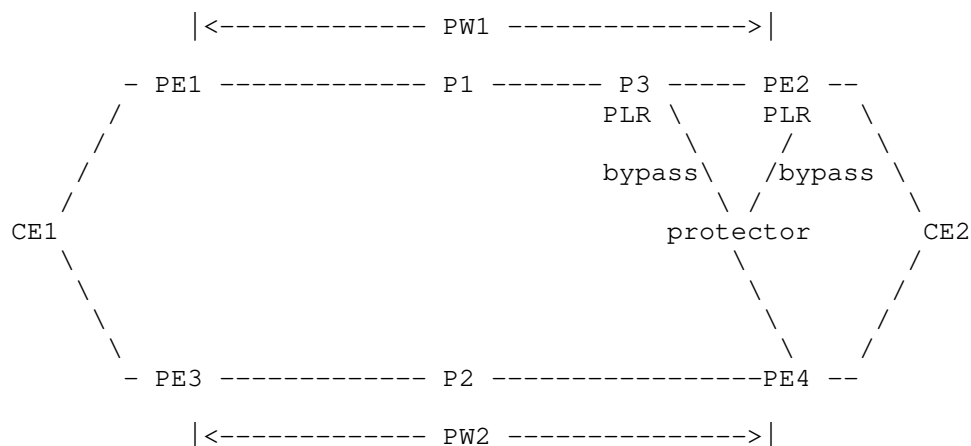


Figure 10

In S-PE node protection, when the protector receives traffic from the PLR, if the protector is a backup S-PE of the backup PW, it MUST forward the traffic via the next segment of the backup PW, and the T-PE of the backup PW MUST forward the traffic to the CE via a backup AC, which is similar to Figure 9. Otherwise, the protector MUST first forward the traffic to the backup S-PE, which MUST then forward the traffic via the next segment of the backup PW. Finally, the T-PE of the backup PW MUST forward the

traffic to the CE via a backup AC. This is shown in Figure 11, where the protector forwards traffic to SPE2 (the backup S-PE). The protector may be protecting other PW segments as well, which are not shown in this figure.

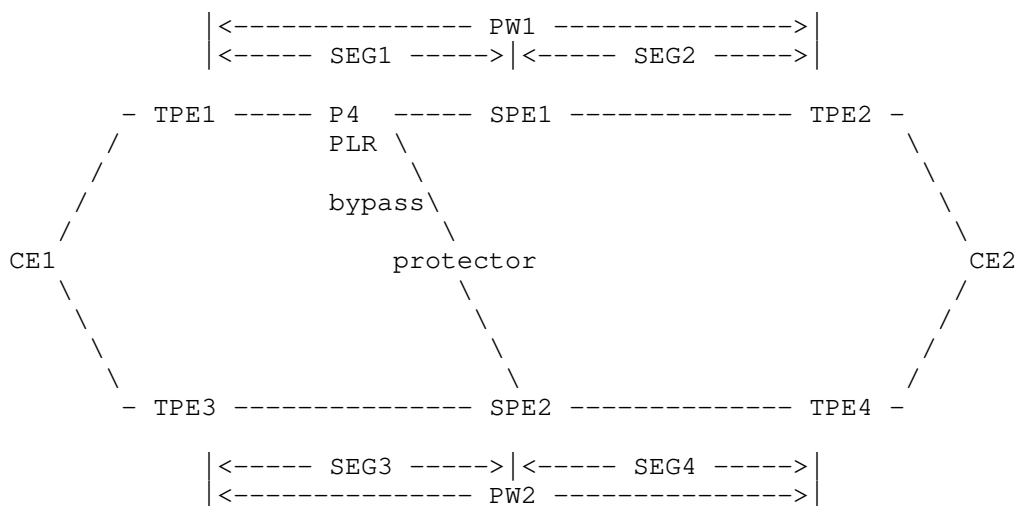


Figure 11

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

4.4. Transport Tunnel

The ingress PE of a primary PW (or PW segment) associates the PW with the primary egress PE through LDP signaling. The ingress PE MUST also associate the transport tunnel of the PW with the context identifier of the {primary PE, protector} of the PW. In particular, the destination of the transport tunnel MUST be the context identifier (Section 4.2.1). This not only ensures PW traffic to be transported to the primary PE, but also facilitates bypass tunnel establishment for PLR(s), as the context identifier implies both the primary PE and the protector.

The association between the transport tunnel and the context identifier MAY be achieved by configuration or an auto-discovery mechanism. In the later case, the ingress PE MAY learn the context identifier from the primary PE, if the primary PE advertises the

context identifier as "third party next hop" in an IPv4/v6 Interface_ID TLV (RFC 3471, RFC 3472) in LDP Label Mapping message.

4.5. Bypass Tunnel

A PLR may provide protection for multiple primary PWs associated with one or multiple pairs of {primary PE, protector}. The PLR MUST establish a bypass tunnel to each protector for each distinct context identifier associated with the protector. The destination of the bypass tunnel MUST be the context identifier, as described in Section 4.2.1. The association between the destination and the context identifier can be achieved by PLR learning or inheriting destination address from the transport tunnel.

For examples, in Figure 8 and Figure 10, a bypass tunnel is established from PE2 (PLR for egress AC failure) to the protector, and another bypass tunnel is established from P3 (PLR for egress node failure) to the protector. In Figure 9 and Figure 11, a bypass tunnel is established from P4 (PLR for switching node failure) to the protector.

During a local repair, the PLR forwards traffic to the protector through the bypass tunnel with PW label intact. This normally involves pushing an MPLS label to the label stack, if the bypass tunnel is an MPLS tunnel, or pushing an IP header to the packet, if the bypass tunnel is an IP tunnel. The protector MUST then forward the traffic based on this PW label, i.e. an upstream assigned label. In order to perform such forwarding, the protector MUST treat the bypass tunnel as a context to determine the primary PE's label space. Specifically, if the bypass tunnel is an MPLS tunnel, the protector MUST assign a non-reserved label for the bypass tunnel, and use this label as the context. If the bypass tunnel is an IP tunnel, the destination address in its IP header should be the context identifier.

A bypass tunnel MUST have the property that it is not affected by the topology changes caused by the failure that the bypass tunnel protects against. Therefore, it can be used to transmit traffic during the convergence period of routing protocols and the delay of global repair. It will remain effective, until the current transport tunnel is rerouted around the failure, or the traffic is moved to another PW or transport tunnel.

4.6. PW Forwarding State on Protector

A protector MUST be able to forward traffic based on the PW label assigned by a primary PE. Therefore, it MUST learn the PW labels from all the primary PEs that it protects (Section 4.7), and maintain

the PW labels in separate label spaces for the primary PEs.

In the control plane, a primary PE's label space is identified by the context identifier of the {primary PE, protector}. When the protector learns a PW label from the primary PE, it MUST associate the PW label with the label space via this context identifier. In the forwarding plane, the label space is indicated by bypass tunnels that are destined for the context identifier.

4.6.1. Co-located Protector

In Figure 12, PE4 is a co-located protector that protects PW1 against egress AC failure and egress node failure. It maintains a label space for PE2, which is identified by the context identifier of {PE2, PE4}. It learns from PE2 the label that PE2 has assigned to PW1, and installs an forwarding entry for the label in the label space. The nexthop of the forwarding entry indicates a label pop with outgoing interface pointing to the backup AC CE2-PE4.

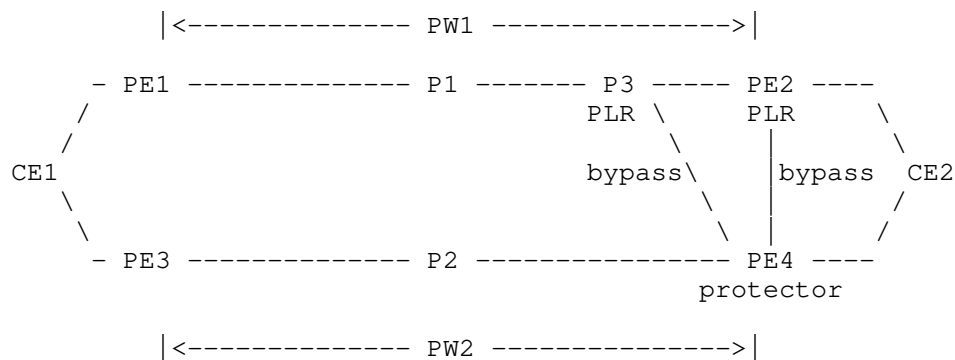


Figure 12

In Figure 13, SPE2 is a co-located protector that protects PW1 against switching node failure. It maintains a label space for SPE1, which is identified by the context identifier of {SPE1, SPE2}. It learns the label that SPE1 has assigned to the PW segment SEG1, and installs a forwarding entry in the label space. The nexthop of the forwarding entry indicates a label swap to the label of the PW segment SEG4.

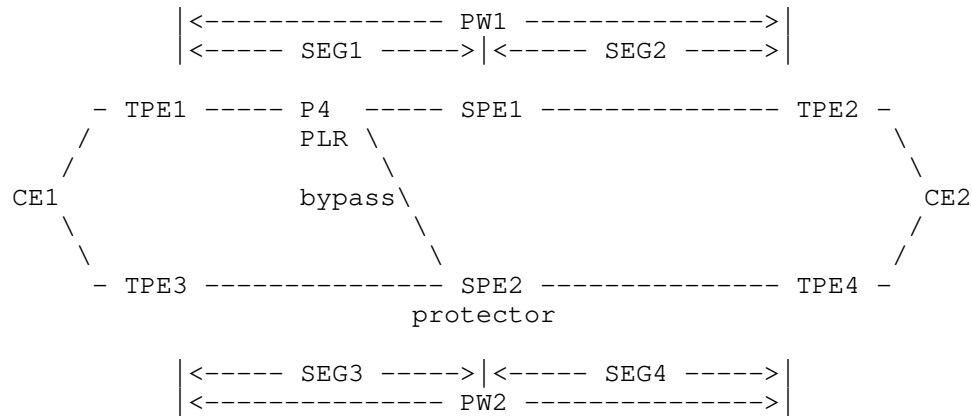


Figure 13

4.6.2. Centralized Protector

In the centralized protector model, for each primary PW of which the protector is not a backup (S-)PE, the protector MUST also learn the label of the backup PW from the backup (S-)PE (Section 4.8). This is the backup (S-)PE that the protector will forward traffic to. The protector MUST use the label as the outgoing label for the forwarding entry of the primary PW label.

In Figure 14, the protector is a centralized protector that protects PW1 against egress AC failure and egress node failure. It maintains a label space for PE2, which is identified by the context identifier of {PE2, protector}. It learns from PE2 the label that PE2 has assigned to PW1, and learns from PE4 the label that PE4 has assigned to PW2. It installs a forwarding entry for PW1's label in the label space. The nexthop of the forwarding entry indicates a label swap to PW2's label.

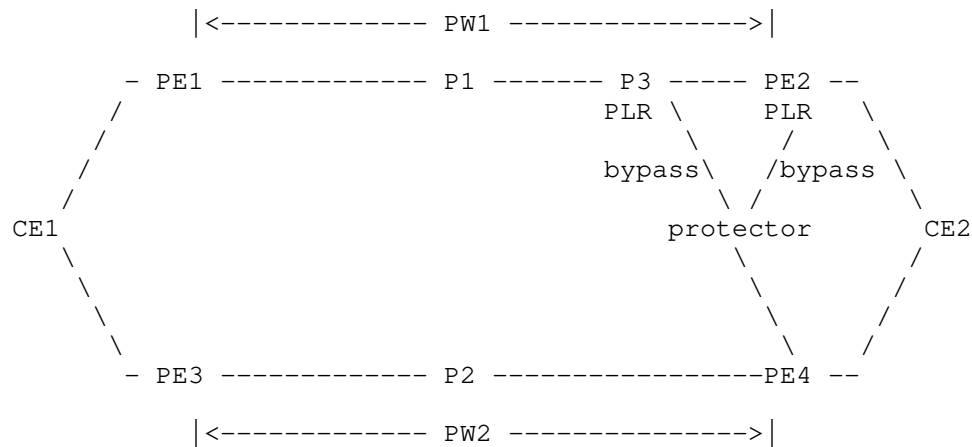


Figure 14

In Figure 15, the protector is a centralized protector that protects the PW segment SEG1 of PW1 against switching node failure of SPE1. It maintains a label space for SPE1, which is identified by the context identifier of {SPE1, protector}. It learns from SPE1 the label that SPE1 has assigned to SEG1, and learns from SPE2 the label that SPE2 has assigned to SEG3. It installs a forwarding entry for SEG1's label in the label space. The nexthop of the forwarding entry indicates a label swap to SEG3's label.

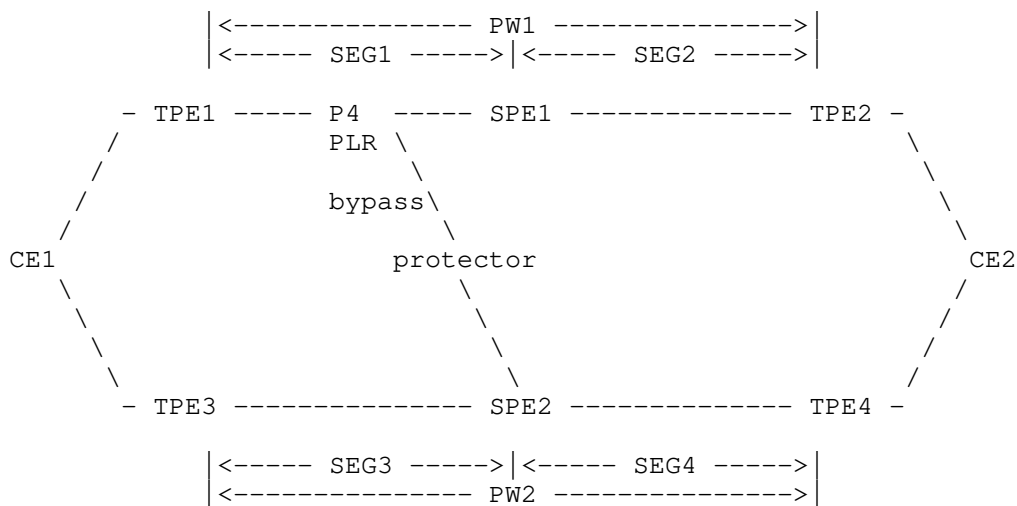


Figure 15

4.7. PW Label Distribution from Primary PE to Protector

A primary PE MUST distribute the label of each primary PW to the protector that protects the PW. To achieve this, the primary PE MUST establish a targeted LDP session with the protector. For each primary PW, the primary PE SHOULD advertise over that session a Protection FEC Element via Label Mapping message. The Protection FEC Element is a new LDP FEC, and its encoding is described below. The PW's label is encoded in the message using the Upstream-Assigned Label TLV defined in (RFC 6389). The Protection FEC Element and the PW label combined represent the primary PE's forwarding state for the PW. The Label Mapping message SHOULD also carry an IPv4/v6 Interface_ID TLV (RFC 6389, RFC 3471) encoded with the context identifier of the {primary PE, protector}.

The protector that receives this Label Mapping message SHOULD install a forwarding entry for the PW label in the label space identified by the context identifier. The nexthop of the forwarding entry SHOULD allow packets to be sent towards the target CE via a backup AC or a backup (S-)PE, depending on the protection model and SS-PW or MS-PW scenario involved.

The Protection FEC Element has type 0x83. It is defined as below:

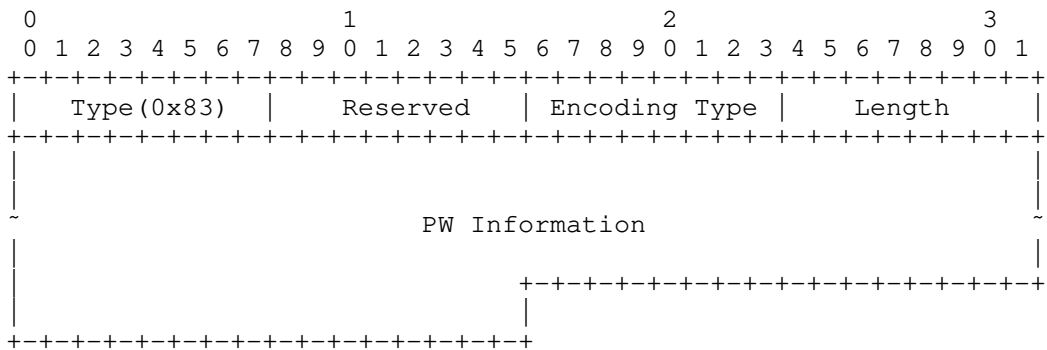


Figure 16

- Encoding Type
Type of format that PW Information field is encoded.
 - Length
Length of PW Information field in octets.
 - PW Information
Field of variable length that specifies a PW
- For Encoding Type, 1 is defined for the PWid FEC Element format, and 2 is defined for the Generalized PWid FEC Element format (RFC 4447).

4.7.1. Protection FEC Element Encoding for PWid

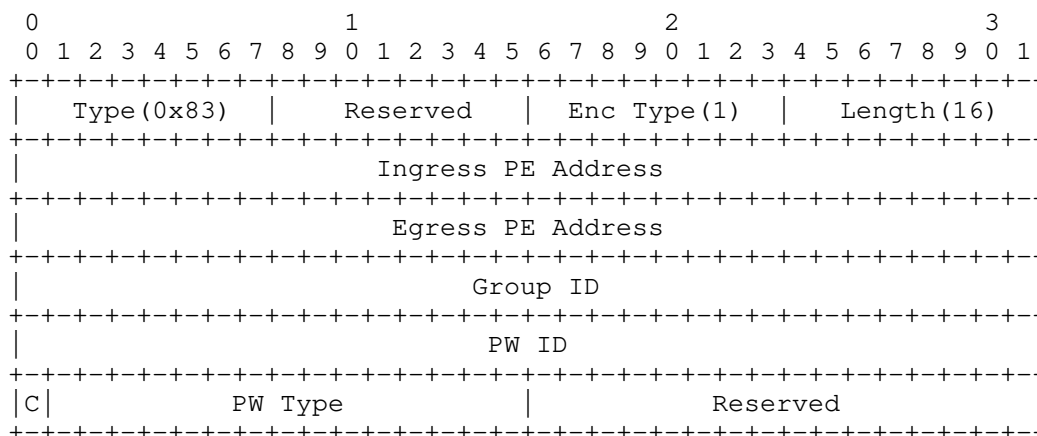


Figure 17

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Group ID

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

- PW ID

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

- Control word bit (C)

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

- PW Type

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

4.7.2. Protection FEC Element Encoding for Generalized PWid

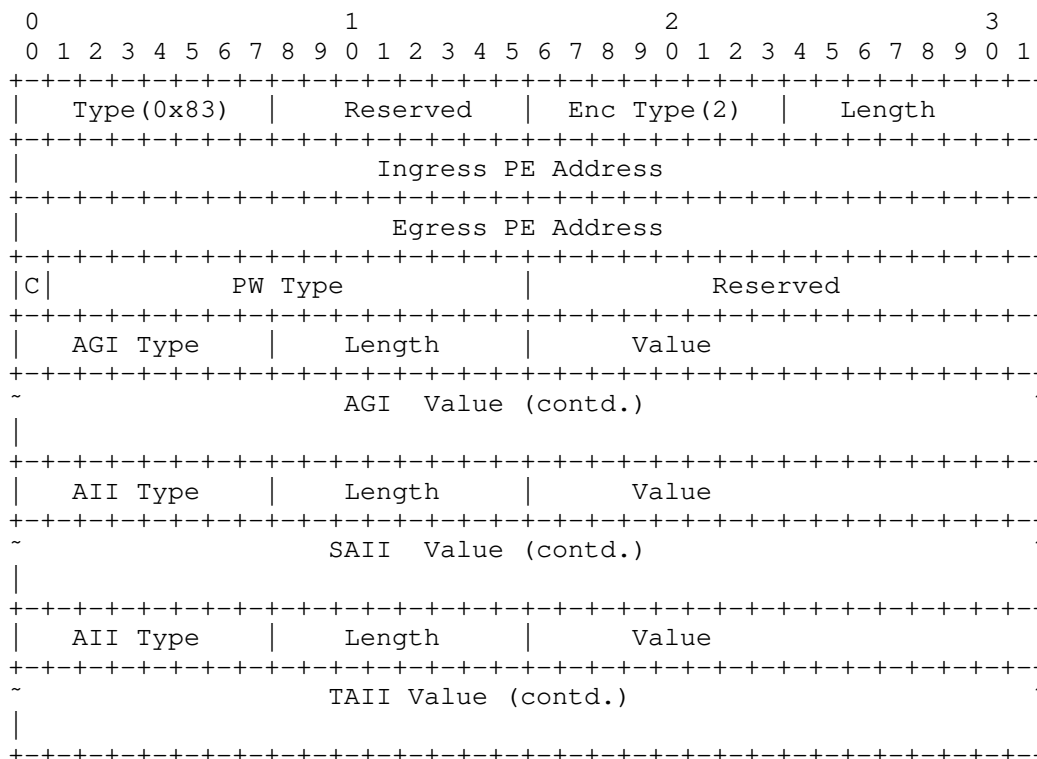


Figure 18

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Control word bit (C)

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

- PW Type

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

- AGI Type, Length, Value, AGI Value

Attachment Group Identifier of PW.

- SAII Type, Length, Value, SAII Value

Source Attachment Individual Identifier of PW.

- TAII Type, Length, Value, TAII Value

Target Attachment Individual Identifier of PW.

4.8. PW Label Distribution from Backup PE to Protector

In the centralized protector model, a protector may not be a backup (S-)PE for some primary PWs. For these PWs, in addition to learning PW labels from the primary PEs, the protector MUST also learn the labels of backup PWs and backup PW segments from backup (S-)PEs.

To achieve this, each backup (S-)PE MUST establish a targeted LDP session with the protector. The backup PE SHOULD advertise over that session a Protection FEC Element for the backup PW via Label Mapping message. The content of this Protection FEC Element MUST match the Protection FEC Element that the primary PE advertises to the protector (section 4.8). The Label Mapping message SHOULD also include a Generic Label TLV encoded with the backup PW's label. The context identifier SHOULD NOT be encoded in Interface_ID TLV in this message. The Protection FEC Element and the backup PW's label combined represent the backup PE's forwarding state for the backup PW.

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

4.9. Revertive Behavior

After a local repair takes effect, PW traffic is redirected from a PLR to a protector and then to target CE. There are three strategies for restoring the traffic to a fully working PW.

- o Global revertive mode

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

- o Control plane revertive mode

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

- o Local revertive mode

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

The fast protection mechanism in this document SHOULD always be used in tandem with the globally revertive mode. Particularly in the case of egress (S-)PE failure, if the ingress PE or the protector loses communication with the (S-)PE for an extensive period of time, the LDP session between them may go down. Consequently, the ingress PE may bring down the primary PW, or the protector may delete the forwarding entry of the primary PW label from the label space. In either case, the service will be disrupted. In other words, although the fast protection can temporarily repair traffic, control plane states may eventually time out if the failure persists. Therefore,

it is recommended that the global revertive mode SHOULD always be established in advance, so that it can move traffic to a fully working backup PW shortly after the local repair.

The control plane revertive mode is optional, because it only applies to the specific scenarios of egress PE failure and S-PE failure.

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

5. IANA Considerations

IANA maintains a registry of LDP FECs at the registry "Label Distribution Protocol" in the sub-registry called "Forwarding Equivalence Class (FEC) Type Name Space".

This document defines a new LDP Protection FEC Element in Section 4.7. IANA has assigned the type value 0x83 to it.

6. Security Considerations

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

7. Acknowledgements

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

8. References

8.1. Normative References

- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC5659] Bocci, M. and S. Bryant, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.

- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC6389] Aggarwal, R. and J.L. Le Roux, "MPLS Upstream Label Assignment for LDP", RFC 6389, November 2011.

[IP-LDP-FRR-MRT]

Atlas, A. and R. Kebler, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees",
draft-ietf-rtgwg-mrt-frr-architecture (work in progress),
2011.

8.2. Informative References

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

Authors' Addresses

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

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

Rahul Aggarwal
Arktan, Inc

Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.be

PWE3
Internet-Draft
Intended status: Informational
Expires: January 16, 2013

YJ. Stein
RAD Data Communications
D. Black
EMC Corporation
B. Briscoe
BT
July 15, 2012

PW Congestion Considerations
draft-stein-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 January 16, 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	9
5. IANA Considerations	9
6. Informative References	10
Appendix A. Loss Probabilities for TDM PWs	11
Authors' Addresses	12

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 shut down 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.

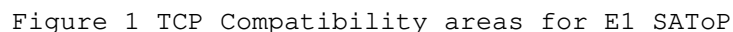
$$D < \frac{(TDM + 4)}{BW f(p) / 4}$$

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-

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.

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.

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 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.

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/>

