Internet Engineering Task Force   Yimin Shen
Internet-Draft   Juniper Networks
Intended status: Standards Track   Rahul Aggarwal
Expires: July 25, 2015   Arktan, Inc
                                                 Wim Henderickx
                                                 Alcatel-Lucent
                                                 Yuanlong Jiang
                                             Huawei Technologies
                                                January 21, 2015

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

Abstract

   This document specifies a fast mechanism for protecting pseudowires
   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 the router upstream adjacent to a failure. In
   particular, the router can restore PW traffic in the order of tens of
   milliseconds, by transmitting the traffic to a protector through a
   pre-established bypass tunnel. Therefore, the mechanism can reduce
   traffic loss before global repair reacts to the failure and the
   network converges on the topology changes due to the failure.

Status of This Memo

Copyright Notice

Table of Contents

1. **Introduction**

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

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

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

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

   a. Egress AC failure.

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

   c. Switching PE failure: Node failure of an S-PE of an MS-PW.

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

2. Specification of Requirements

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

3. Reference Models and Failure Cases

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

3.1. Single-Segment PW

```
                    |<-------------- PW1 --------------->|

               - PE1 -------------- P1 --------------- PE2 -
              /                                             \
             /                                               \
        CE1                                                   CE2
             \                                               /
              \                                             /
               - PE3 -------------- P2 --------------- PE4 -

                    |<-------------- PW2 --------------->|
```
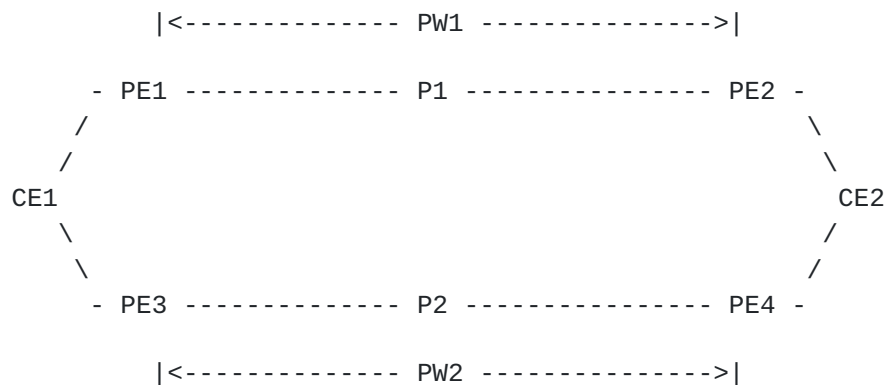
                              Figure 1

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

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

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

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

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

      Primary PW: PW1

Primary PE: PE2

Primary AC: CE2-PE2

Backup PW: PW2

Backup PE: PE4

Backup AC: CE2-PE4

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

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

```
               |<------------- PW1 --------------->|

                 ------------ P1 --------------- PE2 -
               /                                        \
              /                                          \
        CE1 -- PE1                                        CE2
              \                                          /
               \                                        /
                 ------------ P2 --------------- PE4 -

               |<------------- PW2 --------------->|
```
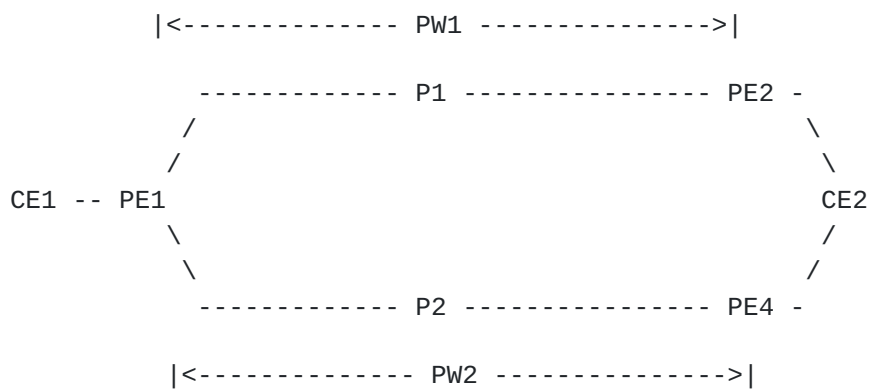
                            Figure 2

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

3.2. Multi-Segment PW

```
                  |<-------------- PW1 --------------->|
                  |<----- SEG1 ----->|<----- SEG2 ----->|

            - TPE1 -------------- SPE1 -------------- TPE2 -
           /                                                \
          /                                                  \
      CE1                                                      CE2
          \                                                  /
           \                                                /
            - TPE3 -------------- SPE2 -------------- TPE4 -

                  |<----- SEG3 ----->|<----- SEG4 ----->|
                  |<-------------- PW2 --------------->|
```
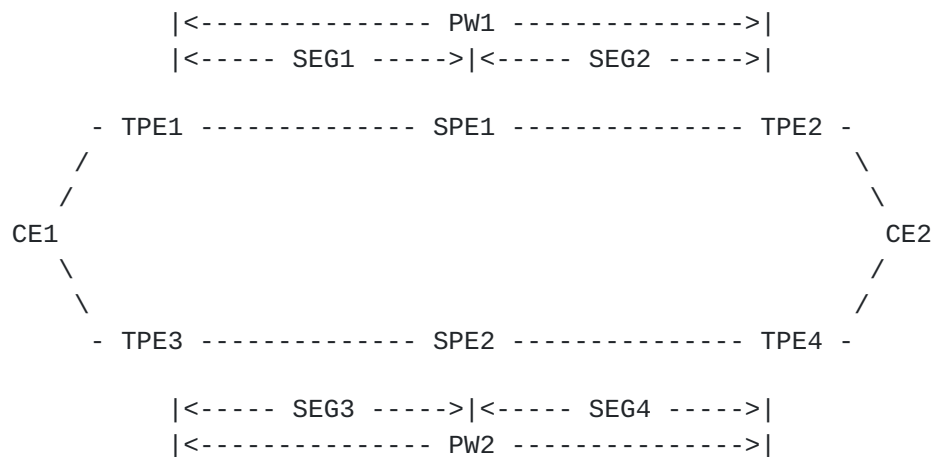
                              Figure 3

   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
   S-PE 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 S-PE failure, as described later in this document.

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

4. Theory of Operation

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

a. Egress AC protection

b. Egress (T-)PE node protection

c. S-PE node protection

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

4.1. Local Repair and Protector

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

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

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

```
                 |<------------- PW1 --------------->|

         - PE1 ------------- P1 --------------- PE2 -
        /                                       PLR  \
       /                                        |     \
    CE1                                  bypass|      CE2
       \                                        |     /
        \                                       |    /
         - PE3 ------------- P2 --------------- PE4 -
                                                protector

                 |<------------- PW2 --------------->|
```

                            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).

```
                 |<------------- PW1 --------------->|

         - PE1 ------------- P1 ------- P3 ----- PE2 -
        /                             PLR \         \
       /                                   \         \
    CE1                                bypass\        CE2
       \                                      \       /
        \                                      \     /
         - PE3 ------------- P2 --------------- PE4 -
                                                protector

                 |<------------- PW2 --------------->|
```
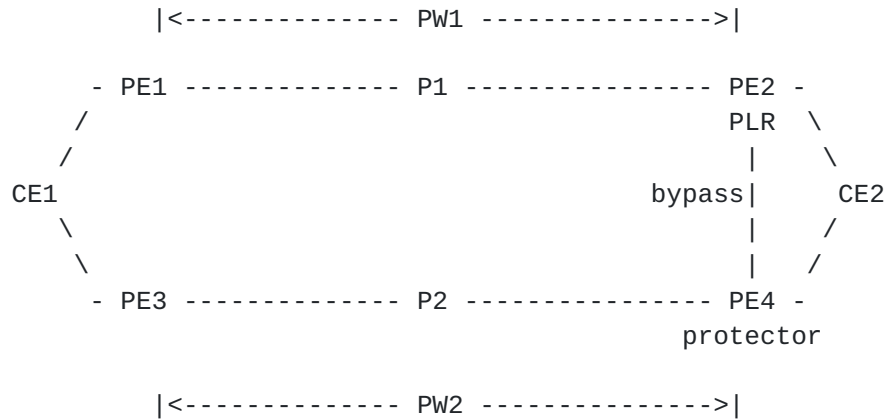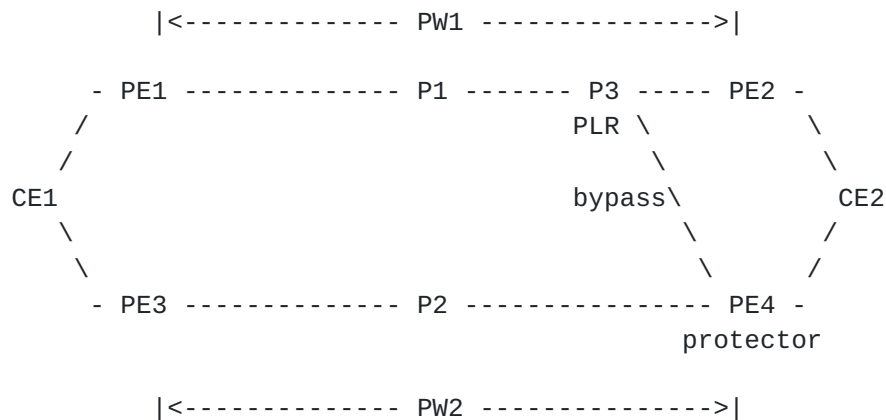
                            Figure 5

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).

```
              |<-------------- PW1 --------------->|
              |<----- SEG1 ----->|<----- SEG2 ----->|

          - TPE1 ----- P4  ----- SPE1 ------------- TPE2 -
         /            PLR \                            \
        /                  \                            \
     CE1                bypass\                           CE2
        \                  \                            /
         \                  \                          /
          - TPE3 -------------- SPE2 ------------- TPE4 -
                           protector

              |<----- SEG3 ----->|<----- SEG4 ----->|
              |<-------------- PW2 --------------->|
```

                            Figure 6

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

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

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

   The above examples have shown the scenarios where the protectors are
   backup (T/S-)PEs. In other scenarios, a protector may be a dedicated
   router that assumes such role, separate from the backup (T/S-)PE of a
   primary PW. During local repair, the PLR MUST still reroute traffic
   to the protector through a bypass tunnel. The protector MUST in turn
   send the traffic to the backup (T/S-)PE, which MUST then send the

   traffic to the target CE via a backup AC or a backup PW segment.
   More detail will be described in Section 4.3.

4.2. Context Identifier

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

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

4.2.1. Semantics

   The semantics of a context identifier is twofold.

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

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

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

      In addition to the normal LDP PW signaling, the primary PE MUST
      have a targeted LDP session with the protector, and advertise PW
      labels to the protector via LDP Label Mapping messages (See

Section 5 for detail). The primary PE MUST attach the context
identifier to each message. Upon receiving the message, the
protector MUST install the advertised PW label in the label space
identified by the context identifier.

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

4.2.2. Advertisement and Path Computation

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

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

o For the bypass tunnel, the PLR MUST choose the protector as the
   actual endpoint. In egress (T-)PE node protection and S-PE node
   protection, the bypass tunnel MUST avoid the primary (S-)PE.

The detail of how the primary PE and the protector may advertise a
context identifier is independent of this mechanism and out of the
scope of this document. One approach would be to advertise it as a
virtual proxy node connected to both routers, with the link between
the proxy node and the primary PE having a more preferable IGP or TE
metric than the link between the proxy node and the protector. The
ultimate goal is for a path computation algorithm, such as CSPF
(constrained shortest path first), LFA (RFC 5286) and MRT ([IP-LDP-
FRR-MRT]), to be able to compute the paths that meet the above
requirements.

4.3. Protection Models

There are two protection models based on the location of a protector.
A network MAY use either 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.

```
                 |<------------- PW1 --------------->|

           - PE1 ------------- P1 ------- P3 ----- PE2 ----
          /                             PLR \     PLR      \
         /                                   \    |         \
     CE1                                 bypass\   |bypass  CE2
        \                                       \  |       /
         \                                       \ |      /
          - PE3 ------------- P2 --------------- PE4 ----
                                                 protector

                 |<------------- PW2 --------------->|
```

                           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 in turn 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.

```
             |<-------------- PW1 --------------->|
             |<----- SEG1 ----->|<----- SEG2 ----->|

         - TPE1 ----- P4  ----- SPE1 ------------- TPE2 -
        /             PLR \                             \
       /                   \                             \
    CE1                  bypass\                          CE2
       \                   \                             /
        \                   \                           /
         - TPE3 -------------- SPE2 ------------- TPE4 -
                        protector

             |<----- SEG3 ----->|<----- SEG4 ----->|
             |<-------------- PW2 --------------->|
```
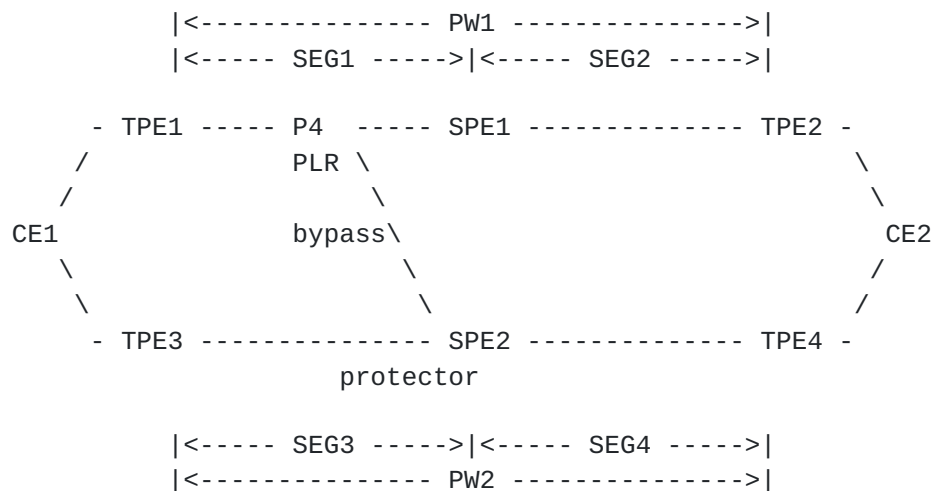
Figure 8

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

4.3.2. Centralized Protector

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

In egress AC protection and egress PE node protection, when the
protector receives traffic from the PLR, if the protector has a
direct connection (i.e. backup AC) to the CE, it MUST forward the
traffic to the CE via the backup AC, which is similar to Figure 7.
Otherwise, it MUST forward the traffic to a backup PE, which MUST
then forward the traffic to the CE via a backup AC. This is shown in
Figure 9, where the protector receives traffic from P3 (the PLR for
egress PE failure) or PE2 (the PLR for egress AC failure) 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 for
clarity.

```
               |<------------- PW1 --------------->|


             - PE1 ------------- P1 ------- P3 ----- PE2 --
            /                              PLR \     PLR   \
           /                                  \    /        \
          /                              bypass\  /bypass \
         /                                    \ /           \
      CE1                                    protector      CE2
         \                                      \          /
          \                                      \        /
           \                                      \      /
            \                                      \    /
             - PE3 ------------- P2 ----------------PE4 --

               |<------------- PW2 --------------->|
```
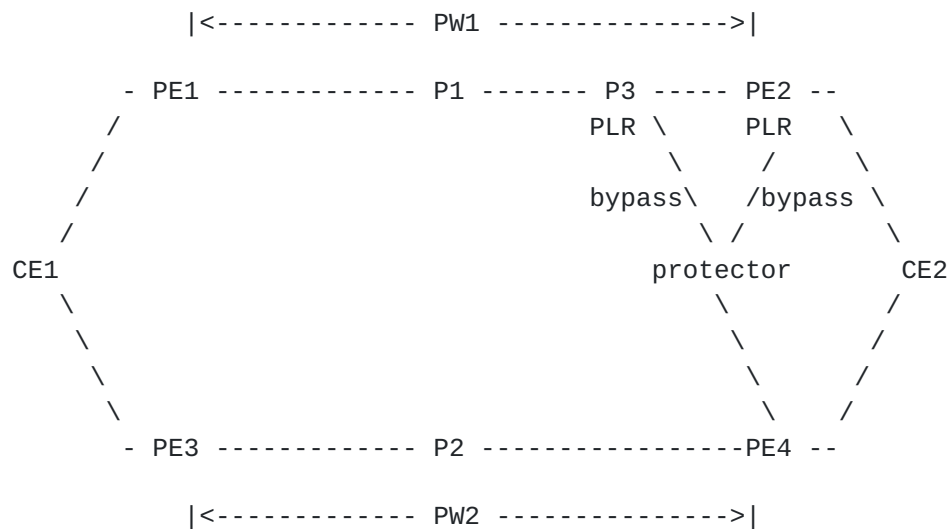
Figure 9

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), SPE2 forwards the traffic to TPE4 via SEG4,
and TPE4 finally forwards traffic to CE2. The protector may be
protecting other PW segments as well, which is not shown in this
figure for clarity.

```
                |<--------------- PW1 --------------->|
                |<----- SEG1 ----->|<----- SEG2 ----->|

          - TPE1 ----- P4  ----- SPE1 ------------- TPE2 -
          /            PLR \                              \
        /                   \                              \
       /            bypass\                                 \
     /                     \                                 \
   CE1                        protector                       CE2
     \                     \                                 /
       \                    \                               /
        \                    \                             /
         \                    \                           /
          - TPE3 -------------- SPE2 -------------- TPE4 -

                |<----- SEG3 ----->|<----- SEG4 ----->|
                |<-------------- PW2 --------------->|
```
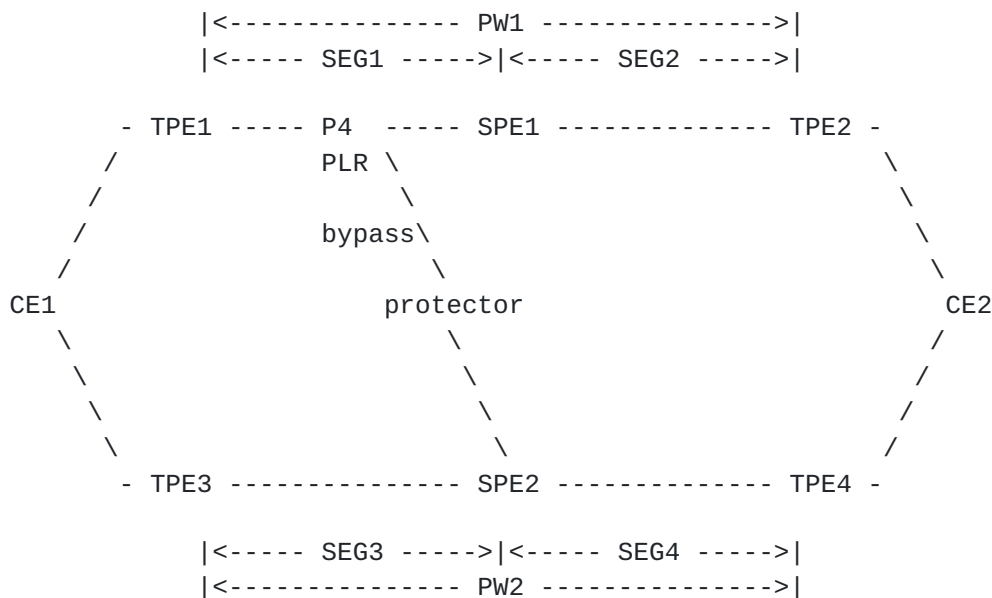
                           Figure 10

   The centralized protector model provides the convenience for multiple
   primary PEs to share one protector. Each primary PE MAY only need
   the one protector to protect all of its PWs. From the perspective of
   scalability, the number of context identifiers needed by a network
   MAY be bound to the number of primary PEs.

4.4. Transport Tunnel

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

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

4.5. **Bypass Tunnel**

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

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

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

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

4.6. **Forwarding State on Protector**

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

4.6.1. Examples of Co-located Protector

   In Figure 7, PE4 is a co-located protector that protects PW1 against
   egress AC failure and egress node failure. It maintains a label
   space for PE2, which is identified by the context identifier of {PE2,
   PE4}. It learns PW1's label from PE2, and installs an forwarding
   entry for the label in that label space. The nexthop of the
   forwarding entry indicates a label pop with outgoing interface
   pointing to the backup AC CE2-PE4.

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

4.6.2. Examples of Centralized Protector

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

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

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

5. LDP Extensions

   As described in previous sections, a targeted LDP session MUST be
   established between each pair of primary PE and protector. The
   primary PE sends Label Mapping message over this session to advertise
   primary PW labels to the protector. In the centralized protector

model, a targeted LDP session MUST also be established between a
backup (S-)PE and a protector. The backup PE sends Label Mapping
message over this session to advertise backup PW labels to the
protector.

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

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

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

5.1. Egress Protection Capability TLV

A protector MUST advertise the Egress Protection Capability TLV in
its Initialization message and Capability message, over the LDP
session with a primary PE. In the centralized protector model, the
protector MUST also advertise the TLV over the LDP session with a
backup PE. The TLV carries one or multiple context identifiers. To
the primary PE, the TLV SHOULD carry the context identifier of the
{primary PE, protector}. In the centralized protector model, the TLV
SHOULD carry to the backup PE multiple context identifiers, one for
each {primary PE, protector} where the backup PE serves as a backup
for the primary PE. This TLV SHOULD NOT be advertised by the primary
PE or the backup PE to the protector.

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

Protection FEC TLVs, when the protector has withdrawn a previously
advertised context identifier or the entire Egress Protection
Capability TLV via Capability message.

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

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |U|F|  Egress Protection (TBD)  |             Length            |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |S| Reserved     |                                               |
 +-+-+-+-+-+-+-+-+-+                                               |
 |                                                               |
 ~               Capability Data = context identifier(s)         ~
 |                                                               |
 |                                       +-+-+-+-+-+-+-+-+-+     |
 |                                       |                       |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
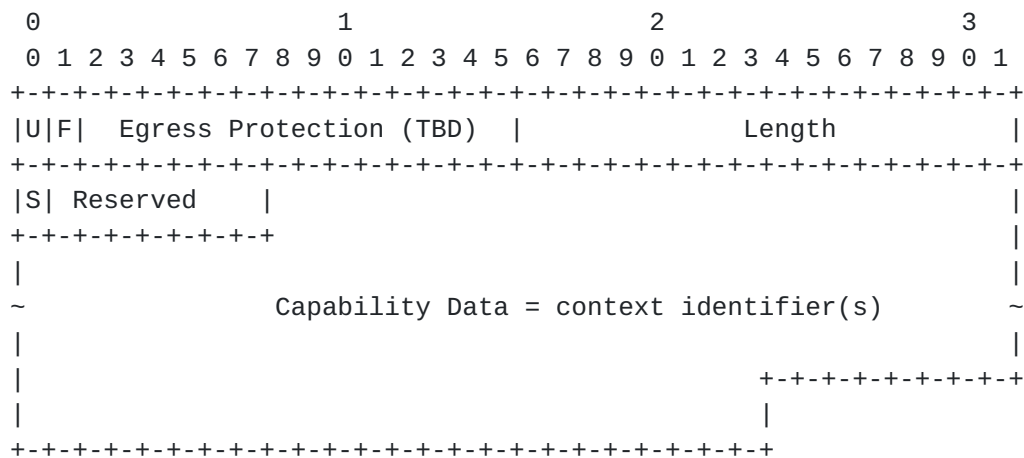
Figure 11

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

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

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

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

The "Capability Data" is encoded with the context identifier of the
{primary PE, protector}.

5.2. PW Label Distribution from Primary PE to Protector

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

6389, [RFC 3471](#)) encoded with the context identifier of the {primary PE, protector}.

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

[5.3](#). PW Label Distribution from Backup PE to Protector

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

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

[5.4](#). Protection FEC Element TLV

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

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   Type(0x83)  |    Reserved   | Encoding Type |    Length     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                                                               |
   |                                                               |
   ~                       PW Information                          ~
   |                                                               |
   |                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                              Figure 12
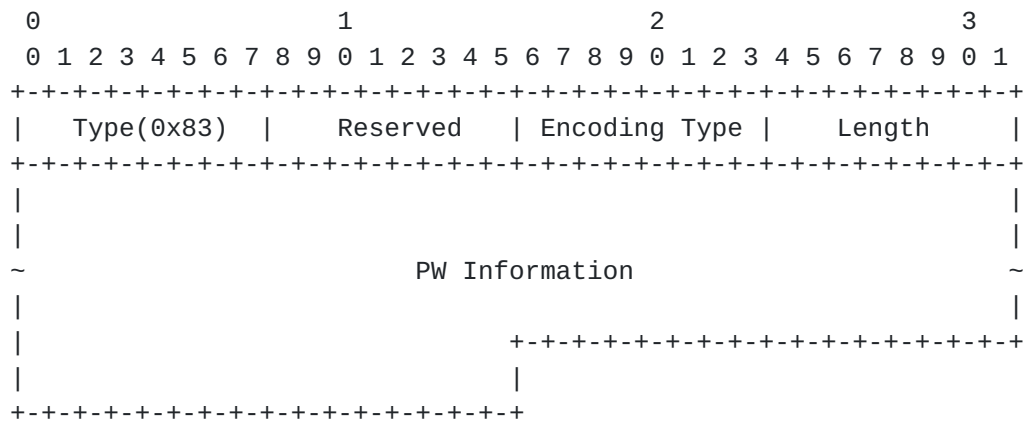
   - Encoding Type

      Type of format that PW Information field is encoded.

   - Length

      Length of PW Information field in octets.

   - PW Information

      Field of variable length that specifies a PW

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

5.4.1. Encoding Format for PWid

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type(0x83)  |   Reserved  | Enc Type(1) |   Length(16)  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Ingress PE Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Egress PE Address                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Group ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          PW ID                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|C|          PW Type        |              Reserved             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
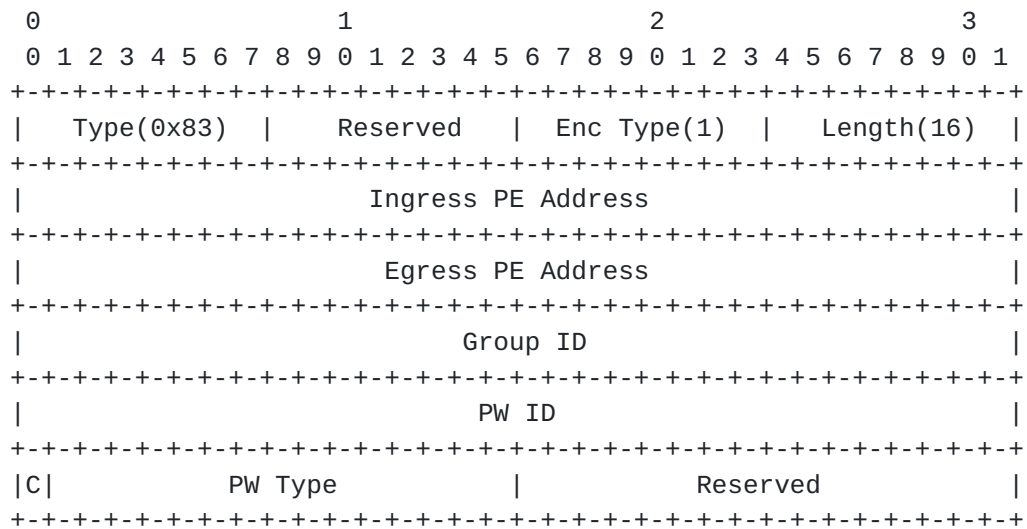
                          Figure 13

- Ingress PE Address

    IP address of the ingress PE of PW.

- Egress PE Address

    IP address of the egress PE of PW.

- Group ID

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

- PW ID

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

- Control word bit (C)

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

- PW Type

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

5.4.2. Encoding Format for Generalized PWid

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Type(0x83)  |   Reserved   |  Enc Type(2)  |   Length      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Ingress PE Address                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Egress PE Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|C|          PW Type          |            Reserved            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   AGI Type   |    Length    |        Value                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     AGI  Value (contd.)                      ~
|                                                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   AII Type   |    Length    |        Value                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     SAII  Value (contd.)                     ~
|                                                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   AII Type   |    Length    |        Value                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     TAII Value (contd.)                      ~
|                                                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

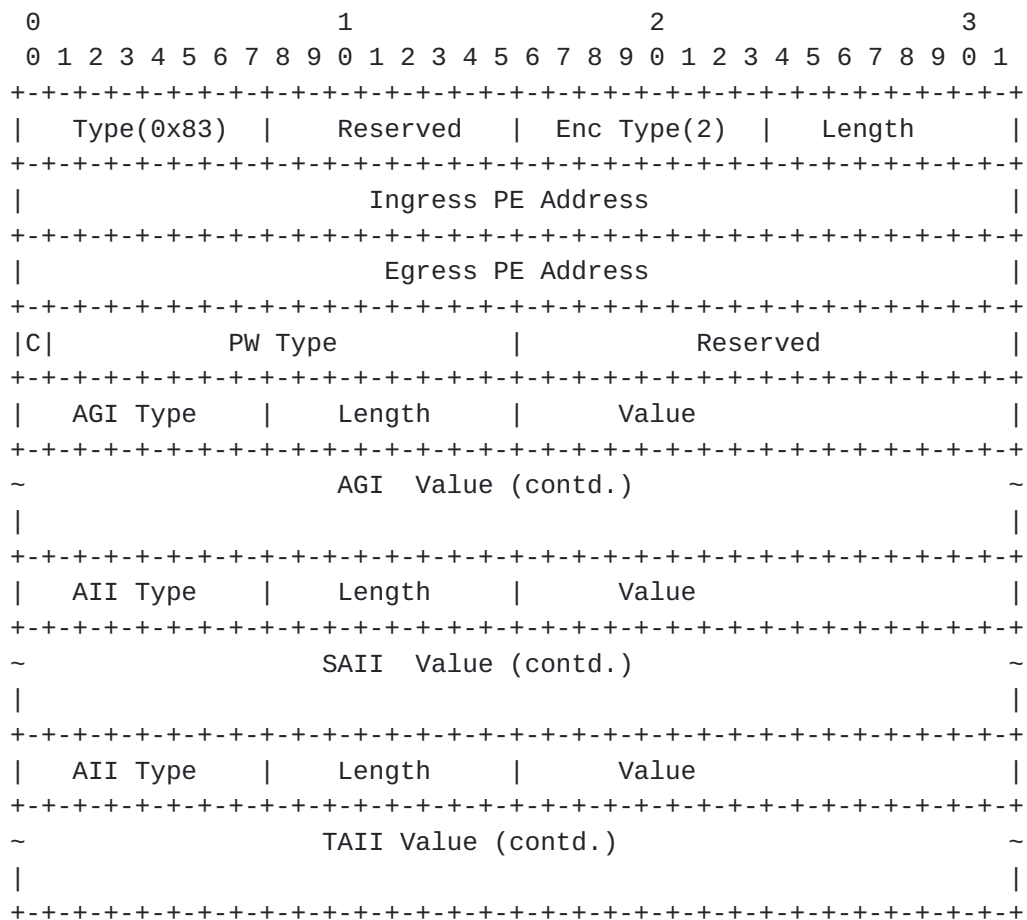   Figure 14

  - Ingress PE Address

      IP address of the ingress PE of PW.

  - Egress PE Address

      IP address of the egress PE of PW.

  - Control word bit (C)

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

  - PW Type

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

   - AGI Type, Length, Value, AGI Value

     Attachment Group Identifier of PW.

   - SAII Type, Length, Value, SAII Value

     Source Attachment Individual Identifier of PW.

   - TAII Type, Length, Value, TAII Value

     Target Attachment Individual Identifier of PW.

6. Revertive Behavior

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

   o Global revertive mode

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

   o Control plane revertive mode

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

   o Local revertive mode

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

the transport tunnel. These procedures are referred to as local reversion.

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

The control plane revertive mode may always happen as part of the convergence of control plane protocols. However, it is only applicable to the specific scenarios described above.

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

7. IANA Considerations

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

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

8. Security Considerations

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

9. Acknowledgements

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

10. References

10.1. Normative References

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

   [RFC5659]   Bocci, M. and S. Bryant, "An Architecture for Multi-
               Segment Pseudowire Emulation Edge-to-Edge", RFC 5659,
               October 2009.

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

   [RFC5331]   Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream
               Label Assignment and Context-Specific Label Space", RFC
               5331, August 2008.

   [RFC5036]   Andersson, L., Minei, I., and B. Thomas, "LDP
               Specification", RFC 5036, October 2007.

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

   [RFC2205]   Braden, B., Zhang, L., Berson, S., Herzog, S., and S.
               Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1
               Functional Specification", RFC 2205, September 1997.

   [RFC3209]   Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,
               and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP
               Tunnels", RFC 3209, December 2001.

   [RFC4090]   Pan, P., Swallow, G., and A. Atlas, "Fast Reroute
               Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May
               2005.

   [RFC5286]   Atlas, A. and A. Zinin, "Basic Specification for IP Fast
               Reroute: Loop-Free Alternates", RFC 5286, September 2008.

   [RFC5714]   Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC
               5714, January 2010.

   [RFC3471]   Berger, L., "Generalized Multi-Protocol Label Switching
               (GMPLS) Signaling Functional Description", RFC 3471,
               January 2003.

   [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-
              Protocol Label Switching (GMPLS) Signaling Constraint-
              based Routed Label Distribution Protocol (CR-LDP)
              Extensions", RFC 3472, January 2003.

   [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol
              Label Switching Architecture", RFC 3031, January 2001.

   [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

   [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection
              (BFD)", RFC 5880, June 2010.

   [RFC6389] Aggarwal, R. and JL. Le Roux, "MPLS Upstream Label
              Assignment for LDP", RFC 6389, November 2011.

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

10.2. Informative References

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

Authors' Addresses

   Yimin Shen
   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

Yuanlong Jiang
Huawei Technologies

Email: jiangyuanlong@huawei.com