

Internet Engineering Task Force
Internet Draft
Intended status: Standards Track
Expires: August 20, 2013

Luca Martini Ed.
Giles Heron Ed.

Cisco

February 20, 2013

Pseudowire Setup and Maintenance using the Label Distribution Protocol

draft-ietf-pwe3-rfc4447bis-01.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 August 20, 2013

Abstract

Layer 2 services (such as Frame Relay, Asynchronous Transfer Mode, and Ethernet) can be "emulated" over an MPLS backbone by encapsulating the Layer 2 Packet Data Units (PDU) and then transmitting them over "pseudowires". It is also possible to use pseudowires to provide low-rate Time Division Multiplexed and Synchronous Optical NETWORKing

circuit emulation over a MPLS-enabled network. This document specifies a protocol for establishing and maintaining the pseudowires, using extensions to the Label Distribution Protocol (LDP). Procedures for encapsulating Layer 2 PDUs are specified in a set of companion documents.

Table of Contents

1	Introduction	4
2	Specification of Requirements	6
3	The Pseudowire Label	6
4	Details Specific to Particular Emulated Services	8
4.1	IP Layer 2 Transport	8
5	LDP	8
5.1	The PWid FEC Element	9
5.2	The Generalized PWid FEC Element	11
5.2.1	Attachment Identifiers	11
5.2.2	Encoding the Generalized PWid FEC Element	13
5.2.2.1	Interface Parameters TLV	14
5.2.2.2	PW Grouping ID TLV	14
5.2.3	Signaling Procedures	15
5.3	Signaling of Pseudo Wire Status	16
5.3.1	Use of Label Mappings Messages.	16
5.3.2	Signaling PW status.	17
5.3.3	Pseudowire Status Negotiation Procedures	18
5.4	Interface Parameters sub-TLV	20
6	Control Word	21
6.1	PW types for which the control word is REQUIRED	21
6.2	PW types for which the control word is NOT mandatory ..	21
6.3	Control-Word Renegotiation by Label Request Message ..	22
6.4	LDP label Withdrawal procedures	23
6.5	Sequencing Considerations	24
6.5.1	Label Advertisements	24
6.5.2	Label Release	25
7	IANA Considerations	25
7.1	LDP TLV TYPE	25
7.2	LDP Status Codes	25
7.3	FEC Type Name Space	26
8	Security Considerations	26
8.1	Data-plane Security	26
8.2	Control-Plane Security	27
9	Acknowledgments	28
10	Normative References	28
11	Informative References	28
12	Author Information	29
13	Additional Contributing Authors	30
Ap A	C-bit Handling Procedures Diagram	33

1. Introduction

In [RFC4619], [RFC4717], [RFC4618], and [RFC4448], it is explained how to encapsulate a Layer 2 Protocol Data Unit (PDU) for transmission over an MPLS-enabled network. Those documents specify that a "pseudowire header", consisting of a demultiplexor field, will be prepended to the encapsulated PDU. The pseudowire demultiplexor field is prepended before transmitting a packet on a pseudowire. When the packet arrives at the remote endpoint of the pseudowire, the demultiplexor is what enables the receiver to identify the particular pseudowire on which the packet has arrived. To transmit the packet from one pseudowire endpoint to another, the packet may need to travel through a "Packet Switched Network (PSN) tunnel"; this will require that an additional header be prepended to the packet.

Accompanying documents [RFC4842, RFC4553] specify methods for transporting time-division multiplexing (TDM) digital signals (TDM circuit emulation) over a packet-oriented MPLS-enabled network. The transmission system for circuit-oriented TDM signals is the Synchronous Optical Network [SDH] (SONET)/Synchronous Digital Hierarchy (SDH) [ITU]. To support TDM traffic, which includes voice, data, and private leased-line service, the pseudowires must emulate the circuit characteristics of SONET/SDH payloads. The TDM signals and payloads are encapsulated for transmission over pseudowires. A pseudowire demultiplexor and a PSN tunnel header is prepended to this encapsulation.

[RFC4553] describes methods for transporting low-rate time-division multiplexing (TDM) digital signals (TDM circuit emulation) over PSNs, while [RFC4842] similarly describes transport of high-rate TDM (SONET/SDH). To support TDM traffic, the pseudowires must emulate the circuit characteristics of the original T1, E1, T3, E3, SONET, or SDH signals. [RFC4553] does this by encapsulating an arbitrary but constant amount of the TDM data in each packet, and the other methods encapsulate TDM structures.

In this document, we specify the use of the MPLS Label Distribution Protocol, LDP [RFC5036], as a protocol for setting up and maintaining the pseudowires. In particular, we define new TLVs, FEC elements, parameters, and codes for LDP, which enable LDP to identify pseudowires and to signal attributes of pseudowires. We specify how a pseudowire endpoint uses these TLVs in LDP to bind a demultiplexor field value to a pseudowire, and how it informs the remote endpoint of the binding. We also specify procedures for reporting pseudowire status changes, for passing additional information about the pseudowire as needed, and for releasing the bindings. These procedures are intended to be independent of the underlying version of IP used for LDP signaling.

In the protocol specified herein, the pseudowire demultiplexor field is an MPLS label. Thus, the packets that are transmitted from one end of the pseudowire to the other are MPLS packets, which must be transmitted through an MPLS tunnel. However, if the pseudowire endpoints are immediately adjacent and penultimate hop popping behavior is in use, the MPLS tunnel may not be necessary. Any sort of PSN tunnel can be used, as long as it is possible to transmit MPLS packets through it. The PSN tunnel can itself be an MPLS LSP, or any other sort of tunnel that can carry MPLS packets. Procedures for setting up and maintaining the MPLS tunnels are outside the scope of this document.

This document deals only with the setup and maintenance of point-to-point pseudowires. Neither point-to-multipoint nor multipoint-to-point pseudowires are discussed.

QoS-related issues are not discussed in this document. The following two figures describe the reference models that are derived from [RFC3985] to support the PW emulated services.

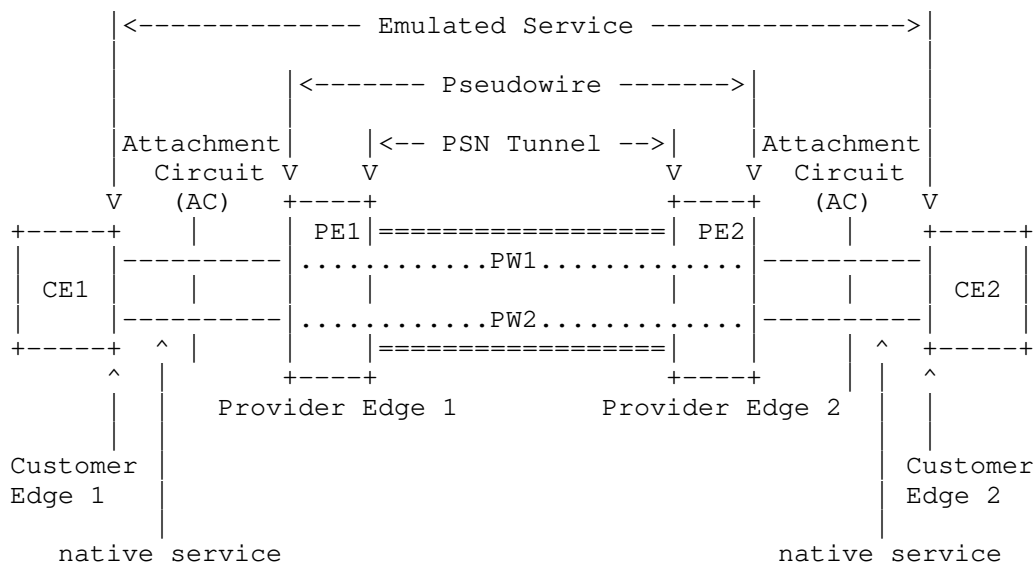


Figure 1: PWE3 Reference Model

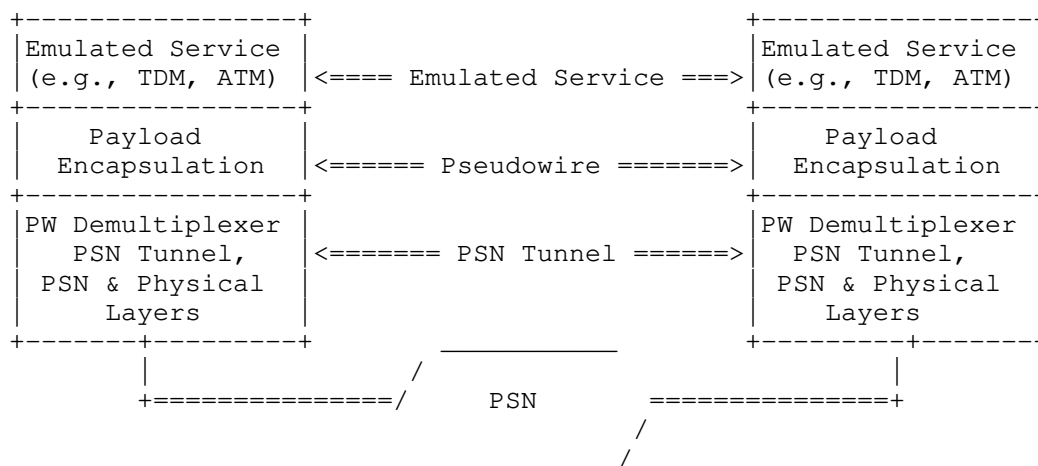


Figure 2: PWE3 Protocol Stack Reference Model

For the purpose of this document, PE1 will be defined as the ingress router, and PE2 as the egress router. A layer 2 PDU will be received at PE1, encapsulated at PE1, transported and decapsulated at PE2, and transmitted out of PE2.

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 [RFC2119].

3. The Pseudowire Label

Suppose that it is desired to transport Layer 2 PDUs from ingress LSR PE1 to egress LSR PE2, across an intervening MPLS-enabled network. We assume that there is an MPLS tunnel from PE1 to PE2. That is, we assume that PE1 can cause a packet to be delivered to PE2 by encapsulating the packet in an "MPLS tunnel header" and sending the result to one of its adjacencies. The MPLS tunnel is an MPLS Label Switched Path (LSP); thus, putting on an MPLS tunnel encapsulation is a matter of pushing on an MPLS label.

We presuppose that a large number of pseudowires can be carried through a single MPLS tunnel. Thus it is never necessary to maintain state in the network core for individual pseudowires. We do not presuppose that the MPLS tunnels are point to point; although the

pseudowires are point to point, the MPLS tunnels may be multipoint to point. We do not presuppose that PE2 will even be able to determine the MPLS tunnel through which a received packet was transmitted. (For example, if the MPLS tunnel is an LSP and penultimate hop popping is used, when the packet arrives at PE2, it will contain no information identifying the tunnel.)

When PE2 receives a packet over a pseudowire, it must be able to determine that the packet was in fact received over a pseudowire, and it must be able to associate that packet with a particular pseudowire. PE2 is able to do this by examining the MPLS label that serves as the pseudowire demultiplexor field shown in Figure 2. Call this label the "PW label".

When PE1 sends a Layer 2 PDU to PE2, it creates an MPLS packet by adding the PW label to the packet, thus creating the first entry of the label stack. If the PSN tunnel is an MPLS LSP, the PE1 pushes another label (the tunnel label) onto the packet as the second entry of the label stack. The PW label is not visible again until the MPLS packet reaches PE2. PE2's disposition of the packet is based on the PW label.

If the payload of the MPLS packet is, for example, an ATM AAL5 PDU, the PW label will generally correspond to a particular ATM VC at PE2. That is, PE2 needs to be able to infer from the PW label the outgoing interface and the VPI/VCI value for the AAL5 PDU. If the payload is a Frame Relay PDU, then PE2 needs to be able to infer from the PW label the outgoing interface and the DLCI value. If the payload is an Ethernet frame, then PE2 needs to be able to infer from the PW label the outgoing interface, and perhaps the VLAN identifier. This process is uni-directional and will be repeated independently for bi-directional operation. It is REQUIRED that the same PW ID and PW type be assigned for a given circuit in both directions. The group ID (see below) MUST NOT be required to match in both directions. The transported frame MAY be modified when it reaches the egress router. If the header of the transported Layer 2 frame is modified, this MUST be done at the egress LSR only. Note that the PW label must always be at the bottom of the packet's label stack, and labels MUST be allocated from the per-platform label space.

This document does not specify a method for distributing the MPLS tunnel label or any other labels that may appear above the PW label on the stack. Any acceptable method of MPLS label distribution will do. This document specifies a protocol for assigning and distributing the PW label. This protocol is LDP, extended as specified in the remainder of this document. An LDP session must be set up between the pseudowire endpoints. LDP MUST exchange PW FEC

label bindings in downstream unsolicited manner, independent of the negotiated label advertisement mode of the LDP session. LDP's "liberal label retention" mode SHOULD be used.

In addition to the protocol specified herein, static assignment of PW labels may be used, and implementations of this protocol SHOULD provide support for static assignment. PW encapsulation is always symmetrical in both directions of traffic along a specific PW, whether the PW uses an LDP control plane or not.

This document specifies all the procedures necessary to set up and maintain the pseudowires needed to support "unswitched" point to point services, where each endpoint of the pseudowire is provisioned with the identity of the other endpoint. There are also protocol mechanisms specified herein that can be used to support switched services and other provisioning models. However, the use of the protocol mechanisms to support those other models and services is not described in this document.

4. Details Specific to Particular Emulated Services

4.1. IP Layer 2 Transport

This mode carries IP packets over a pseudowire. The encapsulation used is according to [RFC3032]. The PW control word MAY be inserted between the MPLS label stack and the IP payload. The encapsulation of the IP packets for forwarding on the attachment circuit is implementation specific, is part of the native service processing (NSP) function [RFC3985], and is outside the scope of this document.

5. LDP

The PW label bindings are distributed using the LDP downstream unsolicited mode described in [RFC5036]. The PEs will establish an LDP session using the Extended Discovery mechanism described in [LDP, sectionn 2.4.2 and 2.5].

An LDP Label Mapping message contains an FEC TLV, a Label TLV, and zero or more optional parameter TLVs.

The FEC TLV is used to indicate the meaning of the label. In the current context, the FEC TLV would be used to identify the particular pseudowire that a particular label is bound to. In this specification, we define two new FEC TLVs to be used for identifying pseudowires. When setting up a particular pseudowire, only one of these FEC TLVs is used. The one to be used will depend on the particular service being emulated and on the particular provisioning model being supported.

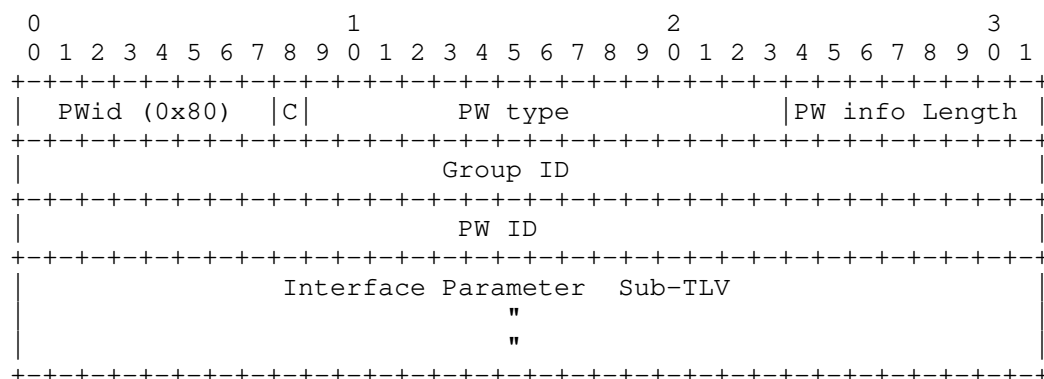
LDP allows each FEC TLV to consist of a set of FEC elements. For setting up and maintaining pseudowires, however, each FEC TLV MUST contain exactly one FEC element.

The LDP base specification has several kinds of label TLVs, including the Generic Label TLV, as specified in [RFC5036], section 3.4.2.1. For setting up and maintaining pseudowires, the Generic Label TLV MUST be used.

5.1. The PWid FEC Element

The PWid FEC element may be used whenever both pseudowire endpoints have been provisioned with the same 32-bit identifier for the pseudowire.

For this purpose, a new type of FEC element is defined. The FEC element type is 0x80 and is defined as follows:



- PW type

A 15 bit quantity containing a value which represents the type of PW. Assigned Values are specified in "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)" [RFC4446].

- Control word bit (C)

The bit (C) is used to flag the presence of a control word as follows:

- C = 1 control word present on this PW.
- C = 0 no control word present on this PW.

Please see the section "C-Bit Handling Procedures" for further explanation.

- PW information length

Length of the PW ID field and the interface parameters sub-TLV in octets. If this value is 0, then it references all PWs using the specified group ID, and there is no PW ID present, nor are there any interface parameter sub-TLVs.

- Group ID

An arbitrary 32 bit value which represents a group of PWs that is used to create groups in the PW space. The group ID is intended to be used as a port index, or a virtual tunnel index. To simplify configuration a particular PW ID at ingress could be part of a Group ID assigned to the virtual tunnel for transport to the egress router. The Group ID is very useful for sending wild card label withdrawals, or PW wild card status notification messages to remote PEs upon physical port failure.

- PW ID

A non-zero 32-bit connection ID that together with the PW type, identifies a particular PW. Note that the PW ID and the PW type MUST be the same at both endpoints.

- Interface Parameter Sub-TLV

This variable length TLV is used to provide interface specific parameters, such as attachment circuit MTU.

Note that as the "interface parameter sub-TLV" is part of the FEC, the rules of LDP make it impossible to change the interface parameters once the pseudowire has been set up. Thus the interface parameters field must not be used to pass information, such as status information, which may change during the life of the pseudowire. Optional parameter TLVs should be used for that purpose.

Using the Pwid FEC, each of the two pseudowire endpoints independently initiates the set up of a unidirectional LSP. An outgoing LSP and an incoming LSP are bound together into a single pseudowire if they have the same PW ID and PW type.

5.2. The Generalized PWid FEC Element

The PWid FEC element can be used if a unique 32-bit value has been assigned to the PW, and if each endpoint has been provisioned with that value. The Generalized PWid FEC element requires that the PW endpoints be uniquely identified; the PW itself is identified as a pair of endpoints. In addition, the endpoint identifiers are structured to support applications where the identity of the remote endpoints needs to be auto-discovered rather than statically configured.

The "Generalized PWid FEC Element" is FEC type 0x81.

The Generalized PWid FEC Element does not contain anything corresponding to the "Group ID" of the PWid FEC element. The functionality of the "Group ID" is provided by a separate optional LDP TLV, the "PW Grouping TLV", described below. The Interface Parameters field of the PWid FEC element is also absent; its functionality is replaced by the optional Interface Parameters TLV, described below.

5.2.1. Attachment Identifiers

As discussed in [RFC3985], a pseudowire can be thought of as connecting two "forwarders". The protocol used to set up a pseudowire must allow the forwarder at one end of a pseudowire to identify the forwarder at the other end. We use the term "attachment identifier", or "AI", to refer to the field that the protocol uses to identify the forwarders. In the PWid FEC, the PWid field serves as the AI. In this section, we specify a more general form of AI that is structured and of variable length.

Every Forwarder in a PE must be associated with an Attachment Identifier (AI), either through configuration or through some algorithm. The Attachment Identifier must be unique in the context of the PE router in which the Forwarder resides. The combination <PE router IP address, AI> must be globally unique.

It is frequently convenient to regard a set of Forwarders as being members of a particular "group", where PWs may only be setup among members of a group. In such cases, it is convenient to identify the Forwarders relative to the group, so that an Attachment Identifier would consist of an Attachment Group Identifier (AGI) plus an Attachment Individual Identifier (AII).

An Attachment Group Identifier may be thought of as a VPN-id, or a VLAN identifier, some attribute that is shared by all the Attachment

PWs (or pools thereof) that are allowed to be connected.

The details of how to construct the AGI and AII fields identifying the pseudowire endpoints are outside the scope of this specification. Different pseudowire applications, and different provisioning models, will require different sorts of AGI and AII fields. The specification of each such application and/or model must include the rules for constructing the AGI and AII fields.

As previously discussed, a (bidirectional) pseudowire consists of a pair of unidirectional LSPs, one in each direction. If a particular pseudowire connects PE1 with PE2, the PW direction from PE1 to PE2 can be identified as:

<PE1, <AGI, AII1>, PE2, <AGI, AII2>>.,

and the PW direction from PE2 to PE1 can be identified by:

<PE2, <AGI, AII2>, PE1, <AGI, AII1>>.

Note that the AGI must be the same at both endpoints, but the AII will in general be different at each endpoint. Thus, from the perspective of a particular PE, each pseudowire has a local or "Source AII", and a remote or "Target AII". The pseudowire setup protocol can carry all three of these quantities:

- Attachment Group Identifier (AGI).
- Source Attachment Individual Identifier (SAII)
- Target Attachment Individual Identifier (TAII)

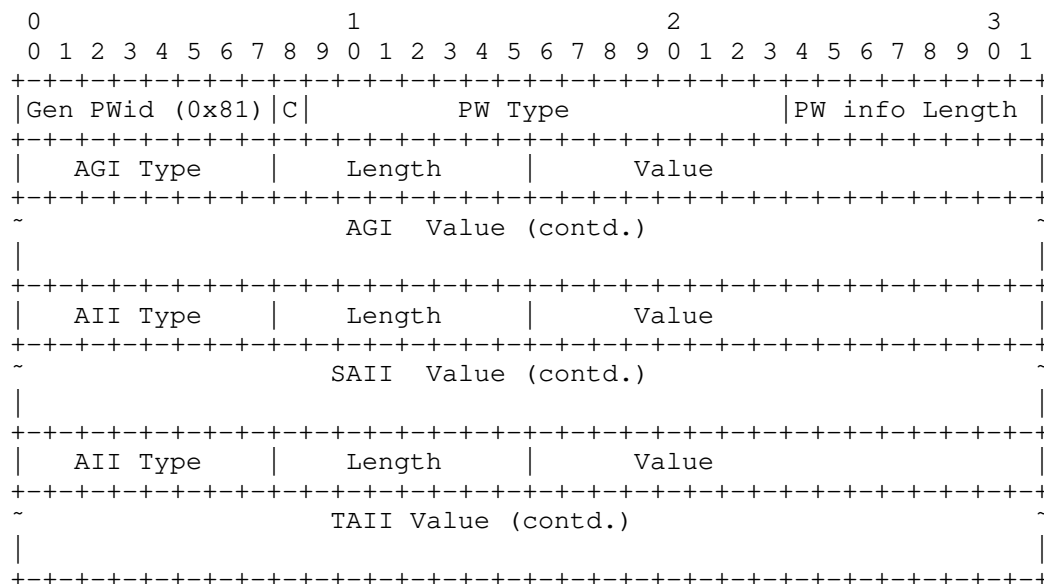
If the AGI is non-null, then the Source AI (SAI) consists of the AGI together with the SAII, and the Target AI (TAI) consists of the TAII together with the AGI. If the AGI is null, then the SAII and TAII are the SAI and TAI, respectively.

The interpretation of the SAI and TAI is a local matter at the respective endpoint.

The association of two unidirectional LSPs into a single bidirectional pseudowire depends on the SAI and the TAI. Each application and/or provisioning model that uses the Generalized PWid FEC element must specify the rules for performing this association.

5.2.2. Encoding the Generalized Pwid FEC Element

FEC element type 0x81 is used. The FEC element is encoded as follows:



This document does not specify the AII and AGI type field values; specification of the type field values to be used for a particular application is part of the specification of that application. IANA has assigned these values using the method defined in the [RFC4446] document.

The SAII, TAI, and AGI are simply carried as octet strings. The length byte specifies the size of the Value field. The null string can be sent by setting the length byte to 0. If a particular application does not need all three of these sub-elements, it MUST send all the sub-elements but set the length to 0 for the unused sub-elements.

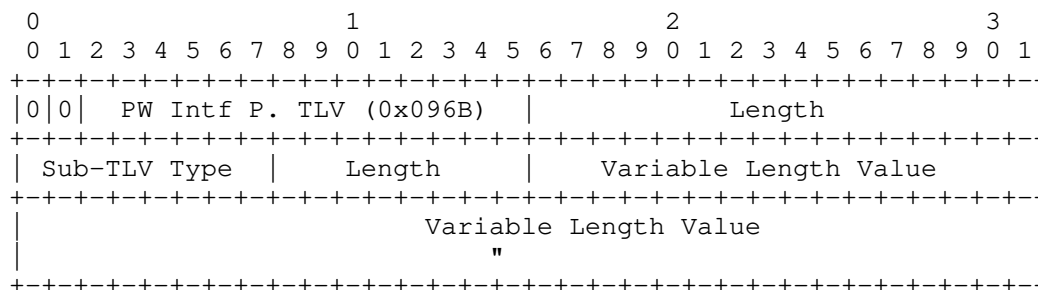
The PW information length field contains the length of the SAII, TAI, and AGI, combined in octets. If this value is 0, then it references all PWs using the specific grouping ID (specified in the PW grouping ID TLV). In this case, there are no other FEC element fields (AGI, SAI, etc.) present, nor any interface parameters TLVs.

Note that the interpretation of a particular field as AGI, SAI, or TAI depends on the order of its occurrence. The type field identifies the type of the AGI, SAI, or TAI. When comparing two

occurrences of an AGI (or SAGI or TAGI), the two occurrences are considered identical if the type, length, and value fields of one are identical, respectively, to those of the other.

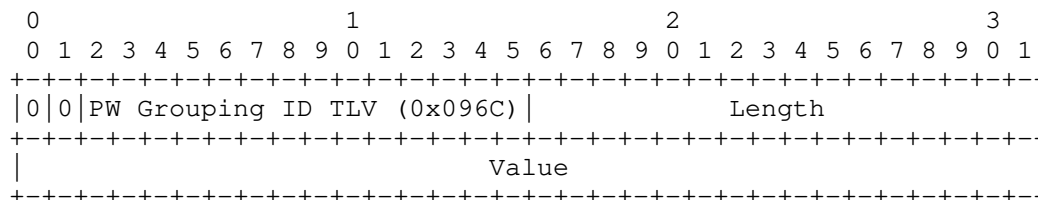
5.2.2.1. Interface Parameters TLV

This TLV MUST only be used when sending the Generalized PW FEC. It specifies interface-specific parameters. Specific parameters, when applicable, MUST be used to validate that the PEs and the ingress and egress ports at the edges of the circuit have the necessary capabilities to interoperate with each other.



A more detailed description of this field can be found in the section "Interface Parameters Sub-TLV", below.

5.2.2.2. PW Grouping ID TLV



The PW Grouping ID is an arbitrary 32-bit value that represents an arbitrary group of PWs. It is used to create group PWs; for example, a PW Grouping ID can be used as a port index and assigned to all PWs that lead to that port. Use of the PW Grouping ID enables one to send "wild card" label withdrawals, or "wild card" status notification messages, to remote PEs upon physical port failure.

Note Well: The PW Grouping ID is different from and has no relation to, the Attachment Group Identifier.

The PW Grouping ID TLV is not part of the FEC and will not be

advertised except in the PW FEC advertisement. The advertising PE MAY use the wild card withdraw semantics, but the remote PEs MUST implement support for wild card messages. This TLV MUST only be used when sending the Generalized PW ID FEC.

To issue a wild card command (status or withdraw):

- Set the PW Info Length to 0 in the Generalized PWid FEC Element.
- Send only the PW Grouping ID TLV with the FEC (No AGI/SAII/TAII is sent).

5.2.3. Signaling Procedures

In order for PE1 to begin signaling PE2, PE1 must know the address of the remote PE2, and a TAI. This information may have been configured at PE1, or it may have been learned dynamically via some autodiscovery procedure.

The egress PE (PE1), which has knowledge of the ingress PE, initiates the setup by sending a Label Mapping Message to the ingress PE (PE2). The Label Mapping message contains the FEC TLV, carrying the Generalized PWid FEC Element (type 0x81). The Generalized PWid FEC element contains the AGI, SAI, and TAI information.

Next, when PE2 receives such a Label Mapping message, PE2 interprets the message as a request to set up a PW whose endpoint (at PE2) is the Forwarder identified by the TAI. From the perspective of the signaling protocol, exactly how PE2 maps AIs to Forwarders is a local matter. In some Virtual Private Wire Services (VPWS) provisioning models, the TAI might, for example, be a string that identifies a particular Attachment Circuit, such as "ATM3VPI4VCI5", or it might, for example, be a string, such as "Fred", that is associated by configuration with a particular Attachment Circuit. In VPLS, the AGI could be a VPN-id, identifying a particular VPLS instance.

If PE2 cannot map the TAI to one of its Forwarders, then PE2 sends a Label Release message to PE1, with a Status Code of "Unassigned/Unrecognized TAI", and the processing of the Label Mapping message is complete.

The FEC TLV sent in a Label Release message is the same as the FEC TLV received in the Label Mapping being released (but without the interface parameter TLV). More generally, the FEC TLV is the same in all LDP messages relating to the same PW. In a Label Release this means that the SAI is the remote peer's AI and the TAI is the sender's local AI.

If the Label Mapping Message has a valid TAI, PE2 must decide whether to accept it. The procedures for so deciding will depend on the particular type of Forwarder identified by the TAI. Of course, the Label Mapping message may be rejected due to standard LDP error conditions as detailed in [RFC5036].

If PE2 decides to accept the Label Mapping message, then it has to make sure that a PW LSP is set up in the opposite (PE1-->PE2) direction. If it has already signaled for the corresponding PW LSP in that direction, nothing more needs to be done. Otherwise, it must initiate such signaling by sending a Label Mapping message to PE1. This is very similar to the Label Mapping message PE2 received, but the SAI and TAI reversed.

Thus, a bidirectional PW consists of two LSPs, where the FEC of one has the SAI and TAI reversed with respect of the FEC of the other.

5.3. Signaling of Pseudo Wire Status

5.3.1. Use of Label Mappings Messages.

The PEs MUST send Label Mapping Messages to their peers as soon as the PW is configured and administratively enabled, regardless of the attachment circuit state. The PW label should not be withdrawn unless the operator administratively configures the pseudowire down (or the PW configuration is deleted entirely). Using the procedures outlined in this section, a simple label withdraw method MAY also be supported as a legacy means of signaling PW status and AC status. In any case, if the label-to-PW binding is not available the PW MUST be considered in the down state.

Once the PW status negotiation procedures are completed and if they result in the use of the label withdraw method for PW status communication, and this method is not supported by one of the PEs, than that PE must send a Label Release Message to its peer with the following error:

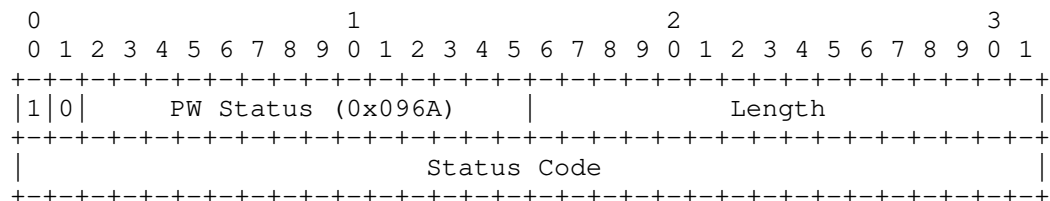
"Label Withdraw PW Status Method Not Supported"

If the label withdraw method for PW status communication is selected for the PW, it will result in the Label Mapping Message being advertised only if the attachment circuit is active. The PW status signaling procedures described in this section MUST be fully implemented.

5.3.2. Signaling PW status.

The PE devices use an LDP TLV to indicate status to their remote peers. This PW Status TLV contains more information than the alternative simple Label Withdraw message.

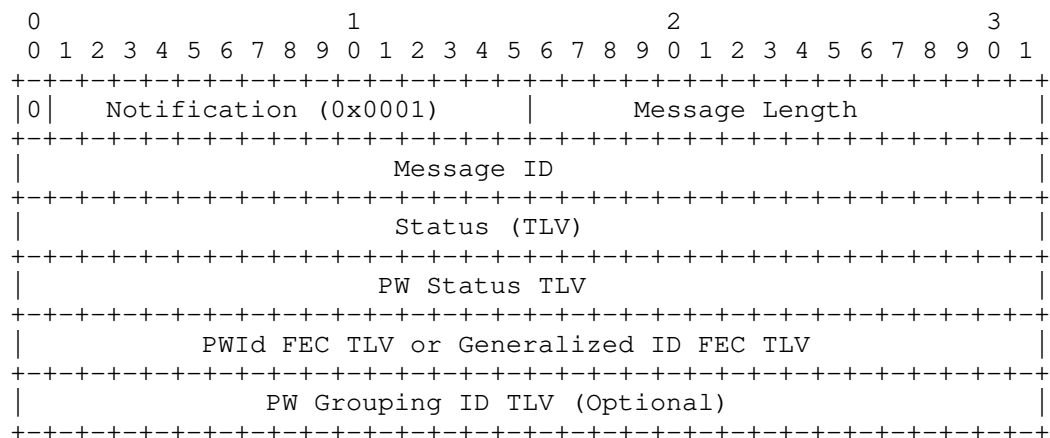
The format of the PW Status TLV is:



The status code is a 4 octet bit field is specified in the PW IANA Allocations document [RFC4446]. The length specifies the length of the Status Code field in octets (equal to 4).

Each bit in the status code field can be set individually to indicate more then a single failure at once. Each fault can be cleared by sending an appropriate Notification message in which the respective bit is cleared. The presence of the lowest bit (PW Not Forwarding) acts only as a generic failure indication when there is a link-down event for which none of the other bits apply.

The Status TLV is transported to the remote PW peer via the LDP Notification message. The general format of the Notification Message is:



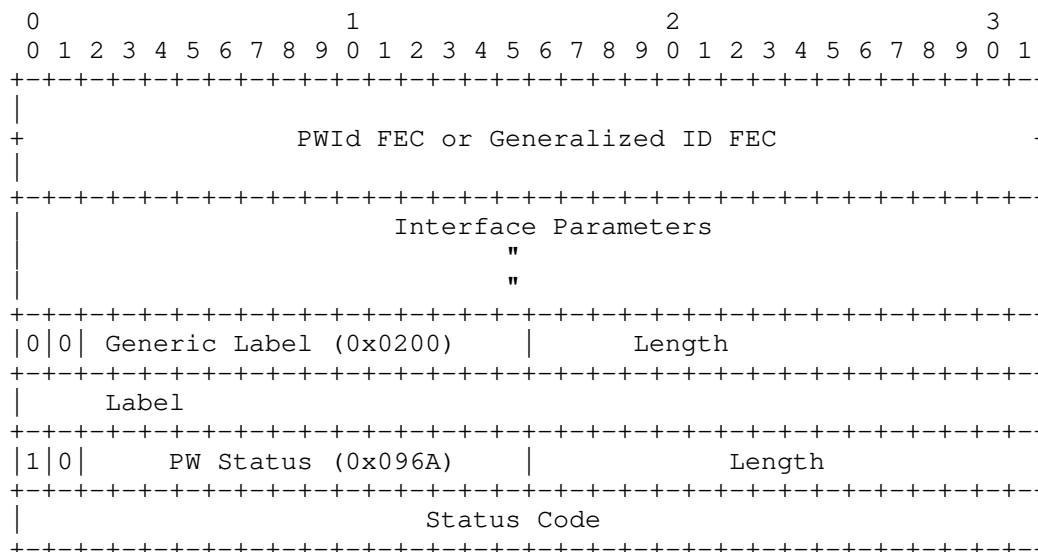
The Status TLV status code is set to 0x00000028, "PW status", to indicate that PW status follows. Since this notification does not refer to any particular message, the Message Id and Message Type fields are set to 0.

The PW FEC TLV SHOULD NOT include the interface parameter sub-TLVs, as they are ignored in the context of this message. When a PE's attachment circuit encounters an error, use of the PW Notification Message allows the PE to send a single "wild card" status message, using a PW FEC TLV with only the group ID set, to denote this change in status for all affected PW connections. This status message contains either the PW FEC TLV with only the group ID set, or else it contains the Generalized FEC TLV with only the PW Grouping ID TLV.

As mentioned above, the Group ID field of the Pwid FEC element, or the PW Grouping ID TLV used with the Generalized Pwid FEC element, can be used to send a status notification for all arbitrary sets of PWs. This procedure is OPTIONAL, and if it is implemented, the LDP Notification message should be as follows: If the Pwid FEC element is used, the PW information length field is set to 0, the PW ID field is not present, and the interface parameter sub-TLVs are not present. If the Generalized FEC element is used, the AGI, SAIL, and TAIL are not present, the PW information length field is set to 0, the PW Grouping ID TLV is included, and the Interface Parameters TLV is omitted. For the purpose of this document, this is called the "wild card PW status notification procedure", and all PEs implementing this design are REQUIRED to accept such a notification message but are not required to send it.

5.3.3. Pseudowire Status Negotiation Procedures

When a PW is first set up, the PEs MUST attempt to negotiate the usage of the PW status TLV. This is accomplished as follows: A PE that supports the PW Status TLV MUST include it in the initial Label Mapping message following the PW FEC and the interface parameter sub-TLVs. The PW Status TLV will then be used for the lifetime of the pseudowire. This is shown in the following diagram:



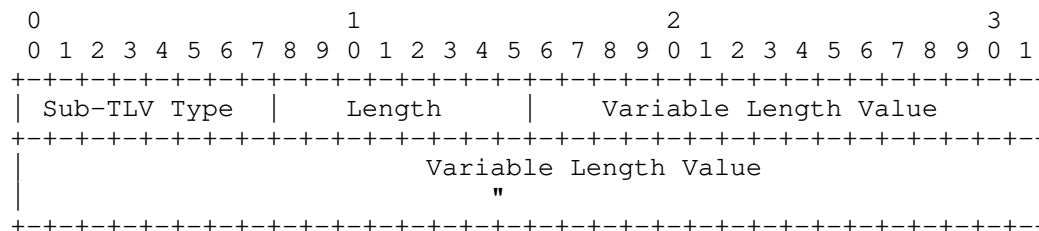
If a PW Status TLV is included in the initial Label Mapping message for a PW, then if the Label Mapping message from the remote PE for that PW does not include a PW status TLV, or if the remote PE does not support the PW Status TLV, the PW will revert to the label withdraw method of signaling PW status. Note that if the PW Status TLV is not supported by the remote peer, the peer will automatically ignore it, since the I (ignore) bit is set in the TLV. The PW Status TLV, therefore, will not be present in the corresponding FEC advertisement from the remote LDP peer, which results in exactly the above behavior.

If the PW Status TLV is not present following the FEC TLV in the initial PW Label Mapping message received by a PE, then the PW Status TLV will not be used, and both PEs supporting the pseudowire will revert to the label withdraw procedure for signaling status changes.

If the negotiation process results in the usage of the PW status TLV, then the actual PW status is determined by the PW status TLV that was sent within the initial PW Label Mapping message. Subsequent updates of PW status are conveyed through the notification message.

5.4. Interface Parameters sub-TLV

This field specifies interface-specific parameters. When applicable, it MUST be used to validate that the PEs and the ingress and egress ports at the edges of the circuit have the necessary capabilities to interoperate with each other. The field structure is defined as follows:



The interface parameter sub-TLV type values are specified in "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)" [RFC4446].

The Length field is defined as the length of the interface parameter including the parameter id and length field itself. Processing of the interface parameters should continue when unknown interface parameters are encountered, and they MUST be silently ignored.

- Interface MTU sub-TLV type

A 2 octet value indicating the MTU in octets. This is the Maximum Transmission Unit, excluding encapsulation overhead, of the egress packet interface that will be transmitting the decapsulated PDU that is received from the MPLS-enabled network. This parameter is applicable only to PWs transporting packets and is REQUIRED for these PW types. If this parameter does not match in both directions of a specific PW, that PW MUST NOT be enabled.

- Optional Interface Description string sub-TLV type

This arbitrary, and OPTIONAL, interface description string is used to send a human-readable administrative string describing the interface to the remote. This parameter is OPTIONAL, and is applicable to all PW types. The interface description parameter string length is variable, and can be from 0 to 80 octets. Human-readable text MUST be provided in the UTF-8 charset using the Default Language [RFC2277].

6. Control Word

6.1. PW types for which the control word is REQUIRED

The Label Mapping messages that are sent in order to set up these PWs MUST have c=1. When a Label Mapping message for a PW of one of these types is received and c=0, a Label Release message MUST be sent, with an "Illegal C-bit" status code. In this case, the PW will not be enabled.

6.2. PW types for which the control word is NOT mandatory

If a system is capable of sending and receiving the control word on PW types for which the control word is not mandatory, then each such PW endpoint MUST be configurable with a parameter that specifies whether the use of the control word is PREFERRED or NOT PREFERRED. For each PW, there MUST be a default value of this parameter. This specification does NOT state what the default value should be.

If a system is NOT capable of sending and receiving the control word on PW types for which the control word is not mandatory, then it behaves exactly as if it were configured for the use of the control word to be NOT PREFERRED.

If a Label Mapping message for the PW has already been received but no Label Mapping message for the PW has yet been sent, then the procedure is as follows:

- i. If the received Label Mapping message has c=0, send a Label Mapping message with c=0; the control word is not used.
- ii. If the received Label Mapping message has c=1, and the PW is locally configured such that the use of the control word is preferred, then send a Label Mapping message with c=1; the control word is used.
- iii. If the received Label Mapping message has c=1, and the PW is locally configured such that the use of the control word is not preferred or the control word is not supported, then act as if no Label Mapping message for the PW had been received (That is: proceed to the next paragraph).

If a Label Mapping message for the PW has not already been received (or if the received Label Mapping message had c=1 and either local configuration says that the use of the control word is not preferred or the control word is not supported), then send a Label Mapping message in which the c bit is set to correspond to the locally configured preference for use of the control word. (That is, set c=1 if locally configured to prefer the control word, and set c=0 if

locally configured to prefer not to use the control word or if the control word is not supported).

The next action depends on what control message is next received for that PW. The possibilities are as follows:

- i. A Label Mapping message with the same c bit value as specified in the Label Mapping message that was sent. PW setup is now complete, and the control word is used if c=1 but is not used if c=0.
- ii. A Label Mapping message with c=1, but the Label Mapping message that was sent has c=0. In this case, ignore the received Label Mapping message and continue to wait for the next control message for the PW.
- iii. A Label Mapping message with c=0, but the Label Mapping message that was sent has c=1. In this case, send a Label Withdraw message with a "Wrong C-bit" status code, followed by a Label Mapping message that has c=0. PW setup is now complete, and the control word is not used.
- iv. A Label Withdraw message with the "Wrong c-bit" status code. Treat as a normal Label Withdraw, but do not respond. Continue to wait for the next control message for the PW.

If at any time after a Label Mapping message has been received a corresponding Label Withdraw or Release is received, the action taken is the same as for any Label Withdraw or Release that might be received at any time.

If both endpoints prefer the use of the control word, this procedure will cause it to be used. If either endpoint prefers not to use the control word or does not support the control word, this procedure will cause it not to be used. If one endpoint prefers to use the control word but the other does not, the one that prefers not to use it has no extra protocol to execute; it just waits for a Label Mapping message that has c=0.

6.3. Control-Word Renegotiation by Label Request Message

The renegotiation process begins when the local PE has received the remote Label Mapping message with the C bit set to 0, and at the point when its use of control word is changed from NOT PREFERRED to PREFERRED. The following additional procedure will be carried out:

- i. The local PE MUST send a Label Release message to remote PE. If local PE has previously sent a Label Mapping message, it MUST send a Label Withdraw message to remote PE and wait until it has received a Label Release message from the remote PE. Note: the above-mentioned sending of the Label Release message and Label Withdraw message does not require a specific sequence.
- ii. The local PE MUST send a Label Request message to the peer PE, and then MUST wait until it receives a Label Mapping message containing the peer's current configured preference for use of control word.
- iii. After receiving the remote peer PE Label Mapping message with the C bit, the local PE MUST follow the procedures defined in the sections above, when sending its Label Mapping message.

Once the remote PE has successfully processed the Label Withdraw message and Label Release message, it will reset its use of control word with the locally configured preference. Then, the remote PE will send a Label Mapping message with locally configured preference for use of control word as a response to Label Request message, as specified in [RFC5036].

It should be noted that the local PE SHOULD wait for the above message-exchanging process to be finished before processing new request to change the configuration. The FEC element in the Label Request message should be the PE's local PW FEC element. As a response to the Label Request message, the peer PE should send a Label Mapping message with its own local PW FEC element. The Label Request message format and procedure is described in [RFC5036].

The diagram in Appendix A illustrates the above procedure.

6.4. LDP label Withdrawal procedures

As mentioned above, the Group ID field of the PWid FEC element, or the PW Grouping ID TLV used with the Generalized PWid FEC element, can be used to withdraw all PW labels associated with a particular PW group. This procedure is OPTIONAL, and if it is implemented, the LDP Label Withdraw message should be as follows: If the PWid FEC element is used, the PW information length field is set to 0, the PW ID field is not present, the interface parameter sub-TLVs are not present, and the Label TLV is not present. If the Generalized FEC element is

used, the AGI, SAIL, and TAIL are not present, the PW information length field is set to 0, the PW Grouping ID TLV is included, the Interface Parameters TLV is not present, and the Label TLV is not present. For the purpose of this document, this is called the "wild card withdraw procedure", and all PEs implementing this design are REQUIRED to accept such withdrawn message but are not required to send it. Note that the PW Grouping ID TLV only applies to PWs using the Generalized ID FEC element, while the Group ID only applies to PWid FEC element.

The interface parameter sub-TLVs, or TLV, MUST NOT be present in any LDP PW Label Withdraw or Label Release message. A wild card Label Release message MUST include only the group ID, or Grouping ID TLV. A Label Release message initiated by a PE router must always include the PW ID.

6.5. Sequencing Considerations

In the case where the router considers the sequence number field in the control word, it is important to note the following details when advertising labels.

6.5.1. Label Advertisements

After a label has been withdrawn by the output router and/or released by the input router, care must be taken not to advertise (re-use) the same released label until the output router can be reasonably certain that old packets containing the released label no longer persist in the MPLS-enabled network.

This precaution is required to prevent the imposition router from restarting packet forwarding with a sequence number of 1 when it receives a Label Mapping message that binds the same FEC to the same label if there are still older packets in the network with a sequence number between 1 and 32768. For example, if there is a packet with a sequence number=n, where n is in the interval [1,32768] traveling through the network, it would be possible for the disposition router to receive that packet after it re-advertises the label. Since the label has been released by the imposition router, the disposition router SHOULD be expecting the next packet to arrive with a sequence number of 1. Receipt of a packet with a sequence number equal to n will result in n packets potentially being rejected by the disposition router until the imposition router imposes a sequence number of n+1 into a packet. Possible methods to avoid this are for the disposition router always to advertise a different PW label, or for the disposition router to wait for a sufficient time before attempting to re-advertise a recently released label. This is only an issue when sequence number processing is enabled at the

disposition router.

6.5.2. Label Release

In situations where the imposition router wants to restart forwarding of packets with sequence number 1, the router shall 1) send to the disposition router a Label Release Message, and 2) send to the disposition router a Label Request message. When sequencing is supported, advertisement of a PW label in response to a Label Request message MUST also consider the issues discussed in the section on Label Advertisements.

7. IANA Considerations

7.1. LDP TLV TYPE

This document uses several new LDP TLV types; IANA already maintains a registry of name "TLV TYPE NAME SPACE" defined by RFC 5036. The following values are suggested for assignment:

TLV type	Description
=====	
0x096A	PW Status TLV
0x096B	PW Interface Parameters TLV
0x096C	Group ID TLV

7.2. LDP Status Codes

This document uses several new LDP status codes; IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by RFC 5036. The following values are suggested for assignment:

Range/Value	E	Description	Reference

0x00000024	0	Illegal C-Bit	[RFC4447]
0x00000025	0	Wrong C-Bit	[RFC4447]
0x00000026	0	Incompatible bit-rate	[RFC4447]
0x00000027	0	CEP-TDM mis-configuration	[RFC4447]
0x00000028	0	PW Status	[RFC4447]
0x00000029	0	Unassigned/Unrecognized TAI	[RFC4447]
0x0000002A	0	Generic Misconfiguration Error	[RFC4447]
0x0000002B	0	Label Withdraw PW Status Method	[RFC4447]

7.3. FEC Type Name Space

This document uses two new FEC element types, 0x80 and 0x81, from the registry "FEC Type Name Space" for the Label Distribution Protocol (LDP RFC 5036).

8. Security Considerations

This document specifies the LDP extensions that are needed for setting up and maintaining pseudowires. The purpose of setting up pseudowires is to enable Layer 2 frames to be encapsulated in MPLS and transmitted from one end of a pseudowire to the other. Therefore we treat the security considerations for both the data plane and the control plane.

8.1. Data-plane Security

With regard to the security of the data plane, the following areas must be considered:

- MPLS PDU inspection.
- MPLS PDU spoofing.
- MPLS PDU alteration.
- MPLS PSN protocol security.
- Access Circuit security.
- Denial of service prevention on the PE routers.

When an MPLS PSN is used to provide pseudowire service, there is a perception that security MUST be at least equal to the currently deployed Layer 2 native protocol networks that the MPLS/PW network combination is emulating. This means that the MPLS-enabled network SHOULD be isolated from outside packet insertion in such a way that it SHOULD NOT be possible to insert an MPLS packet into the network directly. To prevent unwanted packet insertion, it is also important to prevent unauthorized physical access to the PSN, as well as unauthorized administrative access to individual network elements.

As mentioned above, as MPLS-enabled network, should not accept MPLS packets from its external interfaces (i.e., interfaces to CE devices or to other providers' networks) unless the top label of the packet was legitimately distributed to the system from which the packet is being received. If the packet's incoming interface leads to a different SP (rather than to a customer), an appropriate trust relationship must also be present, including the trust that the other SP also provides appropriate security measures.

The three main security problems faced when using an MPLS-enabled network to transport PWs are spoofing, alteration, and inspection.

First, there is a possibility that the PE receiving PW PDUs will get a PDU that appears to be from the PE transmitting the PW into the PSN, but that was not actually transmitted by the PE originating the PW. (That is, the specified encapsulations do not by themselves enable the decapsulator to authenticate the encapsulator.) A second problem is the possibility that the PW PDU will be altered between the time it enters the PSN and the time it leaves the PSN (i.e., the specified encapsulations do not by themselves assure the decapsulator of the packet's integrity.) A third problem is the possibility that the PDU's contents will be seen while the PDU is in transit through the PSN (i.e., the specification encapsulations do not ensure privacy.) How significant these issues are in practice depends on the security requirements of the applications whose traffic is being sent through the tunnel, and how secure the PSN itself is.

8.2. Control-Plane Security

General security considerations with regard to the use of LDP are specified in section 5 of RFC 5036. Those considerations also apply to the case where LDP is used to set up pseudowires.

A pseudowire connects two attachment circuits. It is important to make sure that LDP connections are not arbitrarily accepted from anywhere, or else a local attachment circuit might get connected to an arbitrary remote attachment circuit. Therefore, an incoming LDP session request MUST NOT be accepted unless its IP source address is known to be the source of an "eligible" LDP peer. The set of eligible peers could be pre-configured (either as a list of IP addresses, or as a list of address/mask combinations), or it could be discovered dynamically via an auto-discovery protocol that is itself trusted. (Obviously, if the auto-discovery protocol were not trusted, the set of "eligible peers" it produces could not be trusted.)

Even if an LDP connection request appears to come from an eligible peer, its source address may have been spoofed. Therefore, some means of preventing source address spoofing must be in place. For example, if all the eligible peers are in the same network, source address filtering at the border routers of that network could eliminate the possibility of source address spoofing.

The LDP MD5 authentication key option, as described in section 2.9 of RFC 5036, MUST be implemented, and for a greater degree of security, it must be used. This provides integrity and authentication for the LDP messages and eliminates the possibility of source address spoofing. Use of the MD5 option does not provide privacy, but privacy of the LDP control messages is not usually considered important. As the MD5 option relies on the configuration of pre-

shared keys, it does not provide much protection against replay attacks. In addition, its reliance on pre-shared keys may make it very difficult to deploy when the set of eligible neighbors is determined by an auto-configuration protocol.

When the Generalized PWid FEC Element is used, it is possible that a particular LDP peer may be one of the eligible LDP peers but may not be the right one to connect to the particular attachment circuit identified by the particular instance of the Generalized PWid FEC element. However, given that the peer is known to be one of the eligible peers (as discussed above), this would be the result of a configuration error, rather than a security problem. Nevertheless, it may be advisable for a PE to associate each of its local attachment circuits with a set of eligible peers rather than have just a single set of eligible peers associated with the PE as a whole.

9. Acknowledgments

The authors wish to acknowledge the contributions of Vach Kompella, Vanson Lim, Wei Luo, Himanshu Shah, and Nick Weeds.

10. Normative References

- [RFC2119] Bradner S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, March 1997
- [RFC5036] "LDP Specification." L. Andersson, P. Ed. Minei, I. Ed. B. Thomas. January 2001. RFC5036
- [RFC3032] "MPLS Label Stack Encoding", E. Rosen, Y. Rekhter, D. Tappan, G. Fedorkow, D. Farinacci, T. Li, A. Conta. RFC3032
- [RFC4446] "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)" L. Martini RFC4446 , April 2006

11. Informative References

- [RFC4842] "Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH) Circuit Emulation over Packet (CEP)", A. Malis, P. Pate, R. Cohen, Ed., D. Zelig, RFC4842, April 2007
- [RFC4553] "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", Vainshtein A. Ed. Stein, Ed. YJ. RFC4553, June 2006

- [RFC4619] "Encapsulation Methods for Transport of Frame Relay over Multiprotocol Label Switching (MPLS) Networks", Martini L. Ed. C. Kawa Ed. A. Malis Ed. RFC4619, September 2006
- [RFC4717] "Encapsulation Methods for Transport of Asynchronous Transfer Mode (ATM) over MPLS Networks", Martini L. Jayakumar J. Bocci M. El-Aawar N. Brayley J. Koleyni G. RFC4717 December 2006
- [RFC4618] "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) Frames over MPLS Networks", Martini L. Rosen E. Heron G. Malis A. RFC4618 September 2006
- [RFC4448] "Encapsulation Methods for Transport of Ethernet over MPLS Networks", Martini L. Ed. Rosen E. El-Aawar N. Heron G. RFC4448 April 2006.
- [SDH] American National Standards Institute, "Synchronous Optical Network Formats," ANSI T1.105-1995.
- [ITU] ITU Recommendation G.707, "Network Node Interface For The Synchronous Digital Hierarchy", 1996.
- [RFC3985] "PWE3 Architecture" Bryant, et al., RFC3985.
- [RFC2277] Alvestrand, H., "IETF Policy on Character Sets and Languages", BCP 18, RFC 2277, January 1998.

12. Author Information

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

Nasser El-Aawar
Level 3 Communications, LLC.
1025 Eldorado Blvd.
Broomfield, CO, 80021
e-mail: nna@level3.net

Giles Heron
Tellabs
Abbey Place
24-28 Easton Street
High Wycombe
Bucks
HP11 1NT
UK
e-mail: giles.heron@tellabs.com

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA 01719
e-mail: erosen@cisco.com

Dan Tappan
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA 01719
e-mail: tappan@cisco.com

Toby Smith
Network Appliance
800 Cranberry Woods Dr.
Cranberry Twp, PA 16066
e-mail: Toby.Smith@netapp.com

13. Additional Contributing Authors

Dimitri Stratton Vlachos
Mazu Networks, Inc.
125 Cambridgepark Drive
Cambridge, MA 02140
e-mail: d@mazunetworks.com

Jayakumar Jayakumar,
Cisco Systems Inc.
225, E.Tasman, MS-SJ3/3,
San Jose, CA, 95134
e-mail: jjayakum@cisco.com

Alex Hamilton,
Cisco Systems Inc.
285 W. Tasman, MS-SJCI/3/4,
San Jose, CA, 95134
e-mail: tahamilt@cisco.com

Steve Vogelsang
ECI Telecom
Omega Corporate Center
1300 Omega Drive
Pittsburgh, PA 15205
e-mail: stephen.vogelsang@ecitele.com

John Shirron
ECI Telecom
Omega Corporate Center
1300 Omega Drive
Pittsburgh, PA 15205
e-mail: john.shirron@ecitele.com

Andrew G. Malis
Tellabs
90 Rio Robles Dr.
San Jose, CA 95134
e-mail: Andy.Malis@tellabs.com

Vinai Sirkay
Reliance Infocomm
Dhirubai Ambani Knowledge City
Navi Mumbai 400 709
e-mail: vinai@sirkay.com

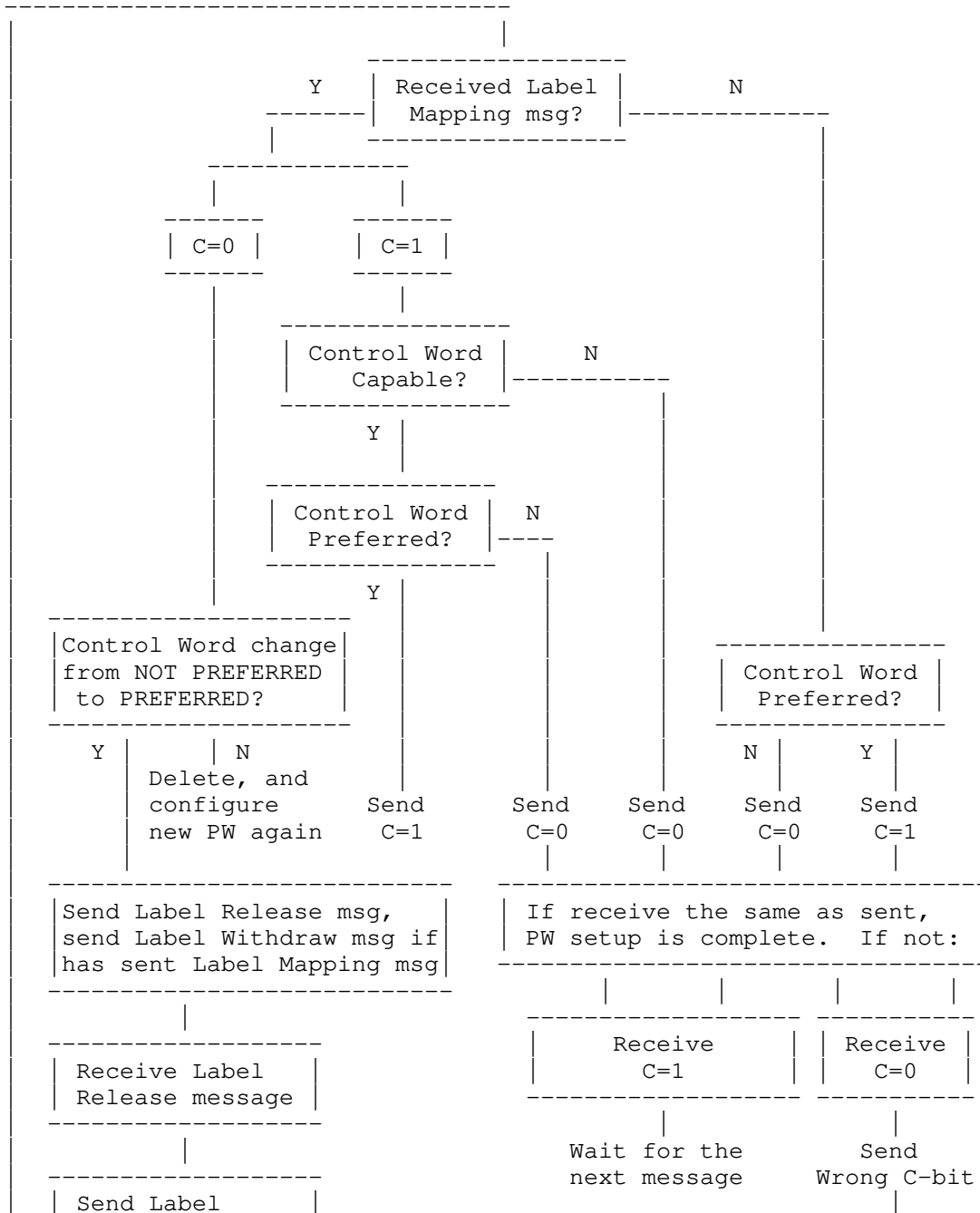
Vasile Radoaca
Nortel Networks
600 Technology Park
Billerica MA 01821
e-mail: vasile@nortelnetworks.com

Chris Liljenstolpe
Alcatel
11600 Sallie Mae Dr.
9th Floor
Reston, VA 20193
e-mail: chris.liljenstolpe@alcatel.com

Dave Cooper
Global Crossing
960 Hamlin Court
Sunnyvale, CA 94089
e-mail: dcooper@gblix.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
e-mail: kireeti@juniper.net

Ap A C-bit Handling Procedures Diagram



```
| | Request message |  
|-----|  
| |  
|-----|
```

Send Label
Mapping message

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.

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: August 2013

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: August 2013

H. Long
M. Ye
Huawei Technologies Co., Ltd.
February 18, 2013

Multi-Segment Pseudowire Signaling with Availability Information
draft-long-pwe3-ms-pw-availability-01.txt

Abstract

This document describes a signaling mechanism for setting up a multi-segment pseudowire between different domains in case that at least one domain has the feature that bandwidth capacity is variable for different availability values. The signaling mechanism is an extension on the multi-segment pseudowire signaling [DYN-MS-PW].

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on August 18, 2013.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. LDP Signaling	3
3. Security Considerations.....	4
4. IANA Considerations	4
5. References	4
5.1. Normative References.....	4
5.2. Informative References.....	4
6. Acknowledgments	5

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. Introduction

[RFC5659] describes the architecture of multi-segment pseudowire. The LDP signaling mechanism for setting up multi-segment pseudowire is described by [DYN-MS-PW]. The mechanism introduces a pseudowire bandwidth TLV for PSN tunnel selection.

In some networks, there may be some links with variable bandwidth. For example, in mobile backhaul network, microwave links are very popular for providing connection of last hops. In case of heavy rain, to maintain the link connectivity, the microwave link will lower the modulation format since demodulating lower modulation format need lower signal-to-noise ratio (SNR). This is called adaptive modulation technology. However, lower modulation format also means lower link bandwidth. Similarly the copper links may change their link bandwidth due to external interference.

The parameter, availability [G.827, F.1703], is often used to describe the link capacity during network planning. A link may provide different bandwidth for different availability requirement. In this case, a LSP across the links should include multiple bandwidth portions with different availability guarantee. As a result, only PW bandwidth requirement information is not enough for the multi-segment pseudowire setup.

An optional PW availability TLV is introduced in this document for providing enough information for S-PE to selecting PSN tunnel.

2. LDP Signaling

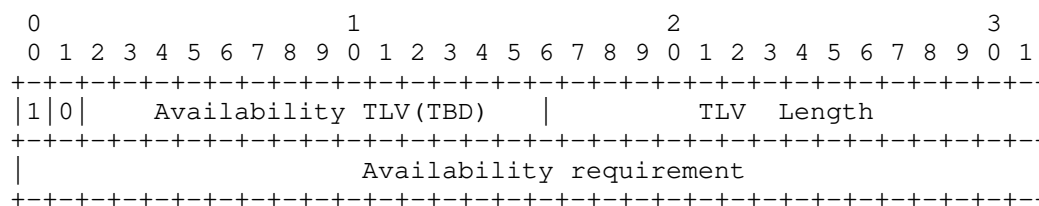


Figure 1 - PW Availability TLV

The optional PW availability TLV described by Figure 1 includes an availability requirement field which is a 32-bit floating point number. This field describes the availability requirement for the multi-segment pseudowire. The value can be inherited from the availability requirement of the emulated service. It should be smaller than 1.

When an S/T-PE receives a PW availability TLV and a PW bandwidth TLV, it will search for an PSN tunnel which can satisfy the bandwidth requirement with the availability requirement. Once the PW next hop is selected, the S/T-PE MUST request the appropriate resources which can guarantee the required availability from the PSN.

In the case where PSN resources towards the next hop cannot satisfy the bandwidth request with the specified the availability requirement; the following procedure MUST be followed:

- i. The PSN MAY allocate more QoS resources, e.g. Bandwidth with required availability guarantee, 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.

3. Security Considerations

This document does not introduce new security considerations.

4. IANA Considerations

TBD.

5. References

5.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5659] Bocci, M., Bryant, S., "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.
- [DYN-MS-PW] Martini, L., Bocci, M., Balus, F., "Dynamic Placement of Multi Segment Pseudo Wires", draft-ietf-pwe3-dynamic-ms-pw-16, October, 2010
- [G.827] ITU-T Recommendation, "Availability performance parameters and objectives for end-to-end international constant bit-rate digital paths", September, 2003.
- [F.1703] ITU-R Recommendation, "Availability objectives for real digital fixed wireless links used in 27 500 km hypothetical reference paths and connections", January, 2005.

5.2. Informative References

6. Acknowledgments

Authors' Addresses

Hao Long
Huawei Technologies Co., Ltd.
No.1899,Xiyuan Avenue, Hi-tech Western District
Chengdu 611731, P.R.China

Email: longhao@huawei.com

Min Ye
Huawei Technologies Co., Ltd.
No.1899,Xiyuan Avenue, Hi-tech Western District
Chengdu 611731, P.R.China

Email: amy.yemin@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: August 12, 2013

Y. Shen, Ed.
Juniper Networks
R. Aggarwal
Arktan, Inc
W. Henderickx
Alcatel-Lucent
February 8, 2013

PW Endpoint Fast Failure Protection
draft-shen-pwe3-endpoint-fast-protection-03

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 be used to reduce traffic loss before a global repair mechanism reacts to the failure or 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 August 12, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Specification of Requirements	4
3. Reference Models and Failure Cases	4
3.1. Single-Segment PW	5
3.2. Multi-Segment PW	7
4. Theory of Operation	8
4.1. Local Repair and Protector	8
4.2. Context Identifier	11
4.2.1. Semantics	11
4.2.2. Advertisement and Path Computation	12
4.3. Protection Models	12
4.3.1. Co-located Protector	12
4.3.2. Centralized Protector	14
4.4. Transport Tunnel	15
4.5. Bypass Tunnel	16
4.6. Forwarding State on Protector	16
4.6.1. Co-located Protector	17
4.6.2. Centralized Protector	18
4.7. PW Label Distribution from Primary PE to Protector	20
4.7.1. Protection FEC Element Encoding for PWid	22
4.7.2. Protection FEC Element Encoding for Generalized PWid	23
4.8. PW Label Distribution from Backup PE to Protector	24
4.9. Revertive Behavior	25
5. IANA Considerations	26
6. Security Considerations	26
7. Acknowledgements	26
8. References	26
8.1. Normative References	26
8.2. Informative References	28
Authors' Addresses	28

1. Introduction

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

In order to protect a layer-2 service against network failures, it is necessary to protect every link and node along the entire data path. From the perspective of 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, it is also desirable that each of these components is protected by a fast protection mechanism based on local repair. Such a mechanism generally involves a bypass path that is pre-computed and pre-installed on a router 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. When the failure happens, 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) and IP/LDP fast-reroute (RFC 5714 and RFC 5286). However, there is a lack of such protection against egress AC failure, egress (T-)PE failure, and S-PE failure. In these cases, service restoration has to rely on 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 failures and restoring traffic.

This document intends to serve the exact need for the above. 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 a SS-PW; Node failure of 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 the failure. Compared with the existing global repair and control plane repair mechanisms, this mechanism can provide faster restoration. However, it is intended to complement those mechanisms, rather than replacing them in any way.

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 a global repair mechanism. In this document, the fast protection mechanism will also take advantage of them for local repair purposes. This SHALL enable local repair and global repair to work in tandem to achieve broader scope and better performance for protection.

3.1. Single-Segment PW

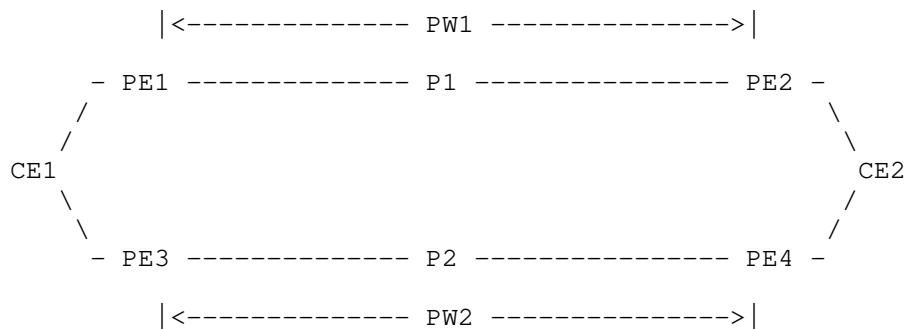


Figure 1

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

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

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

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

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

for the traffic going from CE1 to CE2:

Primary PW: PW1

Primary PE: PE2

Primary AC: CE2-PE2

Backup PW: PW2

Backup PE: PE4

Backup AC: CE2-PE4

In this case, an egress AC failure refers to the failure of the 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 the traffic when CE1 and CE2 switches traffic to PW2 during global repair, or when a local repair takes effect, as described later in this document.

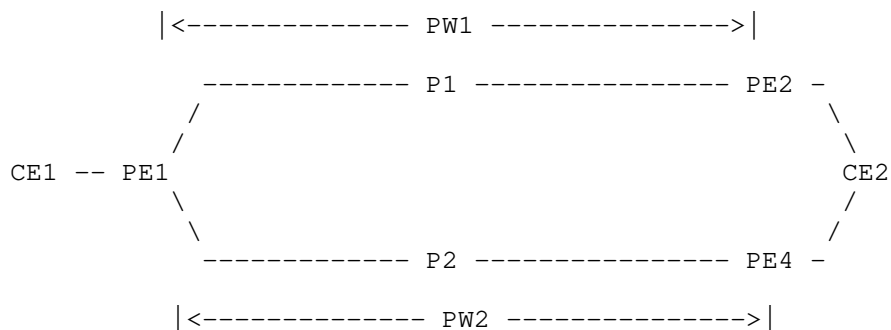


Figure 2

Figure 2 shows another possible scenario, where CE1 is single-homed to PE1, while CE2 remains multi-homed to PE2 and PE4. From the perspective of egress protection for the traffic from CE1 to CE2, this topology is not 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

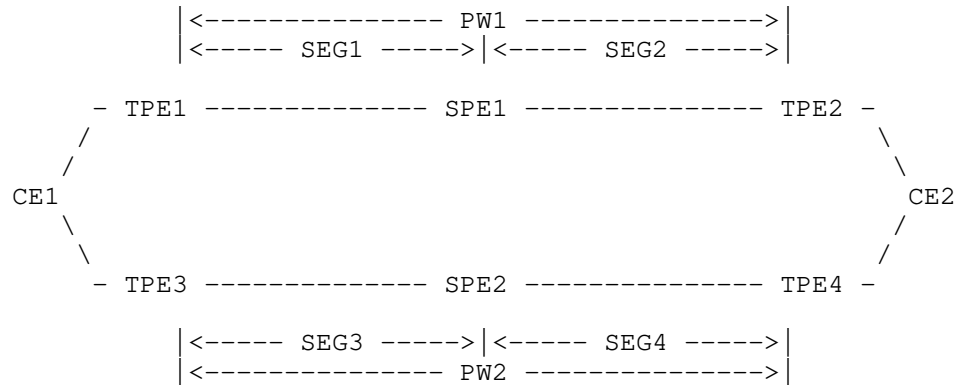


Figure 3

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 that the target CE is multi-homed to a primary PE and a backup PE, and there is a backup PW in the network. In S-PE node protection, it also assumes that there is a backup S-PE on the backup PW.

4.1. Local Repair and Protector

The mechanism relies on local repair to be performed by routers 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 has the property that it is not affected by the topology changes caused by the failure. Upon detecting the failure, the PLR MUST invoke the bypass route and forward PW traffic to the protector through the bypass tunnel. The protector MUST in turn send the traffic to the target CE, which may or may not be directly attached to the protector. This procedure is referred to as local repair.

Different routers may serve as PLR and protector in different

scenarios.

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

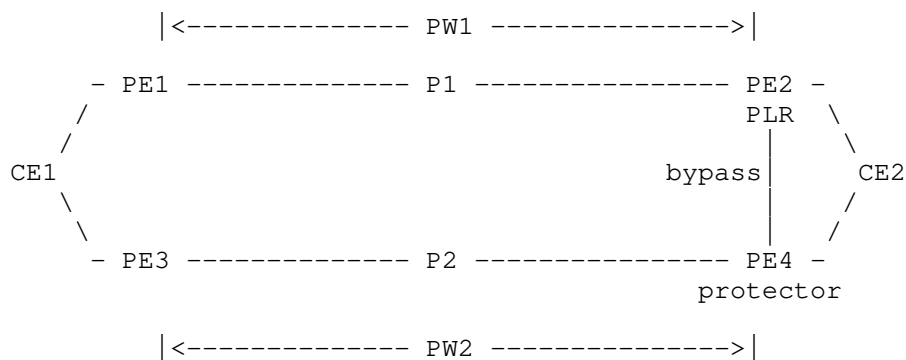


Figure 4

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

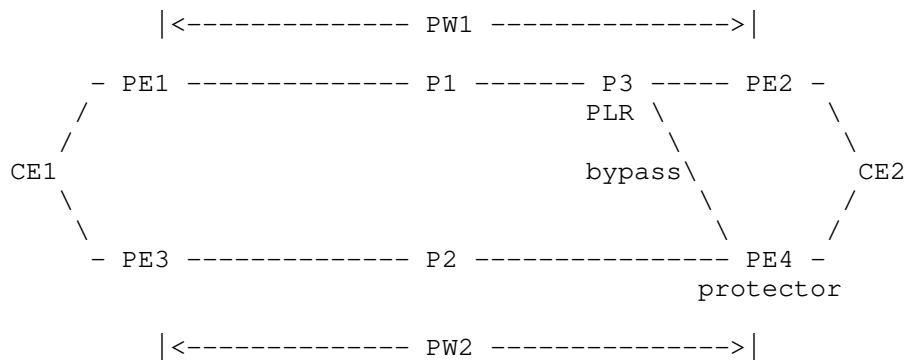


Figure 5

- 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).

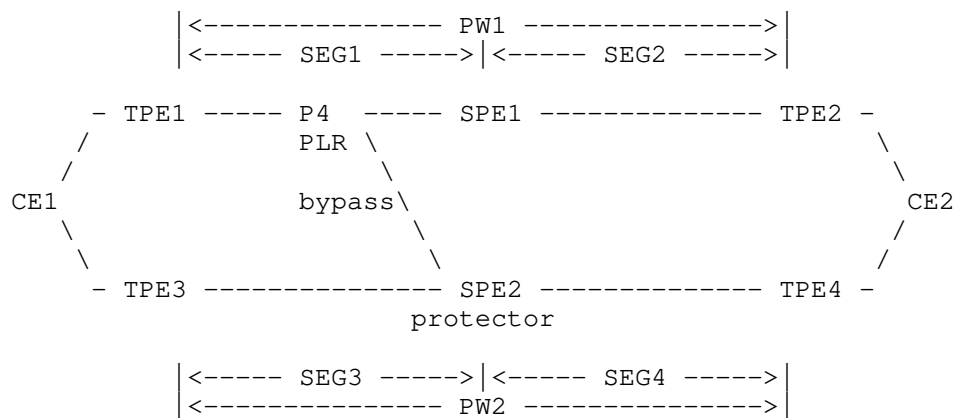


Figure 6

In all scenarios, when a PLR forwards traffic through a bypass tunnel to a protector, it keeps the label of the primary PW intact. This obviates the need for the PLR to maintain forwarding state on a per-PW basis, and allows the bypass tunnel to protect multiple primary PWs.

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

In the above examples, the protectors are backup (S-)PEs. A protector may also be a dedicated router that assumes such a role. In this case, the protector may not be the backup (S-)PE of a given primary PW. During local repair, a PLR still sends traffic to the protector through a bypass tunnel. The protector then sends the traffic to the backup (S-)PE, which finally sends the traffic to the target CE via a backup AC or a backup PW segment. More detail will be described in Section 4.3.

A protector MAY protect PWs for one or multiple primary PEs. The protector MUST maintain a separate label space for each primary PE. Likewise, the PWs of 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}.

4.2. Context Identifier

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

For a primary PW, its ingress PE MUST set up a transport tunnel with destination as the context identifier of the {primary PE, protector}, rather than an IP address of the primary PE. This enables the transport tunnel to follow a path to the primary PE, and also indicates the protector to the PLR(s) along the path.

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

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

A bypass tunnel's destination MUST be set to the context identifier as well, rather than an IP address of the protector. Therefore, the bypass tunnel (either MPLS tunnel label or IP tunnel destination address) can indicate the label space to 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 a transport tunnel and a bypass tunnels demands that the context identifier be advertised by IGP (OSPF or ISIS) in the routing domain and/or the TE domain, as an address reachable via both the primary PE and the protector. This imposes the following requirements on path computation for these tunnels.

- o For the transport tunnel, the ingress PE MUST choose the primary PE as the actual endpoint.
- o For the bypass tunnel, the PLR MUST choose the protector as the actual endpoint. The path MUST avoid the primary PE, with the exception of an egress AC protection bypass tunnel, where the PLR itself is the primary PE.

The detail of how IGP may advertise a context identifier is independent of the protection mechanism, and therefore out of the scope of this document. Some possible approaches are described by [LSP-EGRESS-PROTEC]. The ultimate goal is to allow CSPF (constrained shortest path first), LFA (loop free alternate; RFC 5286) and MRT (maximally redundant trees; [IP-LDP-FRR-MRT]) to compute the expected paths for the transport and bypass tunnels.

4.3. Protection Models

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

4.3.1. Co-located Protector

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

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

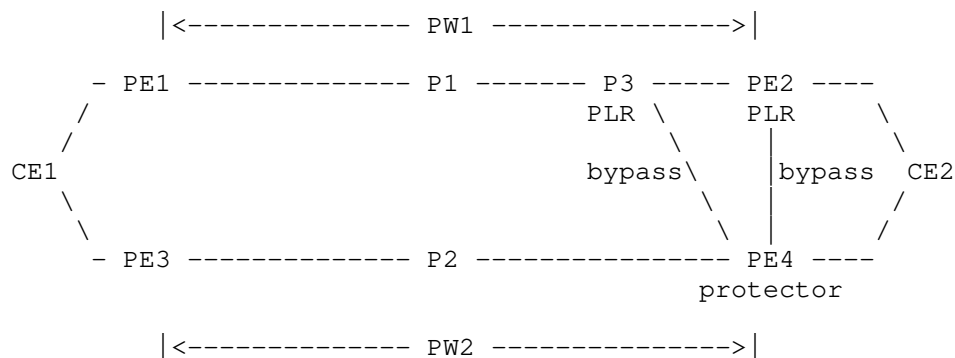


Figure 7

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

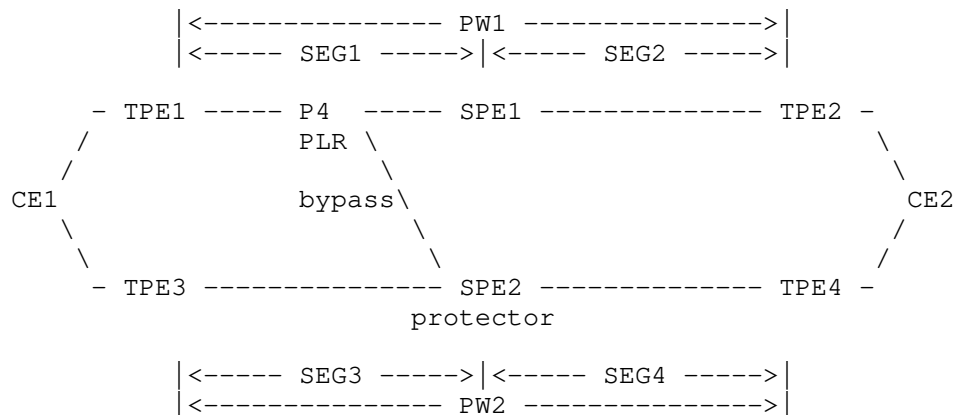


Figure 8

In the co-located protector model, the number of context identifiers needed by a network is the number of distinct {primary PE, backup PE} pairs. Therefore, 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 protects PWs for multiple primary PEs. In egress AC protection and egress PE node protection, the protector MAY or MAY NOT be a backup PE with a direct connection to the target CE. In S-PE node protection, the protector MAY or MAY NOT be a backup S-PE of the backup PW.

In egress AC protection and egress PE node protection, when the protector receives traffic from the PLR, if the protector has a direct connection (i.e. backup AC) to the CE, it MUST forward the traffic to the CE via the backup AC, which is similar to Figure 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.

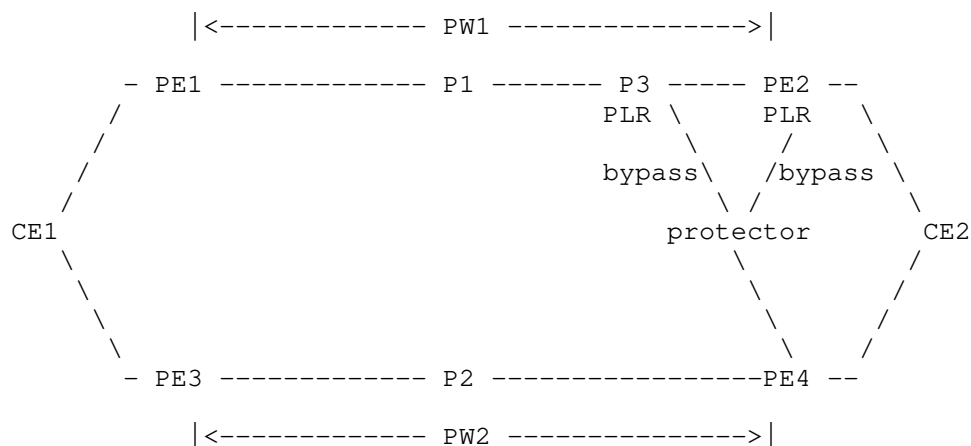


Figure 9

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.

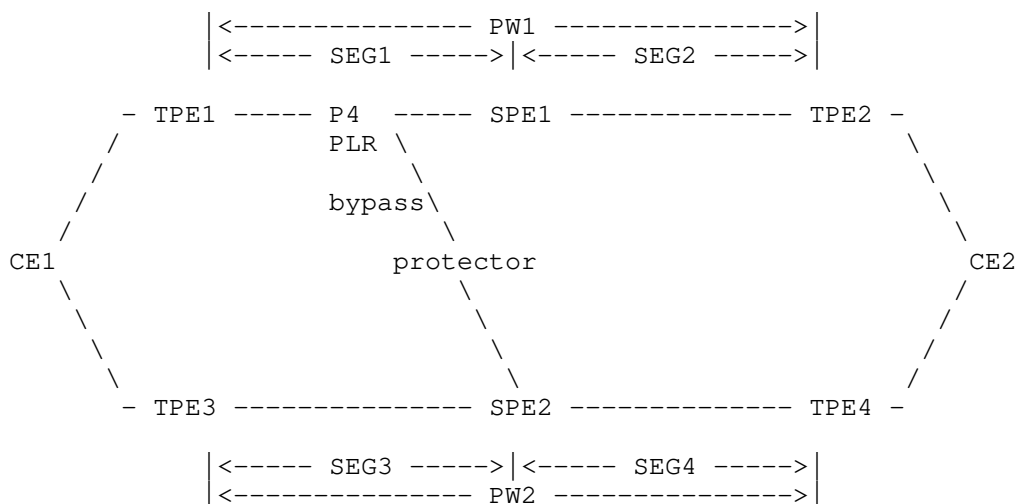


Figure 10

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

4.4. Transport Tunnel

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

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 PE, if the primary PE advertises the context identifier as "third party next hop" in an IPv4/v6 Interface_ID TLV (RFC 3471, RFC 3472) in 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 the protector. The destination of the bypass tunnel MUST be the context identifier (Section 4.2.1). The PLR may learn the context identifier from the destination address from 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 forwards traffic to the protector through the bypass tunnel with PW label intact. 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, i.e. an upstream assigned label. To perform such 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 for the bypass tunnel, and use this label as the context. If the bypass tunnel is an IP tunnel, the protector can decide the context based on the context identifier carried as destination address in the IP header.

A bypass tunnel MUST have the property that it is not affected by the topology change caused by the failure that it protects against. Therefore, it can be used to transmit traffic during the convergence of control plane protocols and the delay of global repair. It will remain effective, until the traffic is moved to another fully functional egress AC, PW or transport tunnel.

4.6. Forwarding State on Protector

A protector MUST learn PW labels from all the primary PEs that it protects (Section 4.7), and maintain the PW labels in a separate label space for each primary PE. In the control plane, a primary PE's label space is identified by the context identifier of the {primary PE, protector}. In the forwarding plane, the label space is indicated by bypass tunnels that are destined for the context identifier.

4.6.1. Co-located Protector

In Figure 11, 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.

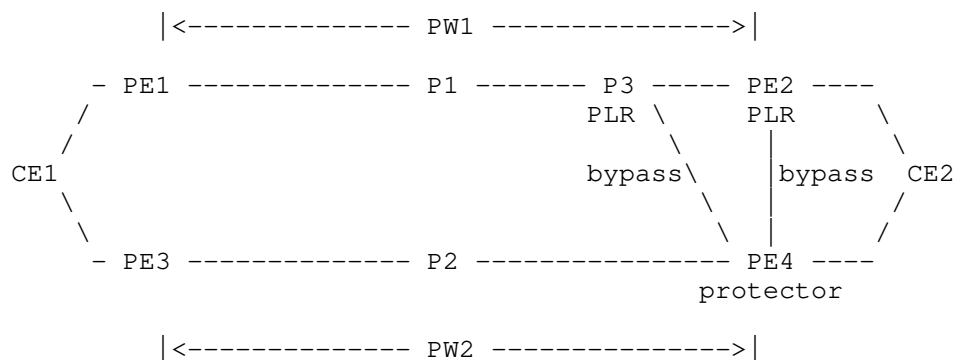


Figure 11

In Figure 12, 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.

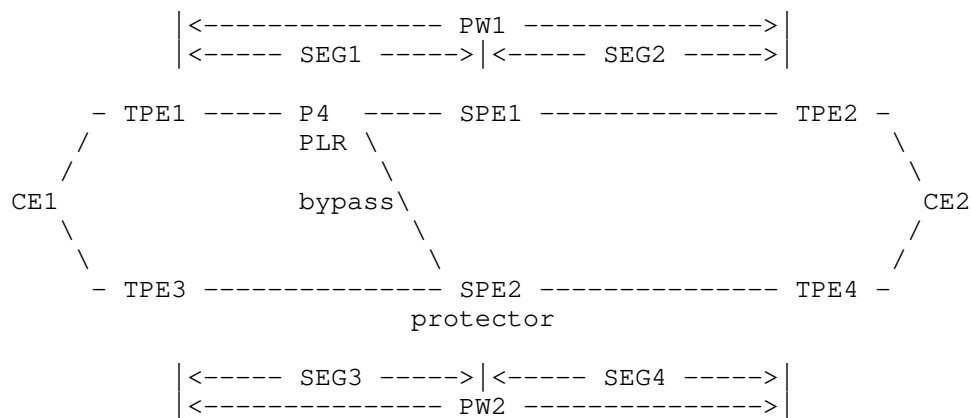


Figure 12

4.6.2. Centralized Protector

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

In Figure 13, 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.

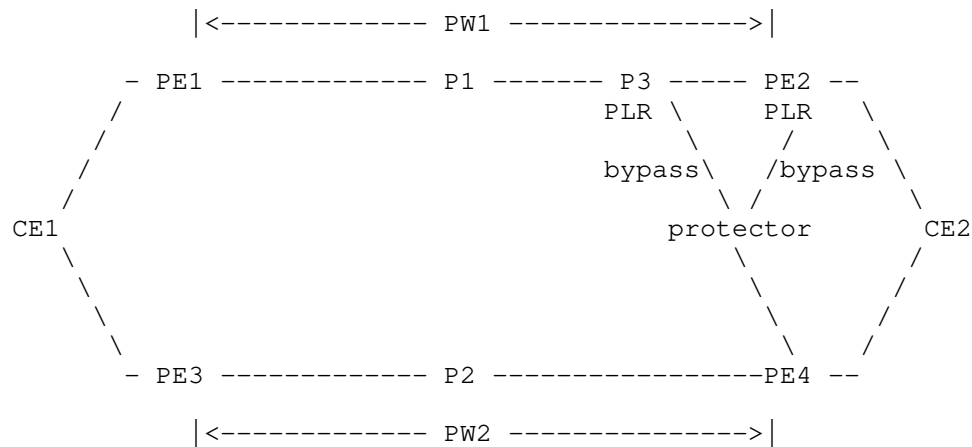


Figure 13

In Figure 14, 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.

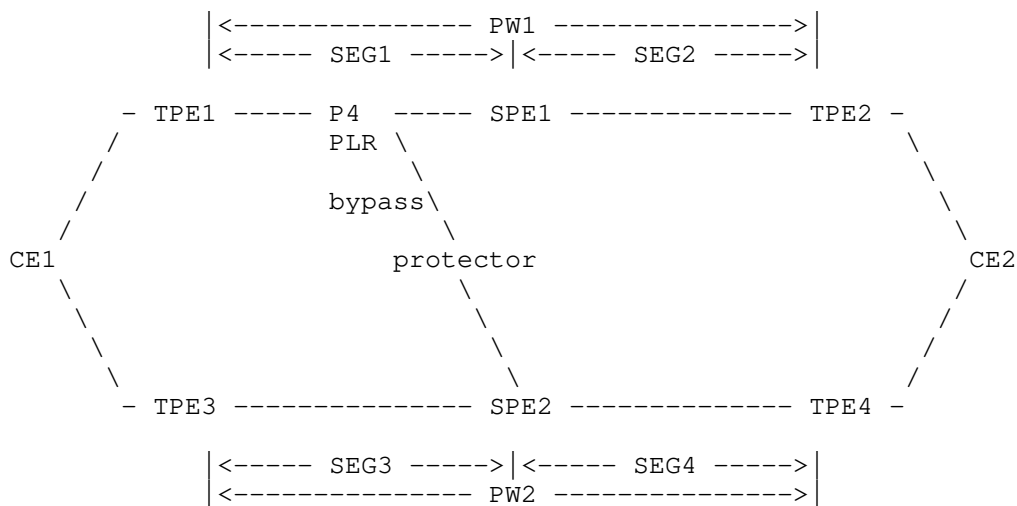


Figure 14

4.7. PW Label Distribution from Primary PE to Protector

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

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

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

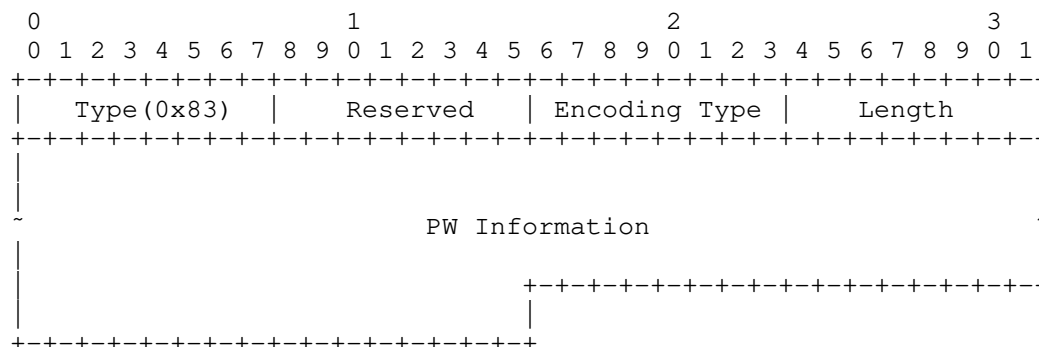


Figure 15

Type of format that PW Information field is encoded.

Length of PW Information field in octets.

Field of variable length that specifies a PW

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

4.7.1. Protection FEC Element Encoding for PWid

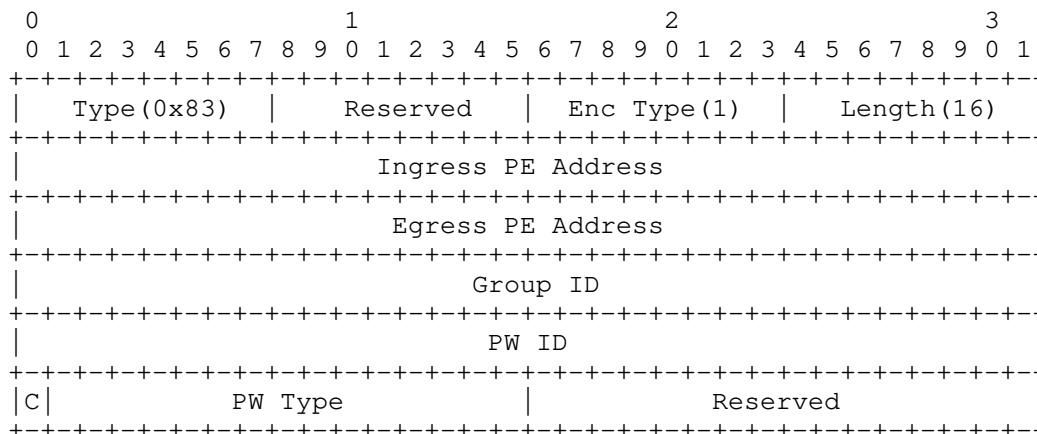


Figure 16

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Group ID

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

- PW ID

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

- Control word bit (C)

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

- PW Type

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

4.7.2. Protection FEC Element Encoding for Generalized PWid

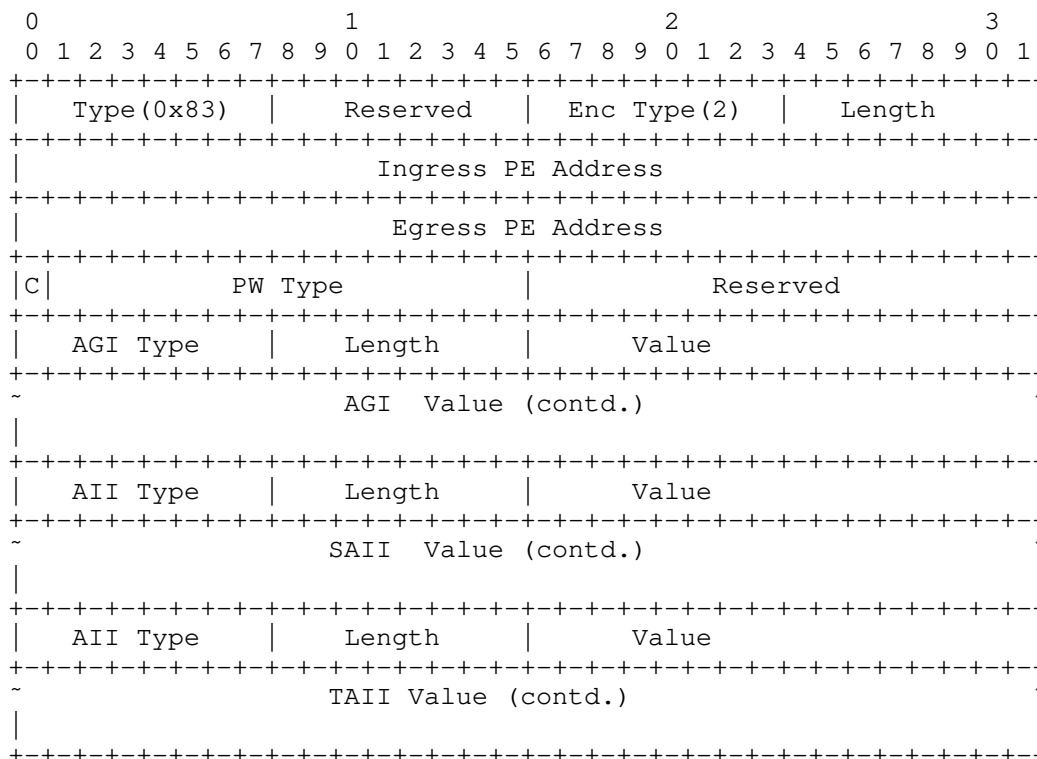


Figure 17

- Ingress PE Address

IP address of the ingress PE of PW.

- Egress PE Address

IP address of the egress PE of PW.

- Control word bit (C)

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

- PW Type

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

- AGI Type, Length, Value, AGI Value

Attachment Group Identifier of PW.

- SAII Type, Length, Value, SAII Value

Source Attachment Individual Identifier of PW.

- TAII Type, Length, Value, TAII Value

Target Attachment Individual Identifier of PW.

4.8. PW Label Distribution from Backup PE to Protector

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

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

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

4.9. Revertive Behavior

After local repair takes effect at a PLR, there are three strategies for restoring traffic to a fully working PW.

- o Global revertive mode

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

- o Control plane revertive mode

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

- o Local revertive mode

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

The fast protection mechanism in this document SHOULD always be used in tandem with the globally revertive mode. Particularly in the case of egress (S-)PE failure, if the ingress PE or the protector loses communication with the (S-)PE for an extensive period of time, the LDP session between them may go down. Consequently, the ingress PE may bring down the primary PW, or the protector may delete the forwarding entry of the primary PW label from the label space. In either case, the service will be disrupted. In other words, although the fast protection can temporarily repair traffic, control plane states may eventually time out if the failure persists. Therefore, it is recommended that the global revertive mode SHOULD always be

established in advance, so that traffic can be moved to a fully working backup PW shortly after the local repair.

The control plane revertive mode may happen as part of the convergence of control plane protocols. However, 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.

5. IANA Considerations

IANA maintains a registry of LDP FECs at the registry "Label Distribution Protocol" in the sub-registry called "Forwarding Equivalence Class (FEC) Type Name Space".

This document defines a new LDP Protection FEC Element in Section 4.7. IANA has assigned the type value 0x83 to it.

6. Security Considerations

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

7. Acknowledgements

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

8. References

8.1. Normative References

- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC5659] Bocci, M. and S. Bryant, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.

- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC6389] Aggarwal, R. and J.L. Le Roux, "MPLS Upstream Label Assignment for LDP", RFC 6389, November 2011.

[LSP-EGRESS-PROTEC]

Jeganathan, J., Gredler, H., and Y. Shen, "RSVP-TE LSP Egress Fast Protection",
draft-minto-rsvp-lsp-egress-fast-protection (work in progress), 2012.

[IP-LDP-FRR-MRT]

Atlas, A. and R. Kebler, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees",
draft-ietf-rtgwg-mrt-frr-architecture (work in progress), 2011.

8.2. Informative References

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

Authors' Addresses

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

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

Rahul Aggarwal
Arktan, Inc

Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.be

MPLS
Internet-Draft
Intended status: Informational
Expires: August 16, 2013

C. Villamizar, Ed.
OCCNC
K. Kompella
Contrail Systems
S. Amante
Level 3 Communications, Inc.
A. Malis
Verizon
C. Pignataro
Cisco
February 12, 2013

MPLS Forwarding Compliance and Performance Requirements
draft-villamizar-mpls-forwarding-01

Abstract

This document provides guidelines for implementors regarding MPLS forwarding and a basis for evaluations of forwarding implementations. Guidelines cover many aspects of MPLS forwarding. Topics are highlighted where implementors might potentially overlook practical requirements which are unstated or underemphasized or are optional for conformance to RFCs.

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 August 16, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	4
1.1. Use of Requirements Language	4
1.2. Apparent Misconceptions	4
1.3. Target Audience	5
2. Forwarding Issues	6
2.1. Forwarding Basics	6
2.1.1. MPLS Reserved Labels	7
2.1.2. MPLS Differentiated Services	7
2.1.3. Time Synchronization	8
2.1.4. Uses of Multiple Label Stack Entries	8
2.1.5. MPLS Link Bundling	9
2.1.6. MPLS Hierarchy	9
2.1.7. MPLS Fast Reroute (FRR)	10
2.1.8. Pseudowire Encapsulation	10
2.1.8.1. Pseudowire Sequence Number	11
2.1.9. Layer-2 and Layer-3 VPN	12
2.2. MPLS Multicast	12
2.3. Packet Rates	13
2.4. MPLS Multipath Techniques	14
2.4.1. Pseudowire Control Word	15
2.4.2. Large Microflows	16
2.4.3. Pseudowire Flow Label	16
2.4.4. MPLS Entropy Label	16
2.4.5. Fields Used for Multipath	17
2.4.5.1. MPLS Fields in Multipath	17
2.4.5.2. IP Fields in Multipath	19
2.4.5.3. Fields Used in Flow Label	20
2.4.5.4. Fields Used in Entropy Label	20
2.5. MPLS-TP and UHP	21
2.6. OAM and DoS Protection	21
2.6.1. DoS Protection	21
2.6.2. MPLS OAM	23
2.6.3. Pseudowire OAM	24
2.6.4. MPLS-TP OAM	25
2.6.5. MPLS OAM and Layer-2 OAM Interworking	26
2.6.6. Extent of OAM Support by Hardware	26

2.7. Number and Size of Flows	27
3. Questions for Suppliers	28
4. Forwarding Compliance and Performance Testing	32
5. Acknowledgements	37
6. IANA Considerations	37
7. Security Considerations	37
8. References	38
8.1. Normative References	38
8.2. Informative References	39
Appendix A. Organization of References Section	43
Authors' Addresses	43

1. Introduction

The initial purpose of this document was to address concerns raised on the MPLS WG mailing list about shortcomings in implementations of MPLS forwarding. Documenting existing misconceptions and potential pitfalls might potentially avoid repeating past mistakes. The document has grown to address a broad set of forwarding requirements.

1.1. Use of Requirements Language

This document is informational. The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are used only where the requirement is specified in an existing RFC. These keywords SHOULD be interpreted as described in RFC 2119 [RFC2119].

Advice given in this document does not use the upper case RFC 2119 keywords, except where explicitly noted that the keywords indicate that operator experiences indicate a requirement, but there are no existing RFC requirements. Such advice may be ignored by implementations. Similarly, implementations not claiming conformance to specific RFCs may ignore the requirements of those RFCs. In both cases, implementators may be doing so at their own peril.

1.2. Apparent Misconceptions

In early generations of forwarding silicon (which might now be behind us), there apparently were some misconceptions about MPLS. The following statements provide clarifications.

1. There are practical reasons to have more than one or two labels in an MPLS label stack. Under some circumstances the label stack can become quite deep. See Section 2.1.
2. The label stack MUST be considered to be arbitrarily deep. Section 3.27.4. "Hierarchy: LSP Tunnels within LSPs" of RFC 3031 [RFC3031] states "The label stack mechanism allows LSP tunneling to nest to any depth." If a the bottom of the label stack cannot be found, but sufficient number of labels exist to forward, an LSR MUST forward the packet. An LSR MUST NOT assume the packet is malformed unless the end of packet is found before bottom of stack. See Section 2.1.
3. In networks where deep label stacks are encountered, they are not rare. Full packet rate performance is required regardless of label stack depth, except where multiple POP operations are required. See Section 2.1.

4. Research has shown that long bursts of short packets with 40 byte or 44 byte IP payload sizes in these bursts are quite common. This is due to TCP ACK compression [ACK-compression].
 - A. A forwarding engine SHOULD, if practical, be able to sustain an arbitrarily long sequence of small packets arriving at full interface rate.
 - B. If indefinite full packet rate for small packets is not practical, a forwarding engine MUST be able to buffer a long sequence of small packets inbound to the on-chip decision engine and sustain full interface rate for some reasonable average packet rate. Absent this small on-chip buffering, QoS agnostic packet drops can occur.

See Section 2.3.

5. The implementor and system designer MUST support pseudowire control word if MPLS-TP is supported or if ACH is being used on a pseudowire [RFC5586]. Deployments SHOULD enable pseudowire control word. See Section 2.4.1.
6. The implementor and system designer SHOULD support adding a pseudowire Flow Label [RFC6391]. Deployments MAY enable this feature for appropriate pseudowire types. See Section 2.4.3.
7. The implementor and system designer SHOULD support adding a MPLS Entropy Label [RFC6790]. Deployments MAY enable this feature. See Section 2.4.4.

1.3. Target Audience

This document is intended for multiple audiences: implementor (implementing MPLS forwarding in silicon or in software); systems designer (putting together a MPLS forwarding systems); deployer (running an MPLS network). These guidelines are intended to serve the following purposes:

1. Explain what to do and what not to do when a deep label stack is encountered. (audience: implementor)
2. Highlight pitfalls to look for when implementing an MPLS forwarding chip. (audience: implementor)
3. Provide a checklist of features and performance specifications to request. (audience: systems designer, deployer)

4. Provide a set of tests to perform. (audience: systems designer, deployer).

The implementor, systems designer, and deployer have a transitive supplier customer relationship. It is in the best interest of the supplier to review their product against their customer's checklist and customer's customer's checklist if applicable.

2. Forwarding Issues

A brief review of forwarding issues is provided in the subsections that follow. This section provides some background on why some of these requirements exist. The questions to ask of suppliers and testing is covered in the following sections, Section 3 and Section 4.

2.1. Forwarding Basics

Basic MPLS architecture and MPLS encapsulation, and therefore packet forwarding is defined in [RFC3031] and [RFC3032]. RFC3031 and RFC3032 are somewhat LDP centric. RSVP-TE supports traffic engineering (TE) and fast reroute, features that LDP lacks. The base document for RSVP-TE based MPLS is [RFC3209].

A few RFCs update RFC3032. Those with impact on forwarding include the following.

1. TTL processing is clarified in [RFC3443].
2. The use of MPLS Explicit NULL is modified in [RFC4182].
3. Differentiated Services is supported by [RFC3270] and [RFC4124]. The "EXP" field is renamed to "Traffic Class" in [RFC5462], removing any misconception that it was available for experimentation or could be ignored.
4. ECN is supported by [RFC5129].
5. The MPLS G-ACh and GAL are defined in [RFC5586].

Other RFCs have implications to MPLS Forwarding and do not update RFC3032 or RFC3209, including:

1. The pseudowire (PW) Associated Channel Header (ACH), defined by [RFC5085], later generalized by the MPLS G-ACh [RFC5586].

2. The Entropy Label Indicator and Entropy Label are defined by [RFC6790].

A few RFCs update RFC3209. Those that are listed as updating RFC3209 generally impact only RSVP-TE signaling. Forwarding is modified by major extension built upon RFC3209.

RFCs which impact forwarding are discussed in the following subsections.

2.1.1. MPLS Reserved Labels

[RFC3032] specifies that label values 0-15 are reserved labels with special meanings. Three values of NULL label are defined (two of which are later updated by [RFC4182]) and a router-alert label is defined. The original intent was that reserved labels, except the NULL labels, could be sent to the routing engine CPU rather than be processed in forwarding hardware. Hardware support is required by new RFCs such as those defining Entropy Label and OAM processed as a result of receiving a GAL. For new reserved labels, some accommodation is needed for LSR that will send the labels to a general purpose CPU. For example, ELI will only be sent to LSR which have signaled support for [RFC6790] and high OAM packet rate must be negotiated among endpoints.

[RFC3429] reserves a label for ITU-T Y.1711, however Y.1711 does not work with multipath and its use is strongly discouraged.

The current list of reserved labels can be found on the "Multiprotocol Label Switching Architecture (MPLS) Label Values" registry reachable at IANA's pages at <<http://www.iana.org>>.

When an unknown reserved label is encountered or a reserved label not directly handled in forwarding hardware is encountered, the packet should be sent to a general purpose CPU by default. If this capability is supported, there must be an option to either drop or rate limit such packets on a per reserved label value basis.

2.1.2. MPLS Differentiated Services

[RFC2474] deprecates the IP Type of Service (TOS) and IP Precedence (Prec) fields and replaces them with the Differentiated Services Field more commonly known as the Differentiated Services Code Point (DSCP) field. [RFC2475] defines the Differentiated Services architecture, which in other forum is often called a Quality of Service (QoS) architecture.

MPLS uses the Traffic Class (TC) field to support Differentiated

Services [RFC5462]. There are two primary documents describing how DSCP is mapped into TC.

1. [RFC3270] defines E-LSP and L-LSP. E-LSP use a static mapping of DSCP into TC. L-LSP use a per LSP mapping of DSCP into TC, with one PHB Scheduling Class (PSC) per L-LSP. Each PSC can use multiple Per-Hop Behavior (PHB) values. For example, the Assured Forwarding service defines three PSC, each with three PHB [RFC2597].
2. [RFC4124] defines assignment of a class-type (CT) to an LSP, where a per CT static mapping of TC to PHB is used. [RFC4124] provides a means to support up to eight E-LSP-like mappings of DSCP to TC.

To meet Differentiated Services requirements specified in [RFC3270], the following forwarding requirements must be met. An ingress LER MUST be able to select an LSP and then apply a per LSP map of DSCP into TC. A midpoint LSR MUST be able to apply a per LSP map of TC to PHB. The number of mappings supported will be far less than the number of LSP supported.

2.1.3. Time Synchronization

PTP or NTP may be carried over MPLS [I-D.ietf-tictoc-1588overmpls]. Generally NTP will be carried within IP with IP carried in MPLS [RFC5905]. Both PTP and NTP benefit from accurate time stamping of incoming packets and the ability to insert accurate time stamps in outgoing packets.

Since the label stack depth may vary, hardware should allow a timestamp to be placed in an outgoing packet at any specified byte position. It may be necessary to modify layer-2 checksums or frame check sequences after insertion. PTP and NTP timestamp formats differ slightly.

Accurate time synchronization in addition to being generally useful is required for MPLS-TP delay measurement (DM) OAM. See Section 2.6.4.

2.1.4. Uses of Multiple Label Stack Entries

MPLS deployments in the early part of the prior decade (circa 2000) tended to support either LDP or RSVP-TE. LDP was favored by some for its ability to scale to a very large number of PE devices at the edge of the network, without adding deployment complexity. RSVP-TE was favored, generally in the network core, where traffic engineering and/or fast reroute were considered important.

Both LDP and RSVP-TE are used simultaneously within major Service Provider networks using a technique known as "LDP over RSVP-TE Tunneling". This technique allows service providers to carry LDP tunnels, originating and terminating at PE's, inside of RSVP-TE tunnels, generally between Inter-City P routers, to take advantage of Traffic Engineering and Fast Re-Route on more expensive Inter-City and Inter-Continental Transport paths. LDP over RSVP-TE tunneling requires a minimum of two MPLS labels: one each for LDP and RSVP-TE.

The use of MPLS FRR [RFC4090] added one more label to MPLS traffic, but only when FRR protection was in use. If LDP over RSVP-TE is in use, and FRR protection is in use, then at least three MPLS labels are present on the label stack on the links through which the Bypass LSP traverses. FRR is covered in Section 2.1.7.

LDP L2VPN, LDP IPVPN, BGP L2VPN, and BGP IPVPN added support for VPN services that are deployed in the vast majority of service providers. These VPN services added yet another label, bringing the label stack depth (when FRR is active) to four.

Pseudowires and VPN are discussed in further detail in Section 2.1.8 and Section 2.1.9.

2.1.5. MPLS Link Bundling

MPLS Link Bundling was the first RFC to address the need for multiple parallel links between nodes [RFC4201]. MPLS Link Bundling is notable in that it tried not to change MPLS forwarding, except in specifying the "All-Ones" component link. MPLS Link Bundling is seldom if ever deployed. Instead multipath techniques described in Section 2.4 are used.

2.1.6. MPLS Hierarchy

MPLS hierarchy is defined in [RFC4206]. Although RFC4206 is considered part of GMPLS, the Packet Switching Capable (PSC) portion of the MPLS hierarchy are applicable to MPLS and may be supported in an otherwise GMPLS free implementation. The MPLS PSC hierarchy remains the most likely means of providing further scaling in an RSVP-TE MPLS network, particularly where the network is designed to provide RSVP-TE connectivity to the edges. This is the case for envisioned MPLS-TP networks. The use of the MPLS PSC hierarchy can add as many as four labels to a label stack, though it is likely that only one layer of PSC will be used in the near future.

2.1.7. MPLS Fast Reroute (FRR)

Fast reroute is defined by [RFC4090]. Two significantly different methods are the "One-to-One Backup" method which uses the "Detour LSP" and the "Facility Backup" which uses a "bypass tunnel". These are commonly referred to as the detour and bypass methods respectively.

The detour method makes use of a presignaled LSP. Hardware assistance is needed for detour FRR only if necessary to accomplish local repair of a large number of LSP within the 10s of milliseconds target. For each affected LSP a SWAP operation must be reprogrammed or otherwise switched over. The use of detour FRR doubles the number of LSP terminating at any given hop and will increase the number of LSP within a network by a factor dependent on the average detour path length.

The bypass method makes use of a tunnel that is unused when no fault exists but may carry many LSP when a local repair is required. There is no presignaling indicating which working LSP will be diverted into any specific bypass LSP. The egress LSR of the bypass LSP MUST use platform label space (as defined in [RFC3031]) so that an LSP working path on any give interface can be backed up using a bypass LSP terminating on any other interface. Hardware assistance is needed if necessary to accomplish local repair of a large number of LSP within the 10s of milliseconds target. For each affected LSP a SWAP operation must be reprogrammed or otherwise switched over with an additional PUSH of the bypass LSP label. In addition the use of platform label space impacts the size of the LSR ILM for LSR with a very large number of interfaces.

2.1.8. Pseudowire Encapsulation

The pseudowire (PW) architecture is defined in [RFC3985]. A pseudowire, when carried over MPLS, adds one or more additional label entries to the MPLS label stack. A PW Control Word is defined in [RFC4385] with motivation for defining the control word in [RFC4928]. The PW Associated Channel defined in [RFC4385] is used for OAM in [RFC5085]. The PW Flow Label is defined in [RFC6391] and is discussed further in this document in Section 2.4.3.

There are numerous pseudowire encapsulations, supporting emulation of services such as Frame Relay, ATM, Ethernet, TDM, and SONET/SDH over packet switched networks (PSNs) using IP or MPLS.

The pseudowire encapsulation is out of scope for this document. Pseudowire impact on MPLS forwarding at midpoint LSR is within scope. The impact on ingress MPLS PUSH and egress MPLS UHP POP are within

scope. While pseudowire encapsulation is out of scope, some advice is given on sequence number support.

2.1.8.1. Pseudowire Sequence Number

Pseudowire (PW) sequence number support is most important for PW payload types with a high expectation of in-order delivery. Resequencing support, rather than dropping at egress on out of order arrival, is most important for PW payload types with a high expectation of lossless delivery. For example, TDM payloads require sequence number support and require resequencing support. The same is true of ATM CBR service. ATM VBR or ABR may have somewhat relaxed requirements, but generally require ATM Early Packet Discard (EPD) or ATM Partial Packet Discard (PPD) [ATM-EPD-and-PPD]. Though sequence number support and resequencing support are beneficial to PW packet oriented payloads such as FR and Ethernet, they are highly desirable but not as strongly required.

Packet reorder should be rare except in a small number of circumstances, most of which are due to network design or equipment design errors:

1. The most common case is where reordering occurs is rare, occurring only when a network or equipment fault forces traffic on a new path with different delay. The packet loss that accompanies a network or equipment fault is generally more disruptive than any reordering which may occur.
2. A path change can be caused by reasons other than a network or equipment fault, such as administrative routing change. This may result in packet reordering but generally without any packet loss.
3. If the edge is not using pseudowire control word (CW) and the core is using multipath, reordering will be far more common. If this is occurring, the best solution is to use CW on the edge, rather than try to fix the reordering using resequencing.
4. Another avoidable case is where some core equipment has multipath and for some reason insists on periodically installing a new random number as the multipath hash seed. If supporting MPLS-TP, equipment MUST provide a means to disable periodic hash reseeding and deployments MUST disable periodic hash reseeding. Even if not supporting MPLS-TP, equipment should provide a means to disable periodic hash reseeding and deployments should disable periodic hash reseeding.

2.1.9. Layer-2 and Layer-3 VPN

Layer-2 VPN [RFC4664] and Layer-3 VPN [RFC4110] add one or more label entry to the MPLS label stack. VPN encapsulations are out of scope for this document. Its impact on forwarding at midpoint LSR are within scope.

Any of these services may be used on an MPLS Entropy Label enabled ingress and egress (see Section 2.4.4 for discussion of Entropy Label) which would add an additional label to the MPLS label stack. The need to provide a useful Entropy Label value impacts the requirements of the VPN ingress LER but is out of scope for this document.

2.2. MPLS Multicast

MPLS Multicast encapsulation is clarified in [RFC5332]. MPLS Multicast may be signaled using RSVP-TE [RFC4875] or LDP [RFC6388].

[RFC4875] defines a root initiated RSVP-TE LSP setup rather than leaf initiated join used in IP multicast. [RFC6388] defines a leaf initiated LDP setup. Both [RFC4875] and [RFC6388] define point to multipoint (P2MP) LSP setup. [RFC6388] also defined multipoint to multipoint (MP2MP) LSP setup.

The P2MP LSP have a single source. An LSR may be a leaf node, an intermediate node, or a "bud" node. A bud serves as both a leaf and intermediate. At a leaf an MPLS POP is performed. The payload may be a IP Multicast packet that requires further replication. At an intermediate node a MPLS SWAP is performed. The bud requires that both a POP and SWAP be performed for the same incoming packet.

One strategy to support P2MP functionality is to POP at the LSR ingress and then optionally PUSH labels at each LSR egress. A given LSR egress chip may support multiple egress interfaces, each of which requires a copy, but each with a different set of added labels and layer-2 encapsulation. Some physical interfaces may have multiple sub-interfaces (such as Ethernet VLAN or channelized interfaces) each requiring a copy.

If packet replication is performed at LSR ingress, then the ingress interface performance may suffer. If the packet replication is performed within a LSR switching fabric and at LSR egress, congestion of egress interfaces cannot make use of backpressure to ingress interfaces using techniques such as virtual output queuing (VOQ). If buffering is primarily supported at egress, then the need for backpressure is minimized. There may be no good solution for high volumes of multicast traffic if VOQ is used.

MP2MP LSP differ in that any branch may provide an input, including a leaf. Packets must be replicated onto all other branches. This forwarding is often implemented as multiple P2MP forwarding trees, one for each potential input.

2.3. Packet Rates

While average packet size of Internet traffic may be large, long sequences of small packets have both been predicted in theory and observed in practice. Traffic compression and TCP ACK compression can conspire to create long sequences of packets of 40-44 bytes in payload length. If carried over Ethernet, the 64 byte minimum payload applies, yielding a packet rate of approximately 150 Mpps (million packets per second) for the duration of the burst on a nominal 100 Gb/s link. The peak rate is higher for other encapsulations can be as high as 250 Mpps (for example IP or MPLS encapsulated using GFP over OTN ODU4).

It is also possible that the packet rates for a minimum payload size, such as 64 byte (64B) payload for Ethernet, is acceptable, but the rate declines for other packet sizes, such as 65B payload. There are other packet rates of interest besides TCP ACK. For example, a TCP ACK carried over an Ethernet PW over MPLS over Ethernet may occupy 82B or 82B plus an increment of 4B if additional MPLS labels are present.

A graph of packet rate vs. packet size often displays a sawtooth. The sawtooth is commonly due to a memory bottleneck and memory widths, sometimes internal cache, but often a very wide external buffer memory interface. In some cases it may be due to a fabric transfer width. A fine packing, rounding up to the nearest 8B or 16B will result in a fine sawtooth with small degradation for 65B, and even less for 82B packets. A coarse packing, rounding up to 64B can yield a sharper drop in performance for 65B packets, or perhaps more important, a larger drop for 82B packets.

The loss of some TCP ACK packets are not the primary concern when such a burst occurs. When a burst occurs, any other packets, regardless of packet length and packet QoS are dropped once on-chip input buffers prior to the decision engine are exceeded. Buffers in front of the packet decision engine are often very small or non-existent (less than one packet of buffer) causing significant QoS agnostic packet drop.

Internet service providers and content providers generally specify full rate forwarding with 40 byte payload packets as a requirement. This requirement often can be waived if the provider can be convinced that when long sequence of short packets occur no packets will be

dropped.

Many equipment suppliers have pointed out that the extra cost in designing hardware capable of processing the minimum size packets at full line rate is significant for very high speed interfaces. If hardware is not capable of processing the minimum size packets at full line rate, then that hardware MUST be capable of handling large burst of small packets, a condition which is often observed. This level of performance is necessary to meet Differentiated Services [RFC2475] requirements for without it, packets are lost prior to inspection of the IP DSCP field [RFC2474] or MPLS TC field [RFC5462].

With adequate on-chip buffers before the packet decision engine, an LSR can absorb a long sequence of short packets. Even if the output is slowed to the point where light congestion occurs, the packets, having cleared the decision process, can make use of larger VOQ or output side buffers and be dealt with according to configured QoS treatment, rather than dropped completely at random.

These on-chip buffers need not contribute significant delay since they are only used when the packet decision engine is unable to keep up, not in response to congestion, plus these buffers are quite small. For example, an on-chip buffer capable of handling 4K packets of 64 bytes in length, or 256KB, corresponds to 2 msec on a 10 Mb/s link and 0.2 usec on a 100 Gb/s link. If the packet decision engine is capable of handling packets at 90% of the full rate for small packets, then the maximum added delay is 0.2 msec and 20 nsec respectively, and this delay only applies if a 4K burst of short packets occurs. When no burst of short packets was being processed, no delay is added.

Packet rate requirements apply regardless of which network tier equipment is deployed in. Whether deployed in the network core or near the network edges, one of the two conditions MUST be met:

1. Packets must be processed at full line rate with minimum sized packets. -OR-
2. Packets must be processed at a rate well under generally accepted average packet sizes, with sufficient buffering prior to the packet decision engine to accommodate long bursts of small packets.

2.4. MPLS Multipath Techniques

In any large provider, service providers and content providers, hash based multipath techniques are used in the core and in the edge. In many of these providers hash based multipath is also used in the

larger metro networks.

The most common multipath techniques are ECMP applied at the IP forwarding level, Ethernet LAG with inspection of the IP payload, and multipath on links carrying both IP and MPLS, where the IP header is inspected below the MPLS label stack. In most core networks, the vast majority of traffic is MPLS encapsulated.

In order to support an adequately balanced load distribution across multiple links, IP header information must be used. Common practice today is to reinspect the IP headers at each LSR and use the label stack and IP header information in a hash performed at each LSR. Further details are provided in Section 2.4.5.

The use of this technique is so ubiquitous in provider networks that lack of support for multipath makes any product unsuitable for use in large core networks. This will continue to be the case in the near future, even as deployment of MPLS Entropy Label begins to relax the core LSR multipath performance requirements given the existing deployed base of edge equipment without the ability to add an Entropy Label.

A generation of edge equipment supporting the ability to add an MPLS Entropy Label is needed before the performance requirements for core LSR can be relaxed. However, it is likely that two generations of deployment in the future will allow core LSR to support full packet rate only when a relatively small number of MPLS labels need to be inspected before hashing. For now, don't count on it.

Common practice today is to reinspect the packet at each LSR and use the label stack and use the IP header field as input to a hash algorithm performed on each packet at each LSR in the network combined with a hash seed that is selected by each LSR. Where flow labels or entropy labels are used, a hash seed must be used.

2.4.1. Pseudowire Control Word

Within the core of a network some form of multipath is almost certain to be used. Multipath techniques deployed today are likely to be looking beneath the label stack for an opportunity to hash on IP addresses.

A pseudowire encapsulated at a network edge must have a means to prevent reordering within the core if the pseudowire will be crossing a network core, or any part of a network topology where multipath is used (see [RFC4385] and [RFC4928]).

Not supporting the ability to encapsulate a pseudowire with a control

word may lock a product out from consideration. A pseudowire capability without control word support might be sufficient for applications that are strictly both intra-metro and low bandwidth. However a provider with other applications will very likely not tolerate having equipment which can only support a subset of their pseudowire needs.

2.4.2. Large Microflows

Where multipath makes use of a simple hash and simple load balance such as modulo or other fixed allocation (see Section 2.4) the presence of large microflows that each consumes 10% of the capacity of a component link of a potentially congested composite link, one such microflow can upset the traffic balance and more than one can in effect reduce the effective capacity of the entire composite link by more than 10%.

When even a very small number of large microflows are present, there is a significant probability that more than one of these large microflows could fall on the same component link. If the traffic contribution from large microflows is small, the probability for three or more large microflows on the same component link drops significantly. Therefore in a network where a significant number of parallel 10 Gb/s links exists, even a 1 Gb/s pseudowire or other large microflow that could not otherwise be subdivided into smaller flows should carry a flow label or entropy label if possible.

Active management of the hash space to better accommodate large microflows has been implemented and deployed in the past, however such techniques are out of scope for this document.

2.4.3. Pseudowire Flow Label

Unlike a pseudowire control word, a pseudowire flow label [RFC6391], is required only for relatively large capacity pseudowires. There are many cases where a pseudowire flow label makes sense. Any service such as a VPN which carries IP traffic within a pseudowire can make use of a pseudowire flow label.

Any pseudowire carried over MPLS which makes use of the pseudowire control word and does not carry a flow label is in effect a single microflow (in [RFC2475] terms).

2.4.4. MPLS Entropy Label

The MPLS Entropy Label simplifies flow group identification [RFC6790] in the network core. Prior to the MPLS Entropy Label core LSR needed to inspect the entire label stack and often the IP headers to provide

an adequate distribution of traffic when using multipath techniques (see Section 2.4.5). With the use of MPLS Entropy Label, a hash can be performed closer to network edges, placed in the label stack, and used within the network core.

The MPLS Entropy Label is capable of avoiding full label stack and payload inspection within the core where performance levels are most difficult to achieve (see Section 2.3). The label stack inspection can be terminated as soon as the first Entropy Label is encountered, which is generally after a small number of labels are inspected.

In order to provide these benefits in the core, LSR closer to the edge must be capable of adding an entropy label. This support may not be required in the access tier, the tier closest to the customer, but is likely to be required in the edge or the border to the network core. LSR peering with external networks will also need to be able to add an Entropy Label.

2.4.5. Fields Used for Multipath

The most common multipath techniques are based on a hash over a set of fields. Regardless of whether a hash is used or some other method is used, there are a limited set of fields which can safely be used for multipath.

2.4.5.1. MPLS Fields in Multipath

If the "outer" or "first" layer of encapsulation is MPLS, then label stack entries are used in the hash. Within a finite amount of time (and for small packets arriving at high speed that time can quite limited) only a finite number of label entries can be inspected. Pipelined or parallel architectures improve this, but the limit is still finite.

The following guidelines are provided for use of MPLS fields in multipath load balancing.

1. Only the 20 bit label field SHOULD be used. The TTL field SHOULD NOT be used. The S bit MUST NOT be used. The TC field (formerly EXP) MUST NOT be used. See below this list for reasons.
2. If an ELI label is found, then if the LSR supports Entropy Label, the EL label field in the next label entry (the EL) SHOULD be used and label entries below that label SHOULD NOT be used and the MPLS payload SHOULD NOT be used. See below this list for reasons.

3. Reserved labels (label values 0-15) MUST NOT be used. In particular, GAL and RA MUST NOT be used so that OAM traffic follows the same path as payload packets with the same label stack.
4. The most entropy is generally found in the label stack entries near the bottom of the label stack (innermost label, closest to S=1 bit). If the entire label stack cannot be used (or entire stack up to an EL), then it is better to use as many labels as possible closest to the bottom of stack.
5. If no ELI is encountered, and the first nibble of payload contains a 4 (IPv4) or 6 (IPv6), an implementation SHOULD support the ability to interpret the payload as IPv4 or IPv6 and extract and use appropriate fields from the IP headers. This feature is considered a hard requirement by many service providers. If supported, there MUST be a way to disable it (if, for example, PW without CW are used). This ability to disable this feature is considered a hard requirement by many service providers. Therefore an implementation has a very strong incentive to support both options.
6. A label which is popped at egress (UHP POP) SHOULD NOT be used. A label which is popped at the penultimate hop (PHP POP) SHOULD be used.

Apparently some chips have made use of the TC (formerly EXP) bits as a source of entropy. This is very harmful since it will reorder Assured Forwarding (AF) traffic [RFC2597] when a subset does not conform to the configured rates and is remarked but not dropped at a prior LSR. Traffic which uses MPLS ECN [RFC5129] can also be reordered if TC is used for entropy. Therefore, as stated in the guidelines above, the TC field (formerly EXP) MUST NOT be used in multipath load balancing as it violates Differentiated Services Ordered Aggregate (OA) requirements in these two instances.

Use of the MPLS label entry S bit would result in putting OAM traffic on a different path if the addition of a GAL at the bottom of stack removed the S bit from the prior label.

If an ELI label is found, then if the LSR supports Entropy Label, the EL label field in the next label entry (the EL) SHOULD be used and the search for additional entropy within the packet SHOULD be terminated. Failure to terminate the search will impact client MPLS-TP LSP carried within server MPLS LSP. A network operator has the option to use administrative attributes as a means to identify LSR which do not terminate the entropy search at the first EL. Administrative attributes are defined in [RFC3209]. Some

configuration is required to support this.

If the PHP POP label is not used, then for any PW for which CW is used, there is no basis for multipath load split. In some networks it is infeasible to put all PW traffic on one component link. Any PW which does not use CW will be improperly split regardless of whether the PHP POP label is used.

2.4.5.2. IP Fields in Multipath

Inspecting the IP payload provides the most entropy in provider networks. The practice of looking past the bottom of stack label for an IP payload is well accepted and documented in [RFC4928] and in other RFCs.

Where IP is mentioned in the document, both IPv4 and IPv6 apply. All LSRs MUST fully support IPv6.

When information in the IP header is used, the following guidelines apply:

1. Both the IP source address and IP destination address SHOULD be used. There MAY be an option to reverse the order of these address, improving the ability to provide symmetric paths in some cases. Many service providers require that both addresses be used.
2. Implementations SHOULD allow inspection of the IP protocol field and use of the UDP or TCP port numbers. For many service providers this feature is considered mandatory, particularly for enterprise, data center, or edge equipment. If this feature is provided, it SHOULD be possible to disable use of TCP and UDP ports. Many service providers consider it a hard requirement that use of UDP and TCP ports can be disabled. Therefore there is a strong incentive for implementations to provide both options.
3. Equipment suppliers MUST NOT make assumptions that because the IP version field is equal to 4 (an IPv4 packet) that the IP protocol will either be TCP (IP protocol 6) or UDP (IP protocol 17) and blindly fetch the data at the offset where the TCP or UDP ports would be found. With IPv6, TCP and UDP port numbers are not at fixed offsets. With IPv4 packets carrying IP options, TCP and UDP port numbers are not at fixed offsets.
4. The IPv6 header flow field SHOULD be used. This is the explicit purpose of the IPv6 flow field, however observed flow fields rarely contains a non-zero value. Some uses of the flow field have been defined such as [RFC6438]. In the absence of MPLS

encapsulation, the IPv6 flow field can serve a role equivalent to Entropy Label.

5. Support other protocols that share a common Layer-4 header such as RTP, UDP-lite, SCTP and DCCP SHOULD be provided, particularly for edge or access equipment where additional entropy may be needed. Equipment SHOULD also use RTP, UDP-lite, SCTP and DCCP headers when creating an Entropy Label.
6. Similar to avoiding TC in MPLS, the IP DSCP, and ECN bits MUST NOT be used. The IPv4 TTL or IPv6 Hop Count SHOULD NOT be used. Note that the IP TOS field was deprecated ([RFC0791] was updated by [RFC2474]). No part of the IP DSCP (formerly IP PREC and IP TOS bits) field can be used.
7. Some IP encapsulations support tunneling, such as IP-in-IP, GRE, L2TPv3, and IPSEC. These provide a greater source of entropy which some provider networks carrying large amounts of tunneled traffic may need. The use of tunneling header information is out of scope for this document.

This document makes the following recommendations. These recommendations are not required to claim compliance to any existing RFC therefore implementors are free to ignore them, but due to service provider requirements may be doing so at their own peril. The use of IP addresses MUST be supported and TCP and UDP ports (conditional on the protocol field and properly located) MUST be supported. The ability to disable use of UDP and TCP ports MUST be available. Though potentially very useful in some networks, it is uncommon to support using payloads of tunneling protocols carried over IP. Though the use of tunneling protocol header information is out of scope for this document, it is not discouraged.

2.4.5.3. Fields Used in Flow Label

The ingress to a pseudowire (PW) can extract information from the payload being encapsulated to create a flow label. [RFC6391] references IP carried in Ethernet as an example. The Native Service Processing (NSP) function defined in [RFC3985] differs with pseudowire type. It is in the NSP function where information for a specific type of PW can be extracted for use in a flow label. Which fields to use for any given PW NSP is out of scope for this document.

2.4.5.4. Fields Used in Entropy Label

An entropy label is added at the ingress to an LSP. The payload being encapsulated is most often MPLS, a PW, or IP. The payload type is identified by the layer-2 encapsulation (Ethernet, GFP, POS, etc).

If the payload is MPLS, then the information used to create an entropy label is the same information used for local load balancing (see Section 2.4.5.1). This information **MUST** be extracted for use in generating an entropy label even if the LSR local egress interface is not a multipath.

Of the non-MPLS payload types, only payloads that are forwarded are of interest. For example, ARP is not forwarded and CNLP (used only for ISIS) is not forwarded.

The non-MPLS payload type of greatest interest are IPv4 and IPv6. The guidelines in Section 2.4.5.2 apply to fields used to create and entropy label.

The IP tunneling protocols mentioned in Section 2.4.5.2 may be more applicable to generation of an entropy label at edge or access where deep packet inspection is practical due to lower interface speeds than in the core where deep packet inspection may be impractical.

2.5. MPLS-TP and UHP

MPLS-TP introduces forwarding demands that will be extremely difficult to meet in a core network. Most troublesome is the requirement for Ultimate Hop Popping (UHP, the opposite of Penultimate Hop Popping or PHP). Using UHP opens the possibility of one or more MPLS POP operation plus an MPLS SWAP operation for each packet. The potential for multiple lookups and multiple counter instances per packet exists.

As networks grow and tunneling of LDP LSPs into RSVP-TE LSPs is used, and/or RSVP-TE hierarchy is used, the requirement to perform one or two or more MPLS POP operations plus a MPLS SWAP operation (and possibly a PUSH or two) increases. If MPLS-TP LM (link monitoring) OAM is enabled at each layer, then a packet and byte count **MUST** be maintained for each POP and SWAP operation so as to offer OAM for each layer.

2.6. OAM and DoS Protection

Denial of service (DoS) protection is an area requiring hardware support that is often overlooked or inadequately considered. Hardware assist is also needed for OAM, particularly the more demanding MPLS-TP OAM.

2.6.1. DoS Protection

Modern equipment supports a number of control plane and management plane protocols. Generally no single means of protecting network

equipment from denial of service (DoS) attacks is sufficient, particularly for high speed interfaces. This problem is not specific to MPLS, but is a topic that cannot be ignored when implementing or evaluating MPLS implementations.

Two types of protections are often cited as primary means of protecting against attacks of all kinds.

Isolated Control/Management Traffic

Control and Management traffic can be carried out-of-band (OOB), meaning not intermixed with payload. For MPLS use of G-ACh and GAL to carry control and management traffic provides a means of isolation from potentially malicious payload. Used along, the compromise of a single node, including a small computer at a network operations center, could compromise an entire network. Implementations which send all G-ACh/GAL traffic directly to a routing engine CPU are subject to DoS attack as a result of such a compromise.

Cryptographic Authentication

Cryptographic authentication can very effectively prevent malicious injection of control or management traffic. Cryptographic authentication can in some circumstances be subject to DoS attack by overwhelming the capacity of the decryption with a high volume of malicious traffic. For very low speed interfaces cryptographic authentication can be performed by the general purpose CPU used as a routing engine. For all other cases, cryptographic hardware may be needed. For very high speed interfaces, even cryptographic hardware can be overwhelmed.

Some control and management protocols are often carried with payload traffic. This is commonly the case with BGP, T-LDP, and SNMP. It is often the case with RSVP-TE. Even when carried over G-ACh/GAL additional measures can reduce the potential for a minor breach to be leveraged to a full network attack.

Some of the additional protections are supported by hardware packet filtering.

GTSM

[RFC5082] defines a mechanism that uses the IPv4 TTL or IPv6 Hop Limit fields to insure control traffic that can only originate from an immediate neighbor is not forged and originating from a distant source. GTSM can be applied to many control protocols which are routable, for example LDP [RFC6720].

IP Filtering

At the very minimum, packet filtering plus classification and use of multiple queues supporting rate limiting is needed for traffic that could potentially be sent to a general purpose CPU used as a routing engine. The first level of filtering only allows connections to be initiated from specific IP prefixes to specific destination ports and then preferably passes traffic directly to a cryptographic engine and/or rate limits. The second level of filtering passes connected traffic, such as TCP connections having received at least one authenticated SYN or having been locally initiated. The second level of filtering only passes traffic to specific address and port pairs to be checked for cryptographic authentication.

The cryptographic authentication is generally the last resort in DoS attack mitigation. If a packet must be first sent to a general purpose CPU, then sent to a cryptographic engine, a DoS attack is possible on high speed interfaces. Only where hardware can identify a signature and the portion of packet covered by the signature is cryptographic authentication highly beneficial in protecting against DoS attacks.

For chips supporting multiple 100 Gb/s interfaces, only a very large number of parallel cryptographic engines can provide the processing capacity to handle a large scale DoS or distributed DoS (DDoS) attack. For many forwarding chips this much processing power requires significant chip real estate and power, and therefore reduces system space and power density. For this reason, cryptographic authentication is not considered a viable first line of defense.

For some networks the first line of defense is some means of supporting OOB control and management traffic. In the past this OOB channel might make use of overhead bits in SONET or OTN or a dedicated DWDM wavelength. G-ACh and GAL provide an alternative OOB mechanism which is independent of underlying layers. In other networks, including most IP/MPLS networks, perimeter filtering serves a similar purpose, though less effective without extreme vigilance.

A second line of defense is filtering, including GTSM. For protocols such as EBGp, GTSM and other filtering is often the first line of defense. Cryptographic authentication is usually the last line of defense and insufficient by itself to mitigate DoS or DDoS attacks.

2.6.2. MPLS OAM

[RFC4377] defines requirements for MPLS OAM that predate MPLS-TP.
[RFC4379] defines what is commonly referred to as LSP Ping and LSP

Traceroute. [RFC4379] is updated by [RFC6424] supporting MPLS tunnels and stitched LSP and P2MP LSP. [RFC4379] is updated by [RFC6425] supporting P2MP LSP. [RFC4379] is updated by [RFC6426] to support MPLS-TP connectivity verification (CV) and route tracing.

[RFC4950] extends the ICMP format to support TTL expiration that may occur when using IP traceroute within an MPLS tunnel. The ICMP message generation can be implemented in forwarding hardware, but if sent to a general purpose CPU must be rate limited to avoid a potential denial or service (DoS) attack.

[RFC5880] defines Bidirectional Forwarding Detection (BFD), a protocol intended to detect faults in the bidirectional path between two forwarding engines. [RFC5884] and [RFC5885] define BFD for MPLS. BFD can provide failure detection on any kind of path between systems, including direct physical links, virtual circuits, tunnels, MPLS Label Switched Paths (LSPs), multihop routed paths, and unidirectional links as long as there is some return path.

The processing requirements for BFD are less than for LSP Ping, making BFD somewhat better suited for relatively high rate proactive monitoring. BFD does not verify that the data plane against the control plane, where LSP Ping does. LSP Ping somewhat better suited for on-demand monitoring including relatively low rate periodic verification of data plane and as a diagnostic tool.

Both BFD and LSP Ping MUST be recognized by hardware and at the very minimum forwarded to the main CPU. Hardware assistance for BFD is often provided and is considered necessary for relatively high rate proactive monitoring. Both BFD and LSP Ping MUST be recognized in any filtering prior to passing traffic to a general purpose CPU and appropriate DoS protection applied (see Section 2.6.1. Failure to recognize BFD and LSP Ping and at least rate limit creates the potential for misconfiguration to cause outages rather than cause errors in the misconfigured OAM.

2.6.3. Pseudowire OAM

Pseudowire OAM makes use of the control channel provided by Virtual Circuit Connectivity Verification (VCCV) [RFC5085]. VCCV makes use of the Pseudowire Control Word. BFD support over VCCV is defined by [RFC5885]. [RFC5885] is updated by [RFC6478] in support of static pseudowires. [RFC4379] is updated by [RFC6829] supporting LSP Ping for Pseudowire FEC advertised over IPv6.

G-ACh/GAL (defined in [RFC5586]) is the preferred MPLS-TP OAM control channel and applies to any MPLS-TP end points, including Pseudowire. See Section 2.6.4 for an overview of MPLS-TP OAM.

2.6.4. MPLS-TP OAM

[RFC6669] summarizes the MPLS-TP OAM toolset, the set of protocols supporting the MPLS-TP OAM requirements specified in [RFC5860] and supported by the MPLS-TP OAM framework defined in [RFC6371].

The MPLS-TP OAM toolset includes:

CC-CV

[RFC6428] defines BFD extensions to support proactive CC-CV applications. [RFC6426] provides LSP ping extensions that are used to implement on-demand connectivity verification.

RDI

Remote Defect Indication (RDI) is triggered by failure of proactive CC-CV, which is BFD based. For fast RDI initiation, RDI SHOULD be initiated and handled by hardware if BFD is handled in forwarding hardware. [RFC6428] provides an extension for BFD that includes the RDI indication in the BFD format and a specification of how this indication is to be used.

Route Tracing

[RFC6426] specifies that the LSP ping enhancements for MPLS-TP on-demand connectivity verification include information on the use of LSP ping for route tracing of an MPLS-TP path.

Alarm Reporting

[RFC6427] describes the details of a new protocol supporting Alarm Indication Signal, Link Down Indication, and fault management. This functionality SHOULD be supported in forwarding hardware on high speed interfaces.

Lock Instruct

Lock instruct is initiated on-demand and therefore need not be implemented in forwarding hardware. [RFC6435] defines a lock instruct protocol.

Lock Reporting

[RFC6427] covers lock reporting. Lock reporting need not be implemented in forwarding hardware.

Diagnostic

[RFC6435] defines protocol support for loopback. Loopback initiation is on-demand and therefore need not be implemented in forwarding hardware. Loopback of packet traffic SHOULD be implemented in forwarding hardware on high speed interfaces.

Packet Loss and Delay Measurement

[RFC6374] and [RFC6375] define a protocol and profile for packet loss measurement (LM) and delay measurement (DM). LM requires a very accurate capture and insertion of packet and byte counters when a packet is transmitted and capture of packet and byte counters when a packet is received. This capture and insertion MUST be implemented in forwarding hardware for LM OAM to be sufficiently accurate. DM requires very accurate capture and insertion of a timestamp on transmission and capture of timestamp when a packet is received. This timestamp capture and insertion MUST be implemented in forwarding hardware for DM OAM to be sufficiently accurate.

See Section 2.6.2 for discussion of hardware support necessary for BFD and LSP Ping.

CC-CV and alarm reporting is tied to protection and therefore SHOULD be supported in forwarding hardware in order to provide protection for a large number of affected LSP within target response intervals. Since CC-CV is supported by BFD, for MPLS-TP, BFD SHOULD be supported in forwarding hardware.

2.6.5. MPLS OAM and Layer-2 OAM Interworking

[RFC6670] provides the reasons for selecting a single MPLS-TP OAM solution and examines the consequences were ITU-T to develop a second OAM solution that is based on Ethernet encodings and mechanisms.

[RFC6310] and [I-D.ietf-pwe3-mpls-eth-oam-iwk] specifies the mapping of defect states between many types of hardware Attachment Circuits (ACs) and associated Pseudowires (PWs). This functionality SHOULD be supported in forwarding hardware.

An MPLS OAM implementation SHOULD interwork with the underlying server layer and provide a means to interwork with a client layer. Where MPLS hierarchy is used both the client and server layer may be MPLS or MPLS-TP. Where the server layer is a Layer-2, such as Ethernet, PPP over SONET/SDH, or GFP over OTN, interwork among layers is also required. For high speed interfaces, this interworking SHOULD be supported in forwarding hardware.

2.6.6. Extent of OAM Support by Hardware

Some OAM functionality must be supported in forwarding hardware while other OAM functionality must be entirely implemented in forwarding hardware.

Where possible, implementation in forwarding hardware should be in

programmable hardware such that if standards are later changed or extended these changes are likely to be accommodated with hardware reprogramming rather than replacement.

Some functions must be implemented in dedicated forwarding hardware. Examples include packet and byte counters needed for LM OAM as well as needed for management protocols. Similarly the capture and insertion of packet and byte counts or timestamps needed for transmitted LM or DM or time synchronization packets MUST be implemented in forwarding hardware to support accurate OAM.

Some functions must be supported in forwarding hardware but may make use of an external general purpose processor if performance criteria can be met. For example origination of AIS to client layers may be triggered by CC-CV server layer hardware but expansion to a large number of client LSP may occur in a general purpose processor. Some forwarding hardware supports one or more on-chip general purpose processors which may be well suited for such a role.

The customer (system supplier or provider) should not dictate design, but should independently validate target functionality and performance. However, it is not uncommon for service providers and system implementors to insist on reviewing design details (under NDA) due to past experiences with suppliers and to reject suppliers who are unwilling to provide details.

2.7. Number and Size of Flows

Service provider networks may carry up to hundreds of millions of flows on 10 Gb/s links. Most flows are very short lived, many under a second. A subset of the flows are low capacity and somewhat long lived. When Internet traffic dominates capacity a very small subset of flows are high capacity and/or very long lived.

Two types of limitations with regard to number and size of flows have been observed.

1. Some hardware cannot handle some very large flows because of internal paths which are limited, such as per packet backplane paths or paths internal or external to chips such as buffer memory paths. Such designs can handle aggregates of smaller flows. Some hardware with acknowledged limitations has been successfully deployed but may be increasingly problematic if the capacity of large microflows in deployed networks continues to grow.
2. Some hardware approaches cannot handle a large number of flows, or a large number of large flows due to attempting to count per

flow, rather than deal with aggregates of flows. Hash techniques scale with regard to number of flows due to a fixed hash size with many flows falling into the same hash bucket. Techniques that identify individual flows have been implemented but have never successfully deployed for Internet traffic.

3. Questions for Suppliers

The following questions should be asked of a supplier. These questions are grouped into broad categories. The questions themselves are intended to be an open ended question to the supplier. The tests in Section 4 are intended to verify whether the supplier disclosed any compliance or performance limitations completely and accurately.

Basic Compliance

- Q#1 Can the implementation forward packets with an arbitrarily large stack depth? What limitations exist, and under what circumstances do further limitations come into play (such as high packet rate or specific features enabled or specific types of packet processing)? See Section 2.1.
- Q#2 Is the entire set of basic MPLS functionality described in Section 2.1 supported?
- Q#3 Are the set of MPLS reserved labels handled correctly and with adequate performance? See Section 2.1.1.
- Q#4 Are mappings of label value and TC to PHB handled correctly, including RFC3270 L-LSP mappings and RFC4124 CT mappings to PHB? See Section 2.1.2.
- Q#5 Is time synchronization adequately supported in forwarding hardware?
 - a. Are both PTP and NTP formats supported?
 - b. Is the accuracy of timestamp insertion and incoming stamping sufficient?See Section 2.1.3.
- Q#6 Is link bundling supported?
 - a. Can LSP be pinned to specific components?

- b. Is the "all-ones" component link supported?

See Section 2.1.5.

Q#7 Is MPLS hierarchy supported?

- a. Are both PHP and UHP supported? What limitations exist on the number of POP operations with UHP?
- b. Are the pipe, short-pipe, and uniform models supported? Are TTL and TC values updated correctly at egress where applicable?

See Section 2.1.6

Q#8 Are pseudowire sequence numbers handled correctly? See Section 2.1.8.1.

Q#9 Is VPN LER functionality handled correctly and without performance issues? See Section 2.1.9.

Q#10 Is MPLS multicast (P2MP and MP2MP) handled correctly?

- a. Are packets dropped on uncongested outputs if some outputs are congested?
- b. Is performance limited in high fanout situations?

See Section 2.2.

Basic Performance

Q#11 Can very small packets be forwarded at full line rate on all interfaces indefinitely? What limitations exist, and under what circumstances do further limitations come into play (such as specific features enabled or specific types of packet processing)?

Q#12 Customers must decide whether to relax the prior requirement and to what extent. If the answer to the prior question indicates that limitations exist, then:

- a. What is the smallest packet size where full line rate forwarding can be supported?
- b. What is the longest burst of full rate small packets that can be supported?

Specify circumstances (such as specific features enabled or specific types of packet processing) often impact these rates and burst sizes.

- Q#13 How many POP operations can be supported along with a SWAP operation at full line rate while maintaining per LSP packet and byte counts for each POP and SWAP? This requirement is particularly relevant for MPLS-TP.
- Q#14 How many PUSH labels can be supported. While this limitation is rarely an issue, it applies to both PHP and UHP, unlike the POP limit which applies to UHP.
- Q#15 For a worst case where all packets arrive on one LSP, what is the counter overflow time? Are any means provided to avoid polling all counters at short intervals? This applies to both MPLS and MPLS-TP.

Multipath Capabilities and Performance

Multipath capabilities and performance do not apply to MPLS-TP but apply to MPLS and apply if MPLS-TP is carried in MPLS.

- Q#16 How are large microflows accommodated? Is there active management of the hash space mapping to output ports? See Section 2.4.2.
- Q#17 How many MPLS labels can be included in a hash based on the MPLS label stack?
- Q#18 Is packet rate performance decreased beyond some number of labels?
- Q#19 Can the IP header and payload information below the MPLS stack be used in the hash? If so, which IP fields, payload types and payload fields are supported?
- Q#20 At what maximum MPLS label stack depth can Bottom of Stack and an IP header appear without impacting packet rate performance?
- Q#21 Are reserved labels excluded from the label stack hash? They MUST be excluded.
- Q#22 How is multipath performance affected by very large flows or an extremely large number of flows, or by very short lived flows? See Section 2.7.

Pseudowire Capabilities and Performance

- Q#23 Is the pseudowire control word supported?
- Q#24 What is the maximum rate of pseudowire encapsulation and decapsulation? Apply the same questions as in Base Performance for any packet based pseudowire such as IP VPN or Ethernet.
- Q#25 Does inclusion of a pseudowire control word impact performance?
- Q#26 Are flow labels supported?
- Q#27 If so, what fields are hashed on for the flow label for different types of pseudowires?
- Q#28 Does inclusion of a flow label impact performance?

Entropy Label Support and Performance

- Q#29 Can an entropy label be added when acting as in ingress LER and can it be removed when acting as an egress LER?
- Q#30 If so, what fields are hashed on for the entropy label?
- Q#31 Does adding or removing an entropy label impact packet rate performance?
- Q#32 Can an entropy label be detected in the label stack, used in the hash, and properly terminate the search for further information to hash on?
- Q#33 Does using an entropy label have any negative impact on performance? It should have no impact or a positive impact.

OAM and DoS Protection

- Q#34 For each control and management plane protocol in use, what measures are taken to provide DoS attack hardenning? Have DoS attack tests been performed? Can compromise of an internal computer on a management subnet be leveraged for any form of attack including DoS attack?

Q#35 What OAM proactive and on-demand mechanisms are supported? What performance limits exist under high proactive monitoring rates? Can excessively high proactive monitoring rates impact control plane performance or cause control plane instability? Ask these questions for each of the following.

- a. MPLS OAM
- b. Pseudowire OAM
- c. MPLS-TP OAM
- d. Layer-2 OAM Interworking

See Section 2.6.

4. Forwarding Compliance and Performance Testing

Packet rate performance of equipment supporting a large number of 10 Gb/s or 100 Gb/s links is not possible using desktop computers or workstations. The use of high end workstations as a source of test traffic was barely viable 20 years ago, but is no longer at all viable. Though custom microcode has been used on specialized router forwarding cards to serve the purpose of generating test traffic and measuring it, for the most part performance testing will require specialized test equipment. There are multiple sources of suitable equipment.

The set of tests listed here do not correspond one-to-one to the set of questions in Section 3. The same categorization is used and these tests largely serve to validate answers provided to the prior questions, and can also provide answers where a supplier is unwilling to disclose compliance or performance.

Performance testing is the domain of the IETF Benchmark Methodology Working Group (BMWG). Below are brief descriptions of conformance and performance tests. Some very basic tests are specified in [RFC5695] which partially cover only the basic performance test T#3.

The following tests should be performed by the systems designer, or deployer, or performed by the supplier on their behalf if it is not practical for the potential customer to perform the tests directly. These tests are grouped into broad categories.

Basic Compliance

- T#1 Test forwarding at a high rate for packets with varying number of label entries. While packets with more than a dozen label entries are unlikely to be used in any practical scenario today, it is useful to know if limitations exists.
- T#2 For each of the questions listed under "Basic Compliance" in Section 3, verify the claimed compliance. For any functionality considered critical to a deployment, where applicable performance using each capability under load should be verified in addition to basic compliance.

Basic Performance

- T#3 Test packet forwarding at full line rate with small packets. See [RFC5695]. The most likely case to fail is the smallest packet size. Also test with packet sizes in four byte increments ranging from payload sizes of 40 to 128 bytes.
- T#4 If the prior tests did not succeed for all packet sizes, then perform the following tests.
 - a. Increase the packet size by 4 bytes until a size is found that can be forwarded at full rate.
 - b. Inject bursts of consecutive small packets into a stream of larger packets. Allow some time for recovery between bursts. Increase the number of packets in the burst until packets are dropped.
- T#5 Send test traffic where a SWAP operation is required. Also set up multiple LSP carried over other LSP where the device under test (DUT) is the egress of these LSP. Create test packets such that the SWAP operation is performed after POP operations, increasing the number of POP operations until forwarding of small packets at full line rate can no longer be supported. Also check to see how many POP operations can be supported before the full set of counters can no longer be maintained. This requirement is particularly relevant for MPLS-TP.
- T#6 Send all traffic on one LSP and see if the counters become inaccurate. Often counters on silicon are much smaller than the 64 bit packet and byte counters in IETF MIB. System developers should consider what counter polling rate is necessary to maintain accurate counters and whether those polling rates are practical. Relevant MIBs for MPLS are

discussed in [RFC4221] and [RFC6639].

Multipath Capabilities and Performance

Multipath capabilities do not apply to MPLS-TP but apply to MPLS and apply if MPLS-TP is carried in MPLS.

- T#7 Send traffic at a rate well exceeding the capacity of a single multipath component link, and where entropy exists only below the top of stack. If only the top label is used this test will fail immediately.
- T#8 Move the labels with entropy down in the stack until either the full forwarding rate can no longer be supported or most or all packets try to use the same component link.
- T#9 Repeat the two tests above with the entropy contained in IP headers or IP payload fields below the label stack rather than in the label stack. Test with the set of IP headers or IP payload fields considered relevant to the deployment or to the target market.
- T#10 Determine whether traffic that contains a pseudowire control word is interpreted as IP traffic. Information in the payload MUST NOT be used in the load balancing if the first nibble of the packet is not 4 or 6 (IPv4 or IPv6).
- T#11 Determine whether reserved labels are excluded from the label stack hash. They MUST be excluded.
- T#12 Perform testing in the presence of combinations of:
 - a. Very large microflows.
 - b. Relatively short lived high capacity flows.
 - c. Extremely large numbers of flows.
 - d. Very short lived small flows.

Pseudowire Capabilities and Performance

- T#13 Ensure that pseudowire can be set up with a pseudowire label and pseudowire control word added at ingress and the pseudowire label and pseudowire control word removed at egress.

- T#14 For pseudowire that contains variable length payload packets, repeat performance tests listed under "Basic Performance" for pseudowire ingress and egress functions.
- T#15 Repeat pseudowire performance tests with and without a pseudowire control word.
- T#16 Determine whether pseudowire can be set up with a pseudowire label, flow label, and pseudowire control word added at ingress and the pseudowire label, flow label, and pseudowire control word removed at egress.
- T#17 Determine which payload fields are used to create the flow label and whether the set of fields and algorithm provide sufficient entropy for load balancing.
- T#18 Repeat pseudowire performance tests with flow labels included.

Entropy Label Support and Performance

- T#19 Determine whether entropy labels can be added at ingress and removed at egress.
- T#20 Determine which fields are used to create an entropy label. Labels further down in the stack, including entropy labels further down and IP headers or IP payload fields where applicable should be used. Determine whether the set of fields and algorithm provide sufficient entropy for load balancing.
- T#21 Repeat performance tests under "Basic Performance" when entropy labels are used, where ingress or egress is the device under test (DUT).
- T#22 Determine whether an ELI is detected when acting as a midpoint LSR and whether the search for further information on which to base the load balancing is used. Information below the entropy label SHOULD NOT be used.
- T#23 Ensure that the Entropy Label Indicator and Entropy Label (ELI and EI) are removed from the label stack during UHP and PHP operations.
- T#24 Insure that operations on the TC field when adding and removing Entropy Label are correctly carried out. If TC is changed during a SWAP operation, the ability to transfer that change MUST be provided. The ability to suppress the

transfer of TC MUST also be provided. See "pipe", "short pipe", and "uniform" models in [RFC3443].

- T#25 Repeat performance tests for midpoint LSR with entropy labels found at various label stack depths.

DoS Protection

- T#26 Actively attack LSR under high protocol churn load and determine control plane performance impact or successful DoS under test conditions. Specifically test for the following.
- a. TCP SYN attack against control plane and management plane protocols using TCP, including CLI access (typically SSH protected login), NETCONF, etc.
 - b. High traffic volume attack against control plane and management plane protocols not using TCP.
 - c. Attacks which can be performed from a compromised management subnet computer, but not one with authentication keys.
 - d. Attacks which can be performed from a compromised peer within the control plane (internal domain and external domain). Assume that per peering keys and per router ID keys rather than network wide keys are in use.

See Section 2.6.1.

OAM Capabilities and Performance

- T#27 Determine maximum sustainable rates of BFD traffic. If BFD requires CPU intervention, determine both maximum rates and CPU loading when multiple interfaces are active.
- T#28 Verify LSP Ping and LSP Traceroute capability.
- T#29 Determine maximum rates of MPLS-TP CC-CV traffic. If CC-CV requires CPU intervention, determine both maximum rates and CPU loading when multiple interfaces are active.
- T#30 Determine MPLS-TP DM precision.

T#31 Determine MPLS-TP LM accuracy.

T#32 Verify MPLS-TP AIS/RDI and PSC functionality, protection speed, and AIS/RDI notification speed when a large number of Management Entities (ME) must be notified with AIS/RDI.

The tests in the "Basic Performance" section of the above list should be repeated under various conditions to retest basic performance when critical capabilities are enabled. Complete repetition of the performance tests enabling each capability and combinations of capabilities would be very time intensive, therefore a reduced set of performance tests can be used to gauge the impact of enabling specific capabilities.

5. Acknowledgements

Numerous very useful comments have been received in private email. Some of these contributions are acknowledged here, approximately in chronologic order.

Paul Doolan provided a brief review resulting in a number of clarifications, most notably regarding on-chip vs. system buffering, 100 Gb/s link speed assumptions in the 150 Mpps figure, and handling of large microflows. Pablo Frank reminded us of the sawtooth effect in PPS vs. packet size graphs, prompting the addition of a few paragraphs on this. Comments from Lou Berger at IETF-85 prompted the addition of Section 2.7.

Valuable comments were received on the BMWG mailing list. Jay Karthik pointed out extraneous methodology hints that belong in an appendix or should be removed.

Nabil Bitar pointed out the need to cover QoS (Differentiated Services), MPLS multicast (P2MP and MP2MP), and MPLS-TP OAM. Nabil also provided a number of clarifications to the questions and tests in Section 3 and Section 4.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This document reviews forwarding behavior specified elsewhere and points out compliance and performance requirements. As such it

introduces no new security requirements or concerns.

Knowledge of potential performance shortcomings may serve to help new implementations avoid pitfalls. It is unlikely that such knowledge could be the basis of new denial of service as these pitfalls are already widely known in the service provider community and among leading equipment suppliers. In practice extreme data and packet rate are needed to affect existing equipment and networks that may be still vulnerable due to failure to implement adequate protection and make this type of denial of service unlikely and make undetectable denial of service of this type impossible.

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.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.
- [RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4182] Rosen, E., "Removing a Restriction on the use of MPLS Explicit NULL", RFC 4182, September 2005.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson,

"Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.

- [RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS", RFC 5129, January 2008.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

8.2. Informative References

- [ACK-compression]
"Observations and Dynamics of a Congestion Control Algorithm: The Effects of Two-Way Traffic", Proc. ACM SIGCOMM, ACM Computer Communications Review (CCR) Vol 21, No 4, 1991, pp.133-147., 1991.
- [ATM-EPD-and-PPD]
"Dynamics of TCP Traffic over ATM Networks", IEEE Journal of Special Areas of Communication Vol 13 No 4, May 1995, pp. 633-641., May 1995.
- [I-D.ietf-pwe3-mpls-eth-oam-iwk]
Mohan, D., Bitar, N., and A. Sajassi, "MPLS and Ethernet OAM Interworking", draft-ietf-pwe3-mpls-eth-oam-iwk-07 (work in progress), January 2013.
- [I-D.ietf-tictoc-1588overmpls]
Davari, S., Oren, A., Bhatia, M., Roberts, P., and L. Montini, "Transporting Timing messages over MPLS Networks", draft-ietf-tictoc-1588overmpls-03 (work in progress), October 2012.
- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474,

December 1998.

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3429] Ohta, H., "Assignment of the 'OAM Alert Label' for Multiprotocol Label Switching Architecture (MPLS) Operation and Maintenance (OAM) Functions", RFC 3429, November 2002.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4110] Callon, R. and M. Suzuki, "A Framework for Layer 3 Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4110, July 2005.
- [RFC4124] Le Faucheur, F., "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4221] Nadeau, T., Srinivasan, C., and A. Farrel, "Multiprotocol Label Switching (MPLS) Management Overview", RFC 4221, November 2005.
- [RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC4950] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "ICMP Extensions for Multiprotocol Label Switching", RFC 4950, August 2007.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.
- [RFC5695] Akhter, A., Asati, R., and C. Pignataro, "MPLS Forwarding Benchmarking Methodology for IP Flows", RFC 5695, November 2009.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [RFC5885] Nadeau, T. and C. Pignataro, "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.
- [RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network

Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

- [RFC6310] Aissaoui, M., Busschbach, P., Martini, L., Morrow, M., Nadeau, T., and Y(J). Stein, "Pseudowire (PW) Operations, Administration, and Maintenance (OAM) Message Mapping", RFC 6310, July 2011.
- [RFC6371] Busi, I. and D. Allan, "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, September 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.
- [RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing Label Switched Path Ping (LSP Ping) over MPLS Tunnels", RFC 6424, November 2011.
- [RFC6425] Saxena, S., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, November 2011.
- [RFC6426] Gray, E., Bahadur, N., Boutros, S., and R. Aggarwal, "MPLS On-Demand Connectivity Verification and Route Tracing", RFC 6426, November 2011.
- [RFC6427] Swallow, G., Fulignoli, A., Vigoureux, M., Boutros, S., and D. Ward, "MPLS Fault Management Operations, Administration, and Maintenance (OAM)", RFC 6427, November 2011.
- [RFC6428] Allan, D., Swallow Ed. , G., and J. Drake Ed. , "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, November 2011.
- [RFC6435] Boutros, S., Sivabalan, S., Aggarwal, R., Vigoureux, M.,

and X. Dai, "MPLS Transport Profile Lock Instruct and Loopback Functions", RFC 6435, November 2011.

- [RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, November 2011.
- [RFC6478] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", RFC 6478, May 2012.
- [RFC6639] King, D. and M. Venkatesan, "Multiprotocol Label Switching Transport Profile (MPLS-TP) MIB-Based Management Overview", RFC 6639, June 2012.
- [RFC6669] Sprecher, N. and L. Fang, "An Overview of the Operations, Administration, and Maintenance (OAM) Toolset for MPLS-Based Transport Networks", RFC 6669, July 2012.
- [RFC6670] Sprecher, N. and KY. Hong, "The Reasons for Selecting a Single Solution for MPLS Transport Profile (MPLS-TP) Operations, Administration, and Maintenance (OAM)", RFC 6670, July 2012.
- [RFC6720] Pignataro, C. and R. Asati, "The Generalized TTL Security Mechanism (GTSM) for the Label Distribution Protocol (LDP)", RFC 6720, August 2012.
- [RFC6829] Chen, M., Pan, P., Pignataro, C., and R. Asati, "Label Switched Path (LSP) Ping for Pseudowire Forwarding Equivalence Classes (FECs) Advertised over IPv6", RFC 6829, January 2013.

Appendix A. Organization of References Section

The References section is split into Normative and Informative subsections. References that directly specify forwarding encapsulations or behaviors are listed as normative. References which describe signaling only, though normative with respect to signaling, are listed as informative. They are informative with respect to MPLS forwarding.

Authors' Addresses

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting, LLC

Email: curtis@occnc.com

Kireeti Kompella
Contrail Systems

Email: kireeti.kompella@gmail.com

Shane Amante
Level 3 Communications, Inc.
1025 Eldorado Blvd
Broomfield, CO 80021

Email: shane@level3.net

Andrew Malis
Verizon
60 Sylvan Road
Waltham, MA 02451

Phone: +1 781-466-2362
Email: andrew.g.malis@verizon.com

Carlos Pignataro
Cisco Systems
7200-12 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: cpignata@cisco.com

TRILL Working Group
INTERNET-DRAFT
Intended status: Proposed Standard

Lucy Yong
Donald Eastlake
Sam Aldrin
Huawei Technologies
Jon Hudson
Brocade
February 18, 2013

Expires: August 17, 2013

TRILL Over Pseudo Wires
<draft-yong-pwe3-trill-o-pw-00.txt>

Abstract

This document describes ways to interconnect a pair of TRILL (Transparent Interconnection of Lots of Links) switch ports with two types of pseudo wires under existing TRILL and PWE3 (pseudowire Emulation End-to-End) standards.

Status of This Memo

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

Distribution of this document is unlimited. Comments should be sent to the authors.

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>.

Table of Contents

1. Introduction.....	3
1.1 Conventions used in this document.....	3
2. PWE3 Interconnection of TRILL Switches.....	4
2.1 PWE3 Type Independent Details.....	4
2.2 TRILL over PPP PWE3.....	4
2.3 TRILL over Ethernet PWE3.....	5
2.4 Preferable Pseudowire Type And Auto-Configuration.....	5
3. IANA Considerations.....	6
4. Security Considerations.....	6
Acknowledgements.....	7
Normative References.....	7
Informative References.....	7
Authors' Addresses.....	9

1. Introduction

The IETF has standardized the TRILL (TRansparent Interconnection of Lots of Links) protocol [RFC6325] that provides optimal pair-wise data frame routing without configuration in multi-hop networks with arbitrary topology. TRILL supports multipathing of both unicast and multicast traffic. Devices that implement TRILL are called TRILL Switches or RBridges (Routing Bridges).

End stations are attached to TRILL switches with Ethernet. But links between TRILL switches can be based on arbitrary link protocols, for example PPP [RFC6361], as well as Ethernet [RFC6325]. A set of connected TRILL switches form a TRILL campus which is bounded by end stations and layer 3 routers. Such a campus may contain bridges.

This document specified the use of two types of PWE3 (Pseudowire Emulation End-to-End) pseudowires as links between TRILL switches. It is assumed that such pseudowires are implemented with MPLS.

1.1 Conventions used in this document

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

Acronyms used in this document include the following:

IS-IS - Intermediate System to Intermediate System [IS-IS]
[RFC1195]

MPLS - Multi-Protocol Label Switching

PPP - Point-to-Point Protocol

PW - Pseudowire

PWE3 - PW Emulation End-to-End

RBridge - Routing Bridge, an alternative name for a TRILL Switch

TRILL - Transparent Interconnection of Lots of Links [RFC6325]

TRILL Switch - A device implementing the TRILL protocol

2. PWE3 Interconnection of TRILL Switches

PPP [RFC4618] or Ethernet [RFC4448] pseudowires may be used to interconnect pairs of TRILL switch ports as described below. The pseudowire between such ports can be auto-configured [RFC4447] or manually configured. The TRILL switches, which are TRILL routers, are also acting as label switched routers for those TRILL switch ports.

In both types, the pseudowire provides transparent transport and the two RBridges appear directly interconnected with a transparent link. With such an interconnection (and negotiation to use TRILL in the PPP case [RFC6361]), the TRILL adjacency over that link is automatically discovered and established through TRILL IS-IS control messages [RFC6325] [RFC6327].

2.1 PWE3 Type Independent Details

The sending pseudowire TRILL switch port MUST copy the priority of the TRILL packets being sent to the 3-bit Class of Service field of the pseudowire label [RFC5462] so the priority will be visible to transit devices that can take the priority into account.

If a pseudowire supports fragmentation and re-assembly, there is no reason to do TRILL MTU testing on it and the pseudowire will not be a constraint on the TRILL campus wide Sz (see Section 4.3.1 [RFC6325]). If the pseudowire does not support fragmentation, then the available TRILL IS-IS packet payload size over the pseudowire (taking into account MPLS encapsulation with a control word) or some lower value, MUST be used in helping to determine Sz (see Section 5 [ClearCorrect]).

An intervening MPLS label switched router or similar device has no awareness of TRILL. Such devices will not change the TRILL Header hop count.

2.2 TRILL over PPP PWE3

For a PPP pseudowire (PW type = 0x0007), the two TRILL switch ports being connected are configured to form a pseudowire with PPP encapsulation [RFC4618]. After the pseudowire is established and TRILL use is negotiated within PPP, the two TRILL switches then appear directly connected with a PPP link [RFC1661].

Behavior for TRILL with a PPP pseudowire continues to follow that of TRILL over PPP as specified in Section 3 of [RFC6361].

2.3 TRILL over Ethernet PWE3

For an Ethernet pseudowire, the two TRILL switch ports being connected are configured to form a pseudowire with Ethernet encapsulation [RFC4448]. The ports MUST use the Raw mode (PW type = 0x0005) and non-service-delimiting, to provide as transparent an Ethernet transport as practical. The two RBridges then appear directly interconnected with an Ethernet link [RFC6325].

Behavior for TRILL with an Ethernet psuedo wire continue to follow that over Ethernet as specified in [RFC6325] and [RFC6327].

2.4 Preferable Pseudowire Type And Auto-Configuration

Use of the PPP pseduowire type is preferable to the Ethernet pseudowire type for the connections discussed in this document. It saves 12 or 16 bytes on every TRILL packet. In particular, the Link Header in the PPP case is simply a 2-byte PPP code point while for the Ethernet case it is 14 or 18 bytes (Outer.MacDA (6), Outer.MacSA (6), sometimes Outer.VLAN (4), and TRILL Ethertype (2)). (While it would also be possible to specify a special custom pseudowire type for TRILL traffic, the authors feel that any efficiency gain over PPP pseudowires would be too small to be worth the complexity of adding such a specification.)

If pseudowire interconnection of two TRILL switch ports is auto-configured [RFC4447] and the initiating RBridge port supports PPP pseudowires, it SHOULD initially attempt the connection set-up with PW type PPP (0x0007). If that pseudowire type is rejected, it SHOULD try again with the Ethernet PW type recommended above (0x0005) if it supports that type. If a responding RBridge port receives a set-up attempt specifying PPP, it SHOULD accept the connection if it supports PPP. If a responding RBridge port receives a set-up attempt specifying Ethernet (PW type = 0x0005), it SHOULD assume that the initiator does not support PPP and accept or reject the Ethernet set-up attempt depending on whether or not it supports Ethernet. SHOULD is specified because local policy as to what pseudowires connections and types are allowed may override these guidelines.

3. IANA Considerations

No IANA action is required by this document. RFC Editor: Please remove this section before publication.

4. Security Considerations

For general TRILL protocol security considerations and those related to Ethernet links, see [RFC6325].

For PPP link TRILL security considerations, see [RFC6361].

For security considerations introduced by carrying Ethernet or PPP TRILL links over pseudowires, see [RFC3985].

Not all implementations need to include specific security mechanisms at the pseudowire layer, for example if they are designed to be deployed only in cases where the networking environment is trusted or where other layers provide adequate security. A complete enumeration of possible deployment scenarios and associated threats and options is not possible and is outside the scope of this document. For applications involving sensitive data, end-to-end security should always be considered, in addition to link security, to provide security in depth.

Acknowledgements

The document was prepared in raw nroff. All macros used were defined within the source file.

Normative References

- [RFC1661] - Simpson, W., Ed., "The Point-to-Point Protocol (PPP)", STD 51, RFC 1661, July 1994.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4447] Martini, L., Ed., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [RFC4618] Martini, L., "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) over MPLS Networks", BCP 116, RFC 4618, September 2006.
- [RFC5462] - Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (Rbridges): Base Protocol Specification", RFC6325, July 2011.
- [RFC6361] - Carlson, J., and D. Eastlake, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC6361, August 2011.
- [ClearCorrect] - Eastlake, D., M. Zhang, A. Ghanwani, V. Manral, and A. Banerjee, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-clear-correct, in RFC Editor's queue.

Informative References

- [IS-IS] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for

providing the connectionless-mode Network Service (ISO 8473)",
ISO/IEC10589:2002, Second Edition, Nov 2002

[RFC1195] - Callon, R., "Use of OSI IS-IS for routing in TCP/IP and
dual environments", RFC 1195, December 1990.

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

[RFC6327] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Dutt, D.,
and V. Manral, "Routing Bridges (RBridges): Adjacency", RFC
6327, July 2011.

Authors' Addresses

Lucy Yong
Huawei R&D USA
5340 Legacy Drive
Plano, TX 75025 USA

Phone: +1-469-227-5837
Email: lucy.yong@huawei.com

Donald E. Eastlake, 3rd
Huawei R&D USA
155 Beaver Street
Milford, MA 01757 USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Sam Aldrin
Huawei R&D USA
2330 Central Expressway
Santa Clara, CA 95050 USA

Phone: +1-408-330-4517
Email: sam.aldrin@huawei.com

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134 USA

Phone: +1-408-333-4062
jon.hudson@brocade.com

Copyright and IPR Provisions

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. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

