

Network Working Group  
Internet Draft  
Expiration Date: December 2013  
Intended status: Standards Track

Luca Martini (Ed.)  
Cisco Systems Inc.  
  
Matthew Bocci (Ed.)  
Florin Balus (Ed.)  
Alcatel-Lucent

June 19, 2013

## Dynamic Placement of Multi Segment Pseudowires

draft-ietf-pwe3-dynamic-ms-pw-17.txt

### Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/ietf/lid-abstracts.txt>.

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

This Internet-Draft will expire on December 19, 2013

### Abstract

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

## Table of Contents

1	Major Co-authors .....	3
2	Acknowledgements .....	3
3	Introduction .....	3
3.1	Scope .....	3
3.2	Specification of Requirements .....	3
3.3	Terminology .....	4
3.4	Architecture Overview .....	4
4	Applicability .....	6
4.1	Changes to Existing PW Signaling .....	6
5	PW Layer 2 Addressing .....	6
5.1	Attachment Circuit Addressing .....	6
5.2	S-PE Addressing .....	7
6	Dynamic Placement of MS-PWs .....	7
6.1	Pseudowire Routing Procedures .....	8
6.1.1	AI PW Routing Table Lookup Aggregation Rules .....	8
6.1.2	PW Static Route .....	9
6.1.3	Dynamic Advertisement with BGP .....	9
6.2	LDP Signaling .....	10
6.2.1	MS-PW Bandwidth Signaling .....	10
6.2.2	Equal Cost Multi Path (ECMP) in PW Routing .....	12
6.2.3	Active/Passive T-PE Election Procedure .....	12
6.2.4	Detailed Signaling Procedures .....	13
7	Failure Handling Procedures .....	14
7.1	PSN Failures .....	14
7.2	S-PE Specific Failures .....	14
7.3	PW Reachability Changes .....	15
8	Operations and Maintenance (OAM) .....	16
9	Security Considerations .....	16
10	IANA Considerations .....	16
10.1	LDP TLV TYPE NAME SPACE .....	17
10.2	LDP Status Codes .....	17
10.3	BGP SAFI .....	17
11	Normative References .....	17
12	Informative References .....	18
13	Author's Addresses .....	18

## 1. Major Co-authors

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

## 2. Acknowledgements

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

## 3. Introduction

### 3.1. Scope

[RFC5254] describes the service provider requirements for extending the reach of pseudowires across multiple PSN domains. This is achieved using a Multi-segment Pseudowire (MS-PW). An MS-PW is defined as a set of two or more contiguous PW segments that behave and function as a single point-to-point PW. This architecture is described in [RFC5659].

The procedures for establishing PWs that extend across a single PSN domain are described in [RFC4447], while procedures for setting up PWs across multiple PSN domains, or control plane domains are described in [RFC6073].

The purpose of this document is to specify extensions to the pseudowire control protocol [RFC4447], and [RFC6073] procedures, to enable multi-segment PWs to be dynamically placed. The proposed procedures follow the guidelines defined in [RFC5036] and enable the reuse of existing TLVs, and procedures defined for SS-PWs in [RFC4447].

### 3.2. Specification of Requirements

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

### 3.3. Terminology

[RFC5659] provides terminology for multi-segment pseudowires.

This document defines the following additional terms:

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

### 3.4. Architecture Overview

The following figure describes the reference models which are derived from [RFC5659] to support PW emulated services across multi-segment PWs.

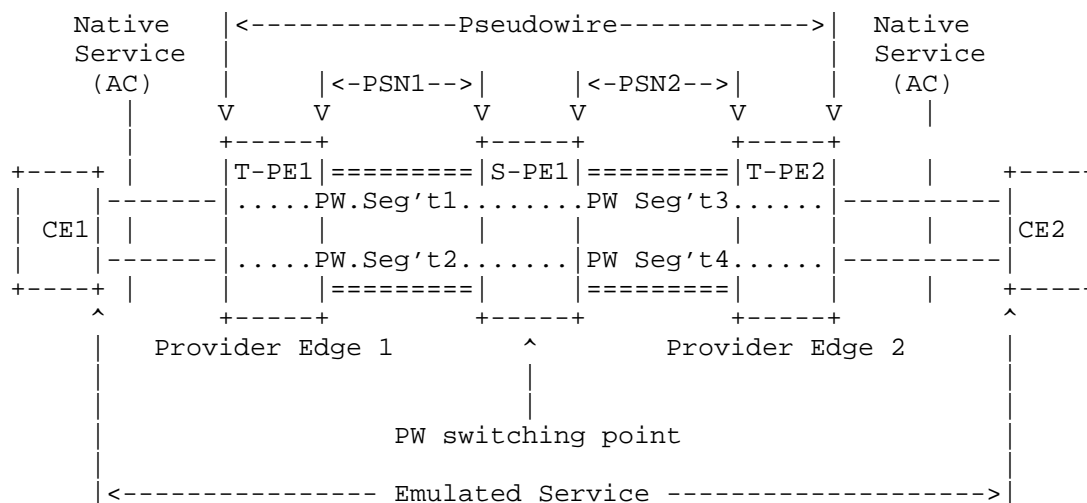


Figure 1: PW switching Reference Model

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

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

#### 4. Applicability

In this document we describe the case where the PSNs carrying the SS-PW are only MPLS PSNs using the generalized FEC 129. Interactions with an IP PSN using L2TPv3 as described in [RFC6073] section 7.4 are left for further study.

##### 4.1. Changes to Existing PW Signaling

The procedures described in this document make use of existing LDP TLVs and related PW signaling procedures described in [RFC4447] and [RFC6073]. The following OPTIONAL TLVs are also defined:

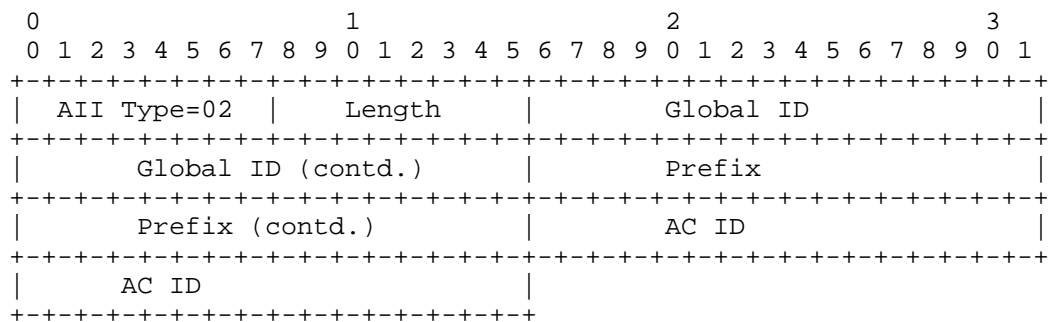
- A Bandwidth TLV to address QoS Signaling requirements (see "MS-PW Next Hop Bandwidth Signaling" section for details).

#### 5. PW Layer 2 Addressing

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

##### 5.1. Attachment Circuit Addressing

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



AII type 2 based addressing schemes permit varying levels of AII

summarization, thus reducing the scaling burden on PW routing. AII Type 2 based PW addressing is suitable for point-to-point provisioning models where it is not required to auto-discover address at Target T-PE (knows the address in priori by provisioning).

Implementations of the following procedure MUST interpret the AII type to determine the meaning of the address format of the AII, irrespective of the number of segments in the MS-PW. All segments of the PW MUST be signaled with same AII Type.

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

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

## 5.2. S-PE Addressing

Each S-PE MUST be uniquely addressable in terms of pseudowires in order to populate the switching point PE TLV specified in [RFC6073]. For this purpose, at least one AI address of the format similar to AII type 2 [RFC5003] composed of the Global ID, and Prefix part, only, MUST be assigned to each S-PE.

If an S-PE is capable of Dynamic MS-PW signaling, but is not assigned with an S-PE address, then on receiving a Dynamic MS-PW label mapping message the S-PE MUST return a Label Release with the "LDP\_RESOURCES\_UNAVAILABLE" ( 0x38)" status code.

## 6. Dynamic Placement of MS-PWs

[RFC6073] describes a procedure for concatenating multiple pseudowires together. This procedure requires each S-PE to be manually configured with the information required for each segment of the MS-PW. The procedures in the following sections describe a method to extend [RFC6073] by allowing the automatic selection of pre-defined S-PEs, and dynamically establishing a MS-PW between two T-PEs.

## 6.1. Pseudowire Routing Procedures

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

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

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

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

### 6.1.1. AII PW Routing Table Lookup Aggregation Rules

All PEs capable of dynamic MS-PW path selection MUST build a PW AII routing table to be used for PW next-hop selection.

The PW addressing scheme (AII type 2 in [RFC5003]) consists of a Global ID, a 32 bit prefix and a 32 bit Attachment Circuit ID.

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

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



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

#### 6.1.2. PW Static Route

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

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

#### 6.1.3. Dynamic Advertisement with BGP

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

Contrary to other l2vpn signaling methods that use BGP [RFC6074], in the case of the dynamically placed MS-PW, the source T-PE knows a-priori (by provisioning) the AC ID on the terminating T-PE to use in signaling. Hence there is no need to advertise a "fully qualified" 96 bit address on a per PW Attachment Circuit basis. Only the T-PE Global ID, Prefix, and prefix length needs to be advertised as part of well known BGP procedures - see [RFC4760].

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

The PW dynamic path NLRI is advertised in BGP UPDATE messages using the MP\_REACH\_NLRI and MP\_UNREACH\_NLRI attributes [RFC4760]. The [AFI, SAFI] value pair used to identify this NLRI is (AFI=25, SAFI=6 (pending IANA allocation)). A route target MAY also be advertised along with the NLRI.



PW QoS objectives can thus be met where the next hop for a PW segment is explicitly configured at each PE, whether the PE is a T-PE or an S-PE in the case of a segmented PW without dynamic path selection (as per RFC6073). In these cases, it is possible to explicitly configure the bandwidth required for a PW so that the T-PE or S-PE can reserve that bandwidth on the PSN tunnel.

Where dynamic path selection is used and therefore the next-hop is not explicitly configured by the operator at the S-PE, a mechanism is required to signal the bandwidth for the PW from the T-PE to the S-PEs. This is accomplished by including an OPTIONAL PW Bandwidth TLV. The PW Bandwidth TLV is specified as follows:

```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 0 |           PW BW   TLV   (0x096E)   |           TLV   Length   |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Forward SENDER_TSPEC                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Reverse SENDER_TSPEC                  |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

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

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

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

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

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

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

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

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

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

#### 6.2.2. Equal Cost Multi Path (ECMP) in PW Routing

A specific PW may find match with a PW route that may have multiple next-hops associated with it. Multiple next-hops may be either configured explicitly as static routes or may be learned through BGP routing procedures. Implementations at and S-PE or T-PE MAY use selection algorithms, such as CRC32 on the FEC TLV, for load balancing of PWs across multiple next-hops. The details of such selection algorithms are outside the scope of this document.

#### 6.2.3. Active/Passive T-PE Election Procedure

When a MS-PW is signaled, each T-PE might independently initiate signaling the MS-PW. This could result in a different path being used be each direction of the PW. To avoid this situation one of the T-PE MUST start the PW signaling (active role), while the other T-PE waits to receive the LDP label mapping message before sending the LDP label

mapping message for the reverse direction of the PW (passive role). The Active T-PE (the ST-PE) and the passive T-PE (the TT-PE) MUST be identified before signaling begins for a given MS-PW.

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

#### 6.2.4. Detailed Signaling Procedures

On receiving a label mapping message, the S-PE MUST inspect the FEC TLV. If the receiving node has no local AII matching the TAI for that label mapping then the finagling should be forwarded on to another S-PE or T-PE. The S-PE will check if the FEC is already installed for the forward direction:

- If it is already installed, and the received mapping was received from the same LDP peer where the forward LDP label mapping was sent, then this label mapping represents signaling in the reverse direction for this MS-PW segment.
- If it is already installed, and the received mapping was received from a different LDP peer where the forward LDP label mapping was sent, then the received label mapping MUST be released with status code as "PW\_LOOP\_DETECTED". If the already installed PW segment has not signaled explicit intent for active role then installed PW segment MUST be released with status code "PW\_LOOP\_DETECTED".
- If the FEC is not already installed, then this represents signaling in the forward direction.

For the forward direction:

- i. Determine the next hop S-PE or T-PE according to the procedures above. If next-hop reachability is not found in the PW AII routing table in the S-PE then label release MUST be sent with status code "AII\_UNREACHABLE". If the next-hop S-PE or T-PE is found and is the same LDP Peer that has sent the label mapping message then a label Release MUST be returned with the status code "PW\_LOOP\_DETECTED". If the SAI in the received label mapping is local to the S-PE then a label released MUST be returned with status code "PW\_LOOP\_DETECTED".
- ii. Check that a PSN tunnel exists to the next hop S-PE or T-PE. If no tunnel exists to the next hop S-PE or T-PE the S-PE MAY attempt to setup a PSN tunnel.

- iii. Check that a PSN tunnel exists to the previous hop. If no tunnel exists to the previous hop S-PE or T-PE the S-PE MAY attempt to setup a PSN tunnel.
- iv. If the S-PE cannot get enough PSN resources to setup the segment to the next or previous S-PE or T-PE, a label release MUST be returned to the T-PE with a status message of "Resources Unavailable".
- v. If the label mapping message contains a Bandwidth TLV, allocate the required resources on the PSN tunnels in the forward and reverse directions according to the procedures above.
- vi. Allocate a new PW label for the forward direction.
- vii. Install the FEC for the forward direction.
- viii. Send the label mapping message with the new forward label and the FEC to the next hop S-PE/T-PE.

For the reverse direction:

- i. Install the received FEC for the reverse direction.
- ii. Determine the next signaling hop by referencing the LDP sessions used to setup the PW in the Forward direction.
- iii. Allocate a new PW label for the reverse direction.
- iv. Install the FEC for the reverse direction.
- v. Send the label mapping message with a new label and the FEC to the next hop S-PE/T-PE.

## 7. Failure Handling Procedures

### 7.1. PSN Failures

Failures of the PSN tunnel MUST be handled by PSN mechanisms. If the PSN is unable to re-establish the PSN tunnel, then the S-PE SHOULD follow the procedures defined in Section 8 of [RFC6073].

### 7.2. S-PE Specific Failures

For defects in an S-PE, the procedures defined in [RFC6073] SHOULD be followed. A T-PE or S-PE may receive an unsolicited label release message from another S-PE or T-PE with various failure codes such "LOOP\_DETECTED", "PW\_LOOP\_DETECTED", "RESOURCE\_UNAVAILABLE", "BAD\_STRICT\_HOP", "ALL\_UNREACHABLE" etc. All these failure codes indicate a generic class of PW failures at an S-PE or T-PE.

When an unsolicited label release message with such a failure status code is received at T-PE then the T-PE MUST re-attempt to establish the PW immediately. However the T-PE MUST throttle its PW setup message retry attempts with an exponential backoff in situations

where PW setup messages are being constantly released. It is also recommended that a T-PE detecting such a situation take action to notify an operator.

S-PEs that receive an unsolicited label release message with a failure status code should follow the following procedures:

- i. If label release is received from an S-PE or T-PE in the forward signaling direction then S-PE MUST tear down both segments of the PW. The status code received in label release SHOULD be propagated while sending label release for the next-segment.
- ii. If the label release is received from an S-PE or T-PE in the reverse Signaling direction do as follows:

If the PW is set-up at S-PE with an Explicit Intent of Role then label release MUST be sent to the next PW segment with same status code. The forward signaling path SHOULD NOT be tear down in such case.

If the PW is set-up at S-PE without an Explicit Intent of Role then tear down both segments of the PW as described in i.

### 7.3. PW Reachability Changes

In general an established MS-PW will not be affected by next-hop changes in L2 PW reachability information.

If there is a change in next-hop of the L2 PW reachability information in the forward direction, the T-PE MAY elect to tear down the MS-PW by sending a label withdraw message to downstream S-PE or T-PE. The teardown MUST be also accompanied by a unsolicited label release message, and will be followed by and attempt to re-establish of the MS-PW by T-PE.

If there is a change in the L2 PW reachability information in the forward direction at S-PE, the S-PE MAY elect to tear down the MS-PW in both directions. A label withdrawal is sent on each direction followed by a unsolicited label release. The unsolicited label releases MUST be accompanied by the Status code "AII\_UNREACHABLE". This procedure is OPTIONAL.

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

## 8. Operations and Maintenance (OAM)

The OAM procedures defined in [RFC6073] may be used also for MS-PWs. A PW switching point TLV is used [RFC6073] to record the switching points that the PW traverses.

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

Type	Length	Description
0x06	14	L2 PW address of PW Switching Point

The above field MUST be included together with type 0x02 in the TLV once per individual PW Switching Point following the same rules and procedures as described in [RFC6073]. A more detailed description of this field is also in setion 7.4.1 of [RFC6073]

## 9. Security Considerations

This document specifies only extensions to the protocols already defined in [RFC4447], and [RFC6073]. Each such protocol may have its own set of security issues, but those issues are not affected by the extensions specified herein. Note that the protocols for dynamically distributing PW Layer 2 reachability information may have their own security issues, however those protocols specifications are outside the scope of this document.

## 10. IANA Considerations

IANA needs to correct a minor error in the registry "Pseudowire Switching Point PE sub-TLV Type". The entry 0x06 "L2 PW address of the PW Switching Point" should have Length 14.



#### 10.1. LDP TLV TYPE NAME SPACE

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

TLV type	Description
0x096E	Bandwidth TLV

#### 10.2. LDP Status Codes

This document uses several new LDP status codes, IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by RFC5036. The following values have been pre-allocated:

Range/Value	E	Description	Reference
0x00000037	0	Bandwidth resources unavailable	RFCxxxx
0x00000038	0	Resources Unavailable	RFCxxxx
0x00000039	0	All Unreachable	RFCxxxx

#### 10.3. BGP SAFI

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

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

#### 11. Normative References

- [RFC6073] Martini et.al. "Segmented Pseudowire", RFC6073, January 2011
- [TSPEC] Wroclawski, J. "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997
- [RFC5036] Andersson, Minei, Thomas. "LDP Specification" RFC5036, October 2007

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

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

## 12. Informative References

[RFC5254] Martini et al, "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3)", RFC5254, Bitar, Martini, Bocci, October 2008

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

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

[RFC6074] E. Rosen, W. Luo, B. Davie, V. Radoaca, "Provisioning, Autodiscovery, and Signaling in L2VPNs", rfc6074, January 2011

## 13. Author's Addresses

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

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

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

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

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

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

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

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

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

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

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

Jeff Sugimoto  
Alcatel-Lucent  
701 E. Middlefield Rd.  
Mountain View, CA 94043  
e-mail: jeffery.sugimoto@alcatel-lucent.com

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

Mike Loomis  
Alcatel-Lucent  
701 E. Middlefield Rd.  
Mountain View, CA 94043  
e-mail: mike.loomis@alcatel-lucent.com

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

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

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

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

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

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

#### 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Expiration Date: December 2013

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 17, 2014

R. Schatzmayr  
Deutsche Telekom AG  
G. Heron, Ed.  
M. Konstantynowicz, Ed.  
M. Townsley  
Cisco Systems  
July 16, 2013

Keyed IPv6 Tunnel  
draft-mkonstan-keyed-ipv6-tunnel-00

Abstract

This document describes a simple L2 Ethernet over IPv6 tunnel encapsulation with mandatory 64-bit authentication key for connecting L2 Ethernet attachment circuits identified by IPv6 addresses. The encapsulation is based on L2TPv3 over IP.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [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 January 17, 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. Static 1:1 Mapping Without a Control Plane . . . . .	2
3. 64-bit Authentication Key . . . . .	3
4. Encapsulation . . . . .	3
5. IANA Considerations . . . . .	6
6. Security Considerations . . . . .	6
7. Acknowledgements . . . . .	7
8. References . . . . .	7
8.1. Normative References . . . . .	7
8.2. Informative References . . . . .	7
Authors' Addresses . . . . .	7

## 1. Introduction

L2TPv3, as defined in RFC3931 [RFC3931], provides a dynamic mechanism for tunneling Layer 2 (L2) "circuits" across a packet-oriented data network (e.g., over IP), with multiple attachment circuits multiplexed over a single pair of IP address endpoints (i.e. a tunnel) using the L2TPv3 session ID as a circuit discriminator.

Implementing L2TPv3 over IPv6 provides the opportunity to utilize unique IPv6 addresses to identify Ethernet attachment circuits directly, leveraging the key property that IPv6 offers, a vast number of unique IP addresses. In this case, processing of the L2TPv3 Session ID may be bypassed upon receipt as each tunnel has one and only one associated session. This local optimization does not hinder the ability to continue supporting the multiplexing of circuits via the Session ID on the same router for other L2TPv3 tunnels.

## 2. Static 1:1 Mapping Without a Control Plane

Use of the L2TPv3 Control Plane is optional. When the control plane is not used, local configuration creates a one-to-one mapping between the access-side L2 attachment circuit and the IP address used in the network-side IPv6 encapsulation. Further, circuit monitoring is performed using Ethernet OAM mechanisms (802.1ag and/or Y.1731).



The L2TPv3 encapsulating router identifies each Ethernet L2 attachment circuit by the Ethernet VLAN stack present on Ethernet frames on the access side

- o port mode access - physical port identifies a L2 attachment circuit.
- o single-stack access - s-tag or c-tag with S-VID or C-VID value identifies a L2 attachment circuit.
- o multi-stack access - ( s-tag, c-tag ) with tuple ( S-VID, C-VID ) identifies a L2 attachment circuit.

L2 attachment connection identifiers s-tag or ( s-tag, c-tag ) are treated with local significance and are not required to be forwarded over the IPv6 network (though the operator may prefer to forward tags in some cases).

The L2TPv3 encapsulating router identifies each L2TPv3 tunnel endpoint by a distinct /128 IPv6 address in the packet header of L2TPv3 IPv6 packets received and transmitted on the network side.

In the event that an IPv6 address used in L2TPv3 does not directly correspond to one and only one attachment circuit on both sides of the L2TPv3 tunnel, the Session ID may be used for additional granularity. This allows for other addressing schemes that may require additional bits beyond those which can fit in the IPv6 header address field.

### 3. 64-bit Authentication Key

All packets MUST carry a 64-bit authentication key in the L2TPv3 cookie field. The cookie MUST be 64-bits long in order to provide sufficient protection against a brute force blind insertion attack.

In absence of the L2TPv3 Control Plane, the L2TPv3 encapsulating router must be provided with local configuration of the 64-bit authentication cookie for each local and remote IPv6 endpoint - note that cookies are asymmetric, so local and remote endpoints may send different cookie values. The value of the cookie must be able to be changed at any time in a manner that does not drop any legitimate tunneled packets - i.e. the receiver must be willing to accept both "old" and "new" cookie values during a change of cookie value.

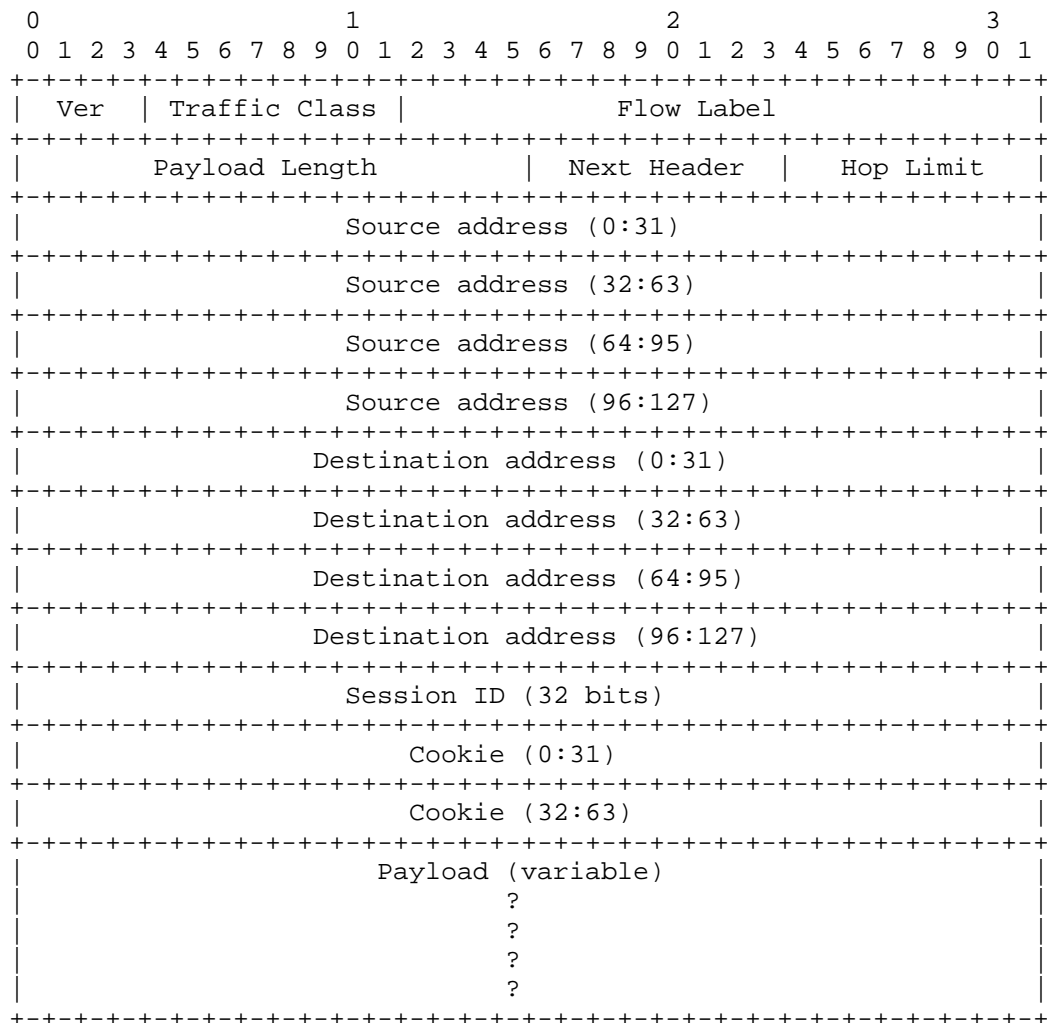
### 4. Encapsulation

RFC4719 [RFC4719] describes encapsulation of Ethernet over L2TPv3. Paraphrasing from this document, the Ethernet frame, without the

preamble or frame check sequence (FCS), is encapsulated in L2TPv3 and is sent as a single packet by the ingress router.

The s-tag (or in the multi-stack access case the s-tag and c-tag) SHOULD be removed before the packet is encapsulated.

The full encapsulation is as follows:



The combined IPv6 and L2TPv3 header contains the following fields:

- o Ver. Set to 0x6 to indicate IPv6.
- o Traffic Class. May be set by the ingress router to ensure correct PHB treatment by transit routers between the ingress and egress, and correct QoS disposition at the egress router.
- o Flow Label. May be set by the ingress router to indicate a flow of packets from the client which may not be reordered by the network (if there is a requirement for finer grained ECMP load balancing than per-circuit load balancing).
- o Payload Length. Set to the length of the packet, excluding the IPv6 header (i.e. the length from the Session ID to the end of the packet).
- o Next Header. Set to 0x73 to indicate that the next header is L2TPv3.
- o Hop Limit. Set to 0xFF, and decremented by one by each router in the path to the egress router.
- o Source Address. IPv6 source address for the tunnel. In the "Static 1:1" case the IPv6 source address may correspond to a port or VLAN being transported as an L2 circuit, or may be a loopback address terminating inside the router (e.g. if L2 circuits are being used within a multipoint VPN) or may be an anycast address terminating on a data center virtual machine.
- o Destination Address. IPv6 destination address for the tunnel. As with the source address this may correspond to a port or VLAN being transported as an L2 circuit or may be a loopback or anycast address.
- o Session ID. In the "Static 1:1 mapping" case described in Section 2, the IPv6 address resolves to an L2TPv3 session immediately, thus the Session ID may be ignored upon receipt. For compatibility with other tunnel termination platforms supporting only 2-stage resolution (IPv6 Address + Session ID), this specification recommends supporting explicit configuration of Session ID to any value other than zero. For cases where both tunnel endpoints support one-stage resolution (IPv6 Address only), this specification recommends setting the Session ID to all ones for easy identification in case of troubleshooting.
- o Cookie. 64 bits, configured and described as in Section 3. All packets for a destined L2 Circuit (or L2TPv3 Session) must match the configured Cookie value or be discarded (see RFC3931 [RFC3931] for more details).

- o Payload. The customer data, with s-tag or s-tag/c-tag removed. As noted above preamble and FCS are stripped before encapsulation. A new FCS will be added at each hop when the IP packet is transmitted.

## 5. IANA Considerations

None.

## 6. Security Considerations

Packet spoofing for any type of Virtual Private Network (VPN) tunneling protocol is of particular concern as insertion of carefully constructed rogue packets into the VPN transit network could result in a violation of VPN traffic separation, leaking data into a customer VPN. This is complicated by the fact that it may be particularly difficult for the operator of the VPN to even be aware that it has become a point of transit into or between customer VPNs.

Keyed IPv6 encapsulation provides traffic separation for its VPNs via use of separate 128-bit IPv6 addresses to identify the endpoints. The mandatory authentication key carried in the L2TPv3 cookie field, provides an additional check to ensure that an arriving packet is intended for the identified tunnel.

In the presence of a blind packet spoofing attack, the authentication key provides security against inadvertent leaking of frames into a customer VPN, like in case of L2TPv3 RFC3931 [RFC3931]. To illustrate the type of security that it is provided in this case, consider comparing the validation of a 64-bit Cookie in the L2TPv3 header to the admission of packets that match a given source and destination IP address pair. Both the source and destination IP address pair validation and Cookie validation consist of a fast check on cleartext header information on all arriving packets. However, since L2TPv3 uses its own value, it removes the requirement for one to maintain a list of (potentially several) permitted or denied IP addresses, and moreover, to guard knowledge of the permitted IP addresses from hackers who may obtain and spoof them. Further, it is far easier to change a compromised L2TPv3 Cookie than a compromised IP address," and a cryptographically random RFC4086 [RFC4086] value is far less likely to be discovered by brute-force attacks compared to an IP address.

For protection against brute-force, blind, insertion attacks, a 64-bit Cookie MUST be used with all tunnels.

Note that the Cookie provides no protection against a sophisticated man-in-the-middle attacker who can sniff and correlate captured data between nodes for use in a coordinated attack.

The L2TPv3 64-bit cookie must not be regarded as a substitute for security such as that provided by IPsec when operating over an open or untrusted network where packets may be sniffed, decoded, and correlated for use in a coordinated attack.

## 7. Acknowledgements

...

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.
- [RFC4086] Eastlake, D., Schiller, J., and S. Crocker, "Randomness Requirements for Security", BCP 106, RFC 4086, June 2005.
- [RFC4719] Aggarwal, R., Townsley, M., and M. Dos Santos, "Transport of Ethernet Frames over Layer 2 Tunneling Protocol Version 3 (L2TPv3)", RFC 4719, November 2006.

### 8.2. Informative References

- [RFC1700] Reynolds, J. and J. Postel, "Assigned Numbers", RFC 1700, October 1994.

## Authors' Addresses

Rainer Schatzmayr  
Deutsche Telekom AG

Email: rainer.schatzmayr@telekom.de

Giles Heron (editor)  
Cisco Systems

Email: giheron@cisco.com

Maciek Konstantynowicz (editor)  
Cisco Systems

Email: [maciek@cisco.com](mailto:maciek@cisco.com)

Mark Townsley  
Cisco Systems

Email: [townsley@cisco.com](mailto:townsley@cisco.com)

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

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

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

Abstract

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

Status of This Memo

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

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

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

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

Copyright Notice

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

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

## Table of Contents

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

## 1. Introduction



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

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

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

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

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

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

## 2. Specification of Requirements

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

## 3. Reference Models and Failure Cases

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

### 3.1. Single-Segment PW

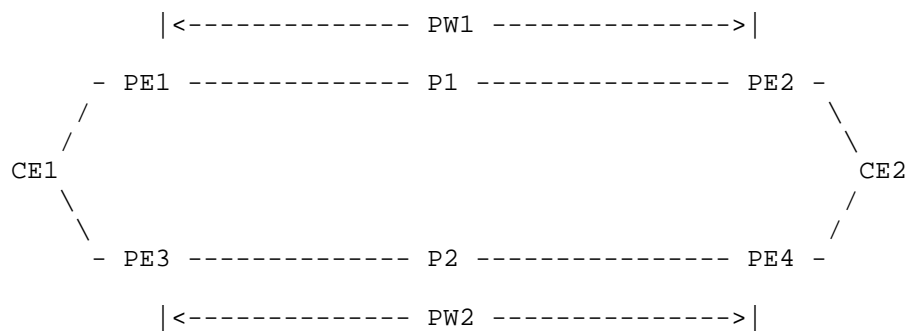


Figure 1

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

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

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

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

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

Primary PW: PW1

Primary PE: PE2

Primary AC: CE2-PE2

Backup PW: PW2

Backup PE: PE4

Backup AC: CE2-PE4

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

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

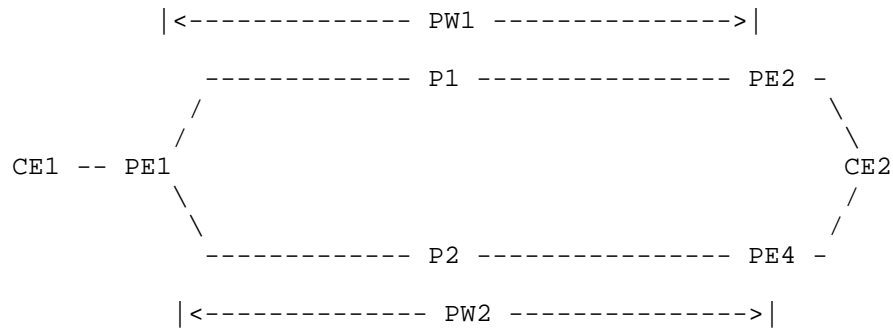


Figure 2

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

### 3.2. Multi-Segment PW

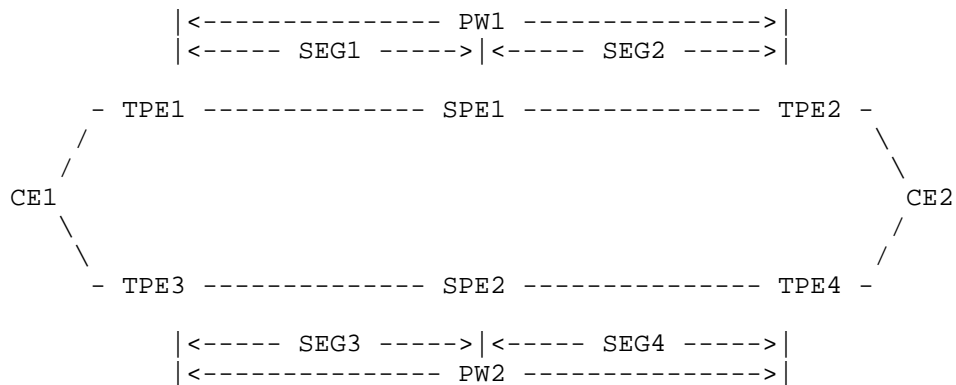


Figure 3

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

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

Primary PW: PW1

Primary T-PE: TPE2

Primary S-PE: SPE1

Primary AC: CE2-TPE2

Backup PW: PW2

Backup T-PE: TPE4

Backup S-PE: SPE2

Backup AC: CE2-TPE4

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

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

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

#### 4. Theory of Operation

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

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

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

#### 4.1. Local Repair and Protector

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

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

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

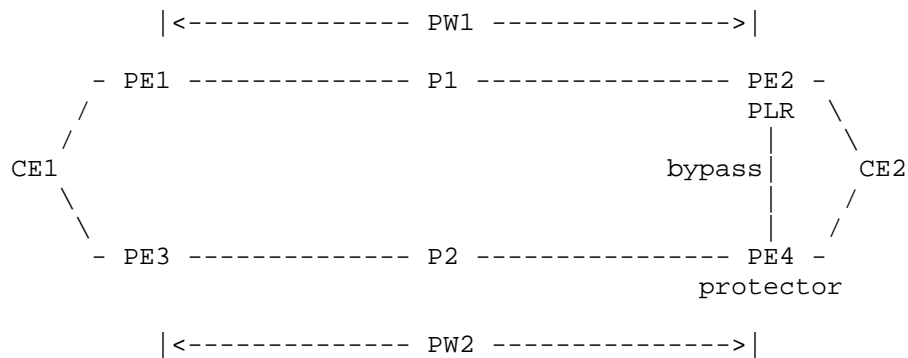


Figure 4

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

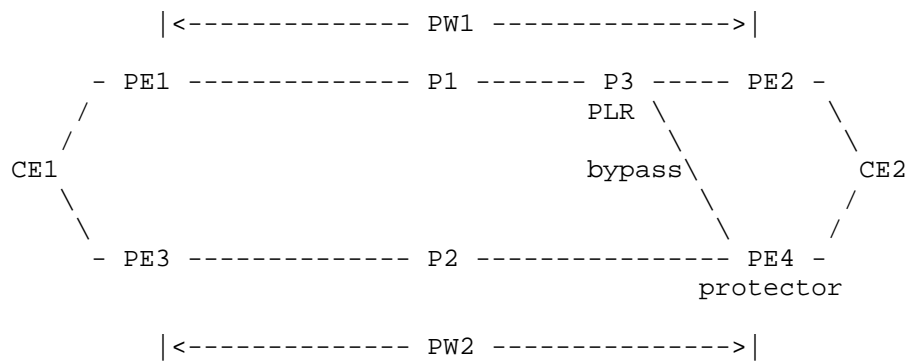


Figure 5

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

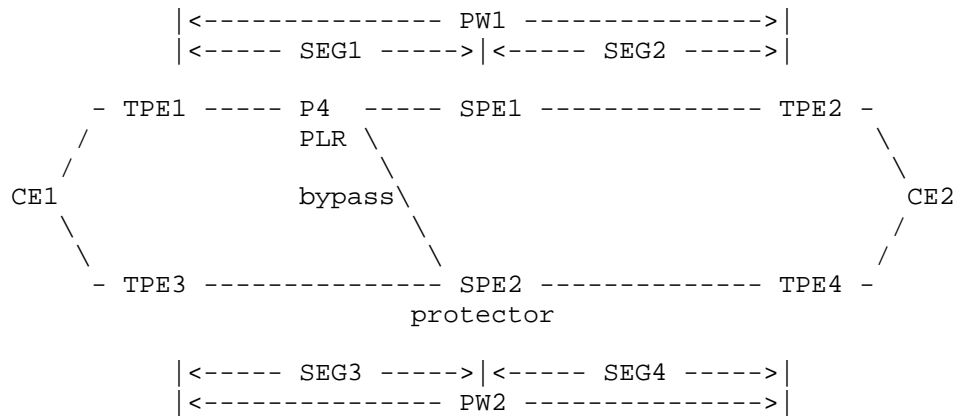


Figure 6

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

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

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

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

#### 4.2. Context Identifier

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

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

##### 4.2.1. Semantics

The semantics of a context identifier is twofold.

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



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

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

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

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

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

#### 4.2.2. Advertisement and Path Computation

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

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



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

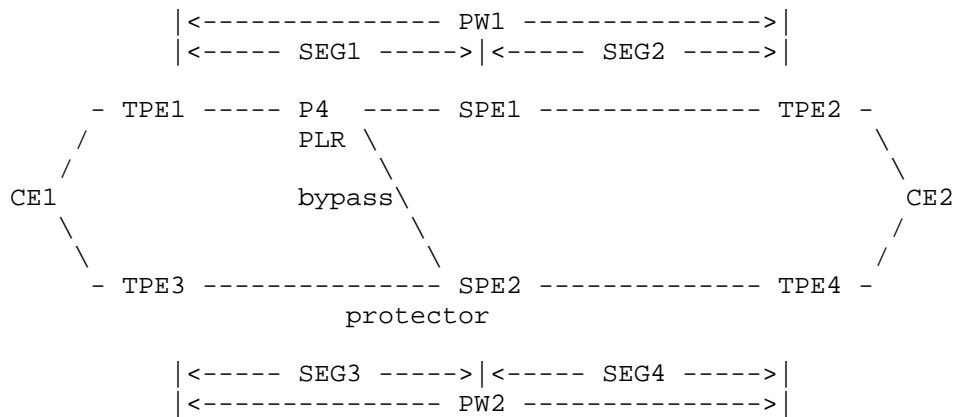


Figure 8

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

#### 4.3.2. Centralized Protector

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

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

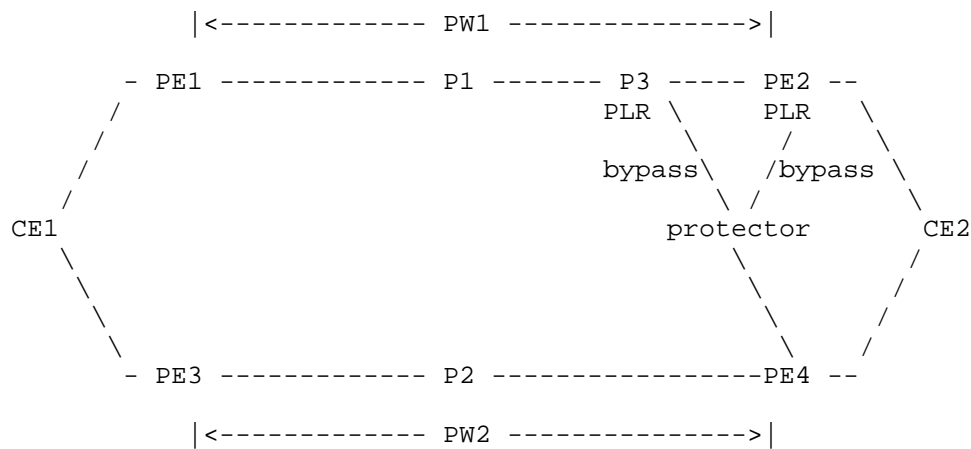


Figure 9

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

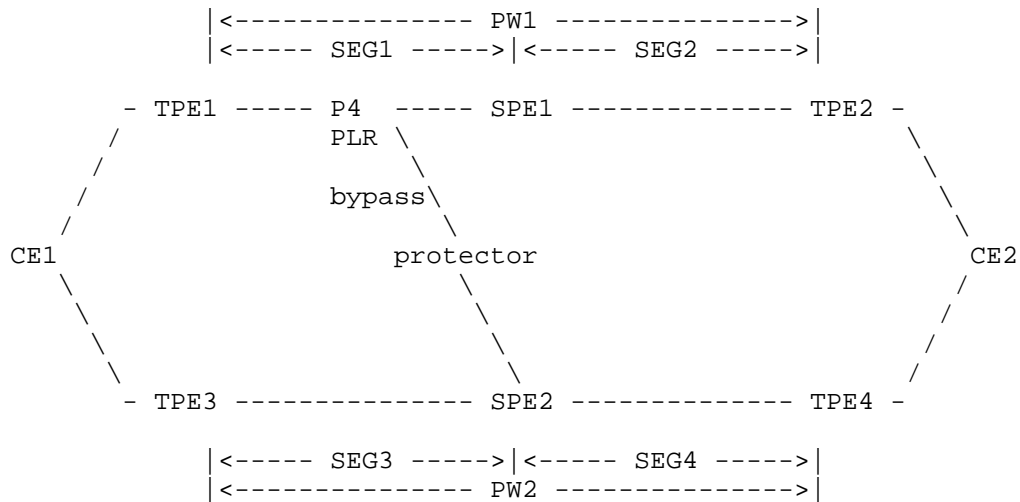


Figure 10

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

#### 4.4. Transport Tunnel

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

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

#### 4.5. Bypass Tunnel

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

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

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

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

#### 4.6. Forwarding State on Protector

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

##### 4.6.1. Examples of Co-located Protector

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

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

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

#### 4.6.2. Examples of Centralized Protector

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

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

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

### 5. LDP Extensions

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

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

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

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

#### 5.1. Egress Protection Capability TLV

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

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

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

```

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

```



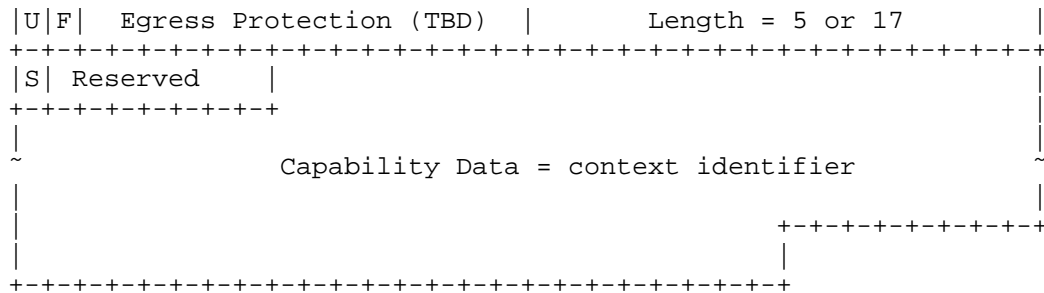


Figure 11

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

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

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

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

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

## 5.2. PW Label Distribution from Primary PE to Protector

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

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

### 5.3. PW Label Distribution from Backup PE to Protector

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

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

### 5.4. Protection FEC Element TLV

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



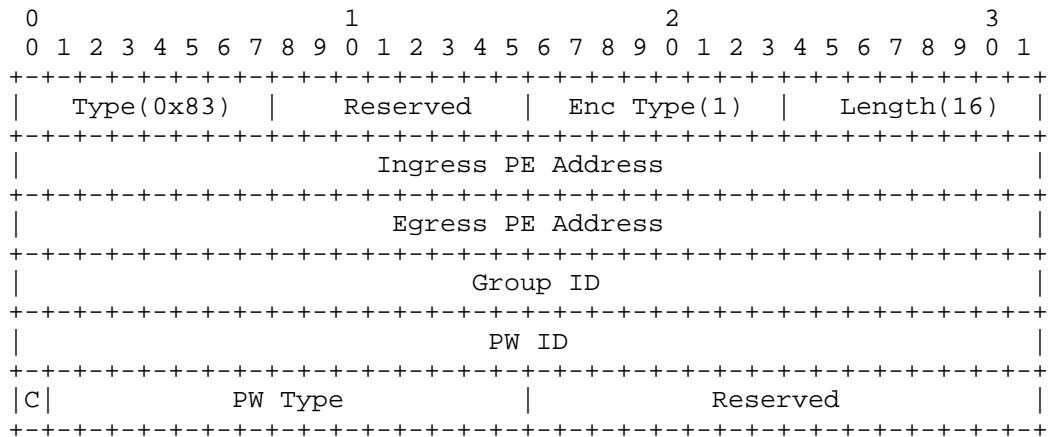


Figure 13

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Group ID

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

- PW ID

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

- Control word bit (C)

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

- PW Type

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

#### 5.4.2. Encoding Format for Generalized PWid

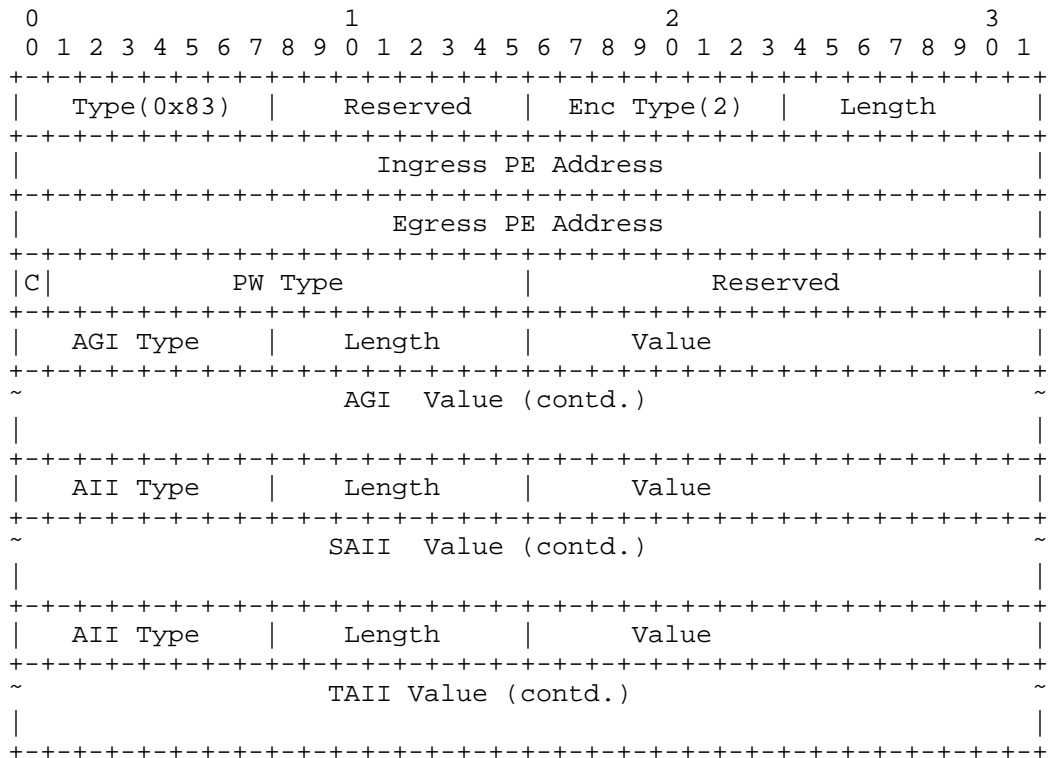


Figure 14

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Control word bit (C)

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

- PW Type

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

- AGI Type, Length, Value, AGI Value

Attachment Group Identifier of PW.

- SAII Type, Length, Value, SAII Value

Source Attachment Individual Identifier of PW.

- TAII Type, Length, Value, TAII Value

Target Attachment Individual Identifier of PW.

## 6. Revertive Behavior

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

- o Global revertive mode

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

- o Control plane revertive mode

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

- o Local revertive mode

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

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

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

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

#### 7. IANA Considerations

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

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

#### 8. Security Considerations

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

#### 9. Acknowledgements

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

#### 10. References

## 10.1. Normative References

- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC5659] Bocci, M. and S. Bryant, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.



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

## 10.2. Informative References

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

## Authors' Addresses

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

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

Rahul Aggarwal  
Arktan, Inc

Email: raggarwa\_1@yahoo.com

Wim Henderickx  
Alcatel-Lucent  
Copernicuslaan 50  
2018 Antwerp  
Belgium

Email: [wim.henderickx@alcatel-lucent.be](mailto:wim.henderickx@alcatel-lucent.be)

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

Mingui Zhang  
Peng Zhou  
Huawei  
July 12, 2013

Label Sharing for Fast PE Protection  
draft-zhang-l3vpn-label-sharing-00.txt

## Abstract

This document describes a method to be used by Service Providers to provide fast protection of VPN connections for a CE. Egress PEs in a redundant group always assign the same label for VPN routes from a VRF. These egress PEs create a BGP virtual Next Hop (vNH) in the domain of the IP/MPLS backbone network as an agent of the CE router. Primary and backup tunnels terminated at the vNH are set up by the BGP/MPLS IP VPN based on IGP FRR. If the primary egress PE fails, the backup egress PEs can recognize the "shared" VPN route label and deliver the failure affected packets accordingly.

## Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

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

## Copyright and License Notice

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

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

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

## Table of Contents

1. Introduction . . . . .	3
1.1. Conventions used in this document . . . . .	3
1.2. Terminology . . . . .	3
2. The Label Sharing Method . . . . .	3
2.1. The Virtual Next Hop . . . . .	4
2.1.1. Generating OSPF LSAs . . . . .	5
2.1.2. Generating ISIS LSPs . . . . .	7
2.2. Link Costs Set Up for IGP FRR . . . . .	9
2.3 Label Assignment and Processing . . . . .	10
2.3.1. The VPN Route Label . . . . .	10
2.3.2. The Tunnel Label . . . . .	10
3. Security Considerations . . . . .	10
4. IANA Considerations . . . . .	11
5. References . . . . .	11
5.1. Normative References . . . . .	11
5.2. Informative References . . . . .	11
Author's Addresses . . . . .	12

## 1. Introduction

For the sake of reliability, ISPs usually connect one CE to multiple PEs. When the primary egress PE fails, a backup egress PE continues to offer VPN connectivity to the CE. If local repair is performed by the upstream neighbor of the primary egress PE on the data path, it's possible to achieve 50msec switchover.

VPN routes learnt from CEs are distributed by egress PEs to ingress PEs that need to know these VPN routes. Egress PEs in a redundant group (RG) MUST allocate the same VPN route label for routes of the same VPN. When the primary egress PE fails, data packets are redirected to a backup egress PE by the PLR router, the backup PE can recognize the VPN route label in these data packets and deliver them correctly. The method developed in this document is so called "Label Sharing for Fast PE Protection". This method requires only software update on egress PE routers while their data plane remains unchanged.

This document supposes BGP/MPLS IP VPN is deployed on the backbone and Label Distribution Protocol (LDP) is used as the tunneling technology. Through generating virtual LSAs/LSPs in OSPF/ISIS, egress PEs in an RG create a virtual router (the vNH) in the IP/MPLS backbone to represent the CE router. When the VPN route is distributed, those egress PEs use vNH as the "BGP next hop". The vNH will be treated as the egress point of the tunnel by other routers. Metrics for the virtual links attached to the vNH are set up in a way that the IGP FRR mechanism defined in [LFA] can be leveraged to achieve local protection.

### 1.1. Conventions used in this document

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

### 1.2. Terminology

VRF: Virtual Routing and Forwarding table  
FRR: Fast ReRouting  
PLR: Point of Local Repair  
LFA: loop-free alternate

## 2. The Label Sharing Method

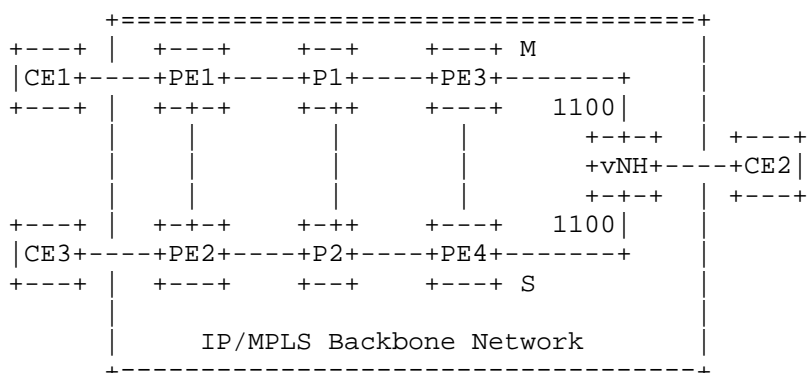


Figure 2.1: Egress PE routers share the same VPN route label.

A CE router is usually connected to multiple PE routers of the IP/MPLS backbone network for the sake of reliability. Figure 2.1 shows such a scenario. In this document, PE1 and PE2 are defined as ingress routers and PE3 and PE4 are defined as egress routers. Suppose PE3 is the primary PE while PE4 is the backup egress PE. In this document, we suppose there are two PEs in one RG. It's possible to expand the method to support more than two PEs in one RG, though it is out the scope of this document.

Those egress PE routers may discover each other as in the same RG from the CE routes learning process which can be a dynamic routing algorithm or a static routing configuration [RFC4364].

### 2.1. The Virtual Next Hop

Egress PEs create a vNH router in IGP to represent the set of CEs dual-homed to the same egress PEs in the Service Provider's backbone. The PE with the highest priority in the RG determines the loopback IP address for the vNH. This loopback IP address can be configured manually or automatically. The SystemID of the vNH under ISIS is composed based on this loopback IP address. The router LSA/LSP for the vNH is generated by the egress PE with the highest priority. This router LSA/LSP also includes the the outgoing links of the vNH. For the incoming links of the vNH, all egress PEs need include these P2P adjacencies in their router LSAs/LSPs.

Egress PEs may create multiple vNHs for one CE. Then multiple tunnels can be set up from ingress PEs to the vNHs. Ingress PEs can choose from these tunnel routes to achieve load balance for the CE.

The overload mode MUST be set so that the rest routers in the network will not route transit traffic through the vNH. In OSPF, the overload

mode can be set up through setting the link weights from the vNH to egress PEs to the maximum link weight which is 0xFFFF. In ISIS, this overload mode is realized as setting the overload bit in the LSP of the vNH.

#### 2.1.1.1. Generating OSPF LSAs

The following Type 1 Router-LSA is flooded by the egress PE with the highest priority. As defined in [RFC2328], this LSA can only be flooded throughout a single area.

0										1										2										3																			
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
LS age										Options										LS type																													
Link State ID																																																	
Advertising Router																																																	
LS sequence number																																																	
LS checksum																				length																													
0					V E B					0										# links																													
Link ID																																																	
Link Data																																																	
Type										# TOS										metric																													
...																																																	
TOS										0										TOS metric																													
Link ID																																																	
Link Data																																																	
...																																																	

#### LS age

The time in seconds since the LSA was originated. (Set to 0x708 by default.)

#### Options

As defined in [RFC2328], options = (E-bit).

LS type  
1

Link State ID  
Same as the Advertising Router

Advertising Router  
The Router ID of the vNH.

LS sequence number  
As defined in [RFC2328].

LS checksum  
As defined and computed in [RFC2328].

length  
The length in bytes of the LSA. This includes the 20 byte LSA header. (As defined and computed in [RFC2328].)

VEB  
As defined in [RFC2328], set its value to 000.

#links  
The number of router links described in this LSA. It equals to the number of Egress PEs in the RG.

The following fields are used to describe each router link connected to an egress PE. Each router link is typed as Type 1 Point-to-point connection to another router.

Link ID  
The Router ID of one of the egress PEs in the RG.

Link Data  
It specifies the interface's MIB-II [RFC1213] ifIndex value. It ranges between 1 and the value of ifNumber. The ifNumber equals to the number of the PEs in the RG. The PE with the highest priority sorts the PEs according to their unsigned integer Router ID in the ascend order and assigns the ifIndex for each.

Type  
Value 1 is used, indicating the router link is a point-to-point connection to another router.

# TOS  
This field is set to 0 for this version.

Metric



It is set to 0xFFFF.

The fields used here to describe the virtual router links are also included in the Router-LSA of each egress PEs. The Link ID is replaced with the Router ID of the vNH. The Link Data specifies the interface's MIB-II [RFC1213] ifIndex value. The "Metric" field is set as defined in Section 2.2.

#### 2.1.2. Generating ISIS LSPs

The primary egress PE generates the following level 1 LSP to describe the vNH node.

	No. of octets
+-----+	
Intradomain Routeing   Protocol Discriminator	1
+-----+	
Length Indicator	1
+-----+	
Version/Protocol ID   Extension	1
+-----+	
ID Length	1
+-----+	
R R R  PDU Type	1
+-----+	
Version	1
+-----+	
Reserved	1
+-----+	
Maximum Area Address	1
+-----+	
PDU Length	2
+-----+	
Remaining Lifetime	2
+-----+	
LSP ID	ID Length + 2
+-----+	
Sequence Number	4
+-----+	
Checksum	2
+-----+	
P ATT LSPDBOL IS Type	1
+-----+	
: Variable Length Fields :	Variable
+-----+	

Intradomain Routeing Protocol Discriminator - 0x83 (as defined in [ISIS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [ISIS]

PDU Type (bits 1 through 5) - 18

Version - 1

Reserved - transmitted as zero, ignored on receipt

Maximum Area Address - same as the primary egress PE

PDU Length - Entire Length of this PDU, in octets, including the header.

Remaining Lifetime - Number of seconds before this LSP is considered expired. (Set to 0x384 by default.)

LSP ID - the system ID of the source of the LSP. It is structured as follows:

+-----+	
Source ID	6
+-----+	
Pseudonode ID	1
+-----+	
LSP Number	1
+-----+	

Source ID - SystemID of the vNH

Pseudonode ID - Transmitted as zero

LSP Number - Fragment number

Sequence Number - sequence number of this LSP (as defined in [ISIS])

Checksum - As defined and computed in [ISIS]

P - Bit 8 - 0

ATT - Bit 7-4 - 0

LSDBOL - Bit 3 - 1

IS Type - Bit 1 and 2 - bit 1 set, indicating the vNH is a Level 1 Intermediate System

In the Variable Length Field, each link outgoing from the vNH to an egress PE is depicted by a Type #22 Extended Intermediate System Neighbors TLV [RFC5305]. The egress PE is identified by the 6 octets SystemID plus one octet of all-zero pseudonode number. The 3 octets metric is set as that in Section 2.2. None sub-TLVs is used by this version, therefore the value of the one octet length of sub-TLVs is 0. The Type #22 TLV requires 11 octets.

The Type #22 TLV is also included in the LSP of each egress PE to depict the incoming link of the vNH. Only the 6 octets SystemID is replaced with the SystemID of the vNH.

## 2.2. Link Costs Set Up for IGP FRR

Tunnel LSPs are set up based on IGP routes through LDP signaling. If the IGP costs for the links between egress PEs and the vNH can be set up in a way that one egress PE appears on the primary path while other PE(s) appears on the backup path, the PLR can make use of the multiple egress PEs to achieve fast failure protection. Suppose [LFA] is being used as the IGP FRR mechanism, the link weights can be set up according to the following rule.

1. This document supposes bidirectional link weights are being used. Assume the weight for the link between PE3 and vNH is "M" and the weight for the link between PE4 and vNH is "S". The weight for the link between PE3 and PE4 is C34.

2. Px is a neighbor of PE3. This Px will act as the PLR. Suppose Pxy is Px's neighbor with the shortest path to PE4, after PE3 is removed from the topology. The cost of this path is Sxy4.

3. Add PE3 back to the topology. The cost of the path from Pxy to PE3 is Sxy3.

4. "M" and "S" can be set up as long as the following two equations hold.

$$\text{eq1: } Sxy4+S < Sxy3+M$$

$$\text{eq2: } C34+S > M$$

Although this document designs the method based on [LFA] which is widely deployed, other IGP FRR mechanisms can also be utilized to

achieve the protection. For example, [MRT] is applicable regardless of how the link weights are set up.

## 2.3 Label Assignment and Processing

### 2.3.1. The VPN Route Label

Egress PEs use BGP to distribute to ingress PEs the routes that they have learnt from CEs [RFC4364]. When egress PEs distribute the routes of the VPN that the CE is in, they MUST assign the same "VPN route label" for one VPN (per VRF label assignment). This label will become the first label of a data packet. The IP address of the vNH is used as the "BGP next hop". For example, in Figure 2.1, both PE3 and PE4 use 1100 as the VPN route label for the routes learnt from CE2.

Suppose PE3 fails and the packet with VPN route label 1100 is redirected to PE4, PE4 recognizes 1100 as the VPN route label it assigned for the VPN that the CE is in. As specified in Section 5 of [RFC4364], PE4 will be able to determine, the attachment circuit over which the packet should be transmitted (to the CE) as well as the data link layer header for that interface. It need to lookup the packet's destination address in the VRF identified by the VPN route label 1100.

When we speak of a PE fails, it may also means that a link to the PE on the primary tunnel fails. In general, we can say that a primary PE fails means that this PE becomes unreachable via its upstream neighbor on the primary tunnel.

The shared label may be manually configured or negotiated through signaling between egress PEs. In [LS-ICCP], application TLVs are defined for [ICCP] to achieve such kind of signaling.

### 2.3.2. The Tunnel Label

This document supposes Label Distribution Protocol is being used as the tunneling technology. The LDP LSP tunnel follows a IGP route from ingress PEs to the vNH. The backup path to vNH can be calculated according to IGP FRR mechanism, such as [MRT] and [LFA].

The ingress PE tunnels the data packet through the backbone network using the "tunnel label" as the second entry of the label stack. The "VPN route label" is not visible again until the MPLS packet reaches the egress PE. The egress PE need pop the second label and deliver the packet according to the "VPN route label".

## 3. Security Considerations

This document raises no new security issues.

#### 4. IANA Considerations

No requirements for IANA.

#### 5. References

##### 5.1. Normative References

- [LFA]      Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [ICCP]      L. Martini, S. Salam, et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11.txt, work in progress.
- [ISIS]      ISO, "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)," ISO/IEC 10589:2002.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.
- [LS-ICCP] M. Zhang, P. Zhou, "ICCP Application TLVs for VPN Route Label Sharing", draft-zhang-pwe3-iccp-label-sharing-00.txt, work in progress

##### 5.2. Informative References

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

## Author's Addresses

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

Email: zhangmingui@huawei.com

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

Email: Jewpon.zhou@huawei.com

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

Mingui Zhang  
Peng Zhou  
Huawei  
July 12, 2013

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

Abstract

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

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

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

Copyright and License Notice

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

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

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

## Table of Contents

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



## 1. Introduction

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

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

### 1.1. Conventions used in this document

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

### 1.2. Terminology

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

## 2. Label Sharing TLVs in ICCP

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

### 2.1. Label Sharing Connect TLV

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

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|U|F|  Type =0x0111(TBD)          |      Length          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|  Protocol Version=0x0001      |A|  Reserved          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               |Optional Sub-TLVs(None for This Version)|
~                               ~                               ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

- U and F Bits

Both are set to 0.

- Type

set to 0x0111 (TBD) for "Label Sharing Connect TLV"

- Length

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

- Protocol Version

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

- A bit

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

- Reserved

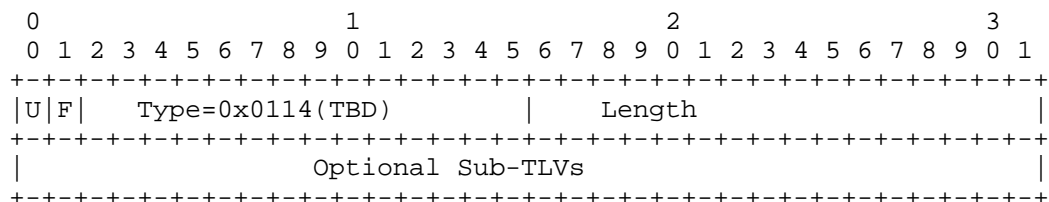
Reserved for future use.

- Optional Sub-TLVs

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

## 2.2. Label Sharing Disconnect TLV

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



- U and F Bits

Both are set to 0.

- Type

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

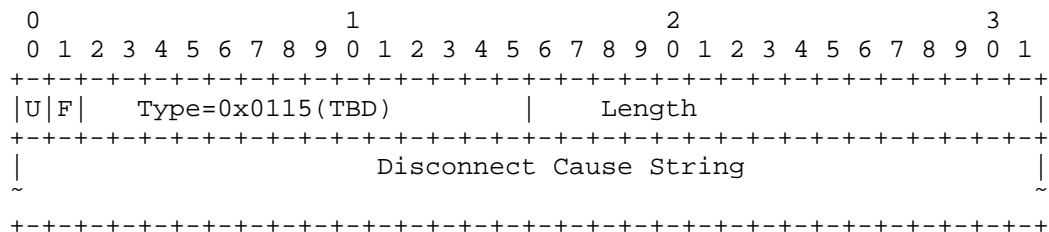
- Length

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

- Optional Sub-TLVs

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

#### 2.2.1. Label Sharing Disconnect Cause TLV



- U and F Bits

Both are set to 0.

- Type

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

- Length

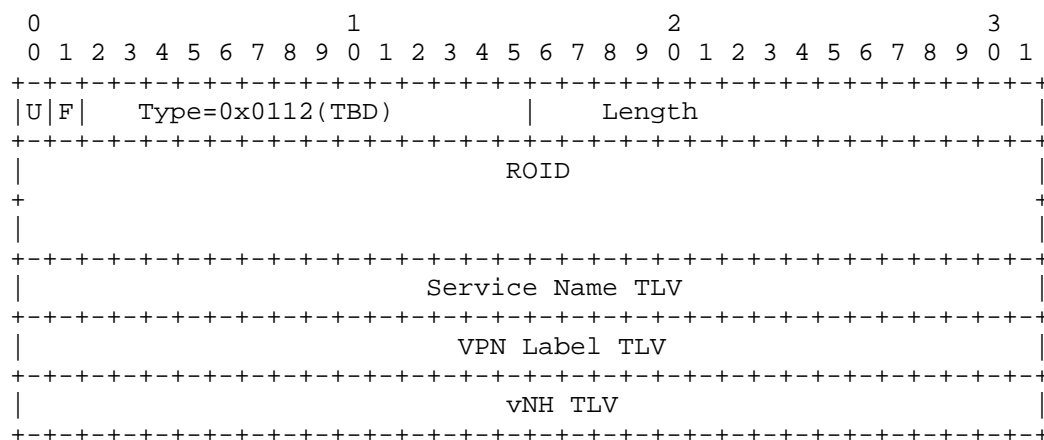
Length of the TLV in octets excluding the U-bit, F-bit, Type, and Length fields.

- Disconnect Cause String

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

### 2.3. Label Sharing Application Data TLVs

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



- U and F Bits

Both are set to 0.

- Type

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

- Length

Length of the MAC address, which is 6 octets.

- ROID

As defined in the ROID section of [ICCP].

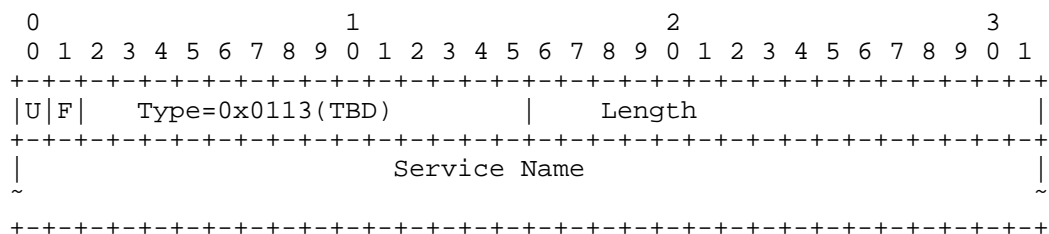
- Sub-TLVs

i Service Name TLV

ii VPN Label TLV

iii vNH TLV

### 2.3.1. Service Name TLV



- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Service Name

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

### 2.3.2. VPN Label TLV

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

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
U F										Type=0x0102(TBD)										Length																			
										Priority										Reserved																			
										VPN Label										Reserved																			
										Lower Label										Upper Label																			

- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Priority

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

- Reserved

Reserved for future use.

- VPN Label

The VPN label to be shared among the RG.

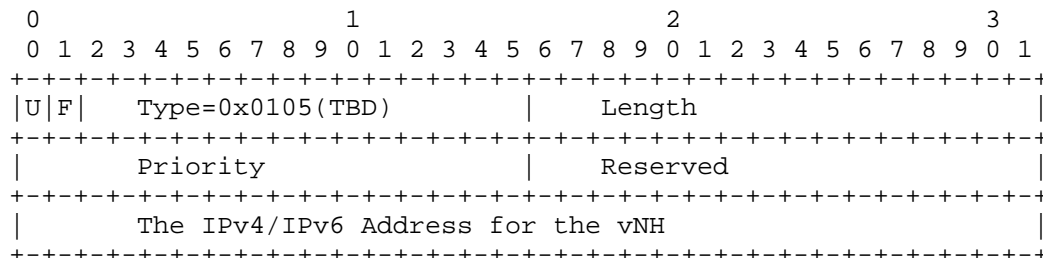
- Lower/Upper Label

The lower/upper bound of a valid VPN label.

### 2.3.3. vNH TLV

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

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



- U and F Bits

Both are set to 0.

- Type

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

- Length

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

- Priority

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

- Reserved

Reserved for future use.

- IPv4/IPv6 Address for the vNH

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

### 3. Security Considerations

This document raises no new security issues.

#### 4. IANA Considerations

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

#### 5. References

##### 5.1. Normative References

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

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

##### 5.2. Informative References

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

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



## Author's Addresses

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

Email: zhangmingui@huawei.com

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

Email: Jewpon.zhou@huawei.com