Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network

Abstract

   Where the payload of a pseudowire comprises a number of distinct
   flows, it can be desirable to carry those flows over the Equal Cost
   Multiple Paths (ECMPs) that exist in the packet switched network.
   Most forwarding engines are able to generate a hash of the MPLS label
   stack and use this mechanism to balance MPLS flows over ECMPs.

   This document describes a method of identifying the flows, or flow
   groups, within pseudowires such that Label Switching Routers can
   balance flows at a finer granularity than individual pseudowires.
   The mechanism uses an additional label in the MPLS label stack.

Status of This Memo

   This is an Internet Standards Track document.

   This document is a product of the Internet Engineering Task Force
   (IETF). It represents the consensus of the IETF community. It has
   received public review and has been approved for publication by the
   Internet Engineering Steering Group (IESG). Further information on
   Internet Standards is available in Section 2 of RFC 5741.

   Information about the current status of this document, any errata,
   and how to provide feedback on it may be obtained at
   http://www.rfc-editor.org/info/rfc6391.

Copyright Notice

Table of Contents

1. Introduction

   A pseudowire (PW) [RFC3985] is normally transported over one single
   network path, even if multiple Equal Cost Multiple Paths (ECMPs)
   exist between the ingress and egress PW provider edge (PE) equipment
   [RFC4385] [RFC4928]. This is required to preserve the
   characteristics of the emulated service (e.g., to avoid misordering
   Structure-Agnostic Time Division Multiplexing over Packet (SAToP) PW
   packets [RFC4553] or subjecting the packets to unusable inter-arrival
   times). The use of a single path to preserve order remains the
   default mode of operation of a PW. The new capability proposed in
   this document is an OPTIONAL mode that may be used when the use of
   ECMPs is known to be beneficial (and not harmful) to the operation of
   the PW.

   Some PWs are used to transport large volumes of IP traffic between
   routers. One example of this is the use of an Ethernet PW to create
   a virtual direct link between a pair of routers. Such PWs may carry
   from hundreds of Mbps to Gbps of traffic. These PWs only require
   packet ordering to be preserved within the context of each individual
   transported IP flow. They do not require packet ordering to be
   preserved between all packets of all IP flows within the pseudowire.

   The ability to explicitly configure such a PW to leverage the
   availability of multiple ECMPs allows for better capacity planning,
   as the statistical multiplexing of a larger number of smaller flows
   is more efficient than with a smaller set of larger flows.

   Typically, forwarding hardware can deduce that an IP payload is being
   directly carried by an MPLS label stack, and it is capable of looking
   at some fields in packets to construct hash buckets for conversations
   or flows. However, when the MPLS payload is a PW, an intermediate
   node has no information on the type of PW being carried in the
   packet. This limits the forwarder at the intermediate node to only
   being able to make an ECMP choice based on a hash of the MPLS label
   stack. In the case of a PW emulating a high-bandwidth trunk, the
   granularity obtained by hashing the label stack is inadequate for
   satisfactory load balancing. The ingress node, however, is in the
   special position of being able to understand the unencapsulated
   packet header to assist with spreading flows among any available
   ECMPs, or even any Loop-Free Alternates [RFC5286]. This document
   defines a method to introduce granularity on the hashing of traffic
   running over PWs by introducing an additional label, chosen by the
   ingress node, and placed at the bottom of the label stack.

In addition to providing an indication of the flow structure for use
in ECMP forwarding decisions, the mechanism described in the document
may also be used to select flows for distribution over an IEEE
802.1AX-2008 (originally specified as IEEE 802.3ad-2000) Link
Aggregation Group (LAG) that has been used in an MPLS network.

NOTE: Although Ethernet is frequently referenced as a use case in
this RFC, the mechanisms described in this document are general
mechanisms that may be applied to any PW type in which there are
identifiable flows, and in which there is no requirement to preserve
the order between those flows.

1.1. Requirements Language

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

1.2. ECMP in Label Switching Routers

Label Switching Routers (LSRs) commonly generate a hash of the label
stack or some elements of the label stack as a method of
discriminating between flows and use this to distribute those flows
over the available ECMPs that exist in the network. Since the label
at the bottom of the stack is usually the label most closely
associated with the flow, this normally provides the greatest
entropy, and hence is usually included in the hash. This document
describes a method of adding an additional Label Stack Entry (LSE) at
the bottom of the stack in order to facilitate the load balancing of
the flows within a PW over the available ECMPs. A similar design for
general MPLS use has also been proposed [MPLS-ENTROPY]; see Section 9
of this document.

An alternative method of load balancing by creating a number of PWs
and distributing the flows amongst them was considered, but was
rejected because:

o It did not introduce as much entropy as can be introduced by
   adding an additional LSE.

o It required additional PWs to be set up and maintained.

1.3. Flow Label

An additional LSE [RFC3032] is interposed between the PW LSE and the
control word, or if the control word is not present, between the PW
LSE and the PW payload. This additional LSE is called the flow LSE,
and the label carried by the flow LSE is called the flow label.

Indivisible flows within the PW MUST be mapped to the same flow label
by the ingress PE. The flow label stimulates the correct ECMP load-
balancing behaviour in the packet switched network (PSN). On receipt
of the PW packet at the egress PE (which knows a flow LSE is
present), the flow LSE is discarded without processing.

Note that the flow label MUST NOT be an MPLS reserved label (values
in the range 0..15) [RFC3032], but is otherwise unconstrained by the
protocol.

It is useful to give consideration to the choice of Time to Live
(TTL) value in the flow LSE [RFC3032]. The flow LSE is at the bottom
of the label stack; therefore, even when penultimate hop popping is
employed, it will always be preceded by the PW label on arrival at
the PE. If, due to an error condition, the flow LSE becomes the top
of the stack, it might be examined as if it were a normal LSE, and
the packet might then be forwarded. This can be prevented by setting
the flow LSE TTL to 1, thereby forcing the packet to be discarded by
the forwarder. Note that setting the TTL to 1 regardless of the
payload may be considered a departure from the TTL procedures defined
in [RFC3032] that apply to the general MPLS case.

This document does not define a use for the Traffic Class (TC) field
[RFC5462] (formerly known as the Experimental Use (EXP) bits
[RFC3032]) in the flow label. Future documents may define a use for
these bits; therefore, implementations conforming to this
specification MUST set the TC field to zero at the ingress and MUST
ignore them at the egress.

2. Native Service Processing Function

The Native Service Processing (NSP) function [RFC3985] is a component
of a PE that has knowledge of the structure of the emulated service
and is able to take action on the service outside the scope of the
PW. In this case, it is REQUIRED that the NSP in the ingress PE
identify flows, or groups of flows within the service, and indicate
the flow (group) identity of each packet as it is passed to the
pseudowire forwarder. As an example, where the PW type is an
Ethernet, the NSP might parse the ingress Ethernet traffic and
consider all of the IP traffic. This traffic could then be
categorised into flows by considering all traffic with the same
source and destination address pair to be a single indivisible flow.
Since this is an NSP function, by definition, the method used to
identify a flow is outside the scope of the PW design. Similarly,
since the NSP is internal to the PE, the method of flow indication to
the PW forwarder is outside the scope of this document.

3. Pseudowire Forwarder

   The PW forwarder must be provided with a method of mapping flows to
   load-balanced paths.

   The forwarder must generate a label for the flow or group of flows.
   How the flow label values are determined is outside the scope of this
   document; however, the flow label allocated to a flow MUST NOT be an
   MPLS reserved label and SHOULD remain constant for the life of the
   flow. It is RECOMMENDED that the method chosen to generate the load-
   balancing labels introduce a high degree of entropy in their values,
   to maximise the entropy presented to the ECMP selection mechanism in
   the LSRs in the PSN, and hence distribute the flows as evenly as
   possible over the available PSN ECMP. The forwarder at the ingress
   PE prepends the PW control word (if applicable), and then pushes the
   flow label, followed by the PW label.

   NOTE: Although this document does not attempt to specify any hash
   algorithms, it is suggested that any such algorithm should be based
   on the assumption that there will be a high degree of entropy in the
   values assigned to the flow labels.

   The forwarder at the egress PE uses the pseudowire label to identify
   the pseudowire. From the context associated with the pseudowire
   label, the egress PE can determine whether a flow LSE is present. If
   a flow LSE is present, it MUST be checked to determine whether it
   carries a reserved label. If it is a reserved label, the packet is
   processed according to the rules associated with that reserved label;
   otherwise, the LSE is discarded.

   All other PW forwarding operations are unmodified by the inclusion of
   the flow LSE.

3.1. Encapsulation

   The PWE3 Protocol Stack Reference Model modified to include flow LSE
   is shown in Figure 1.

```
      +-------------+                          +-------------+
      | Emulated    |                          | Emulated    |
      | Ethernet    |                          | Ethernet    |
      | (including  |     Emulated Service     | (including  |
      | VLAN)       |<========================>| VLAN)       |
      | Services    |                          | Services    |
      +-------------+                          +-------------+
      | Flow        |                          | Flow        |
      +-------------+         Pseudowire        +-------------+
      |Demultiplexer|<========================>|Demultiplexer|
      +-------------+                          +-------------+
      | PSN         |        PSN Tunnel         | PSN         |
      | MPLS        |<========================>| MPLS        |
      +-------------+                          +-------------+
      | Physical    |                          | Physical    |
      +-----+-------+                          +-----+-------+
```
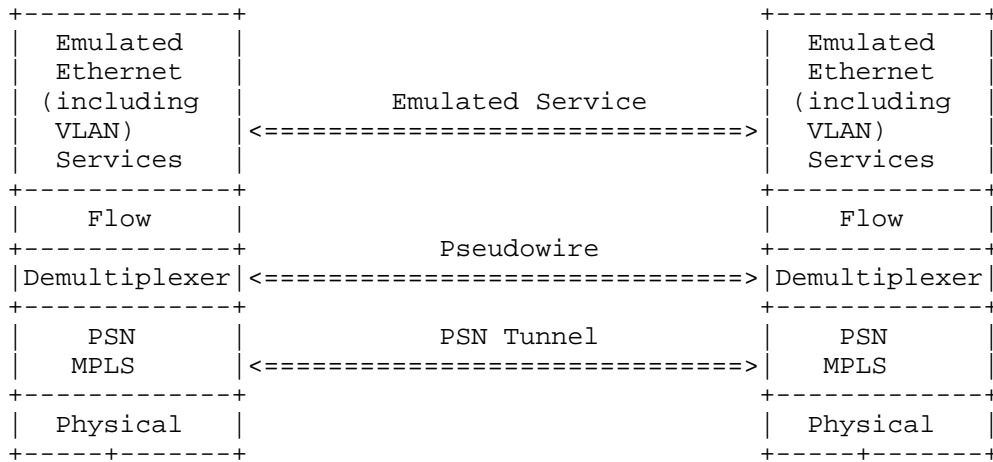
              Figure 1: PWE3 Protocol Stack Reference Model

   The encapsulation of a PW with a flow LSE is shown in Figure 2.

```
      +---------------------------+
      |                           |
      |  Payload                  |
      |                           |  n octets
      |                           |
      +---------------------------+
      | Optional Control Word     |  4 octets
      +---------------------------+
      | Flow LSE                  |  4 octets
      +---------------------------+
      | PW LSE                    |  4 octets
      +---------------------------+
      | MPLS Tunnel LSE (s)       |  n*4 octets (four octets per LSE)
      +---------------------------+
```

    Figure 2: Encapsulation of a Pseudowire with a Pseudowire Flow LSE

4. Signalling the Presence of the Flow Label

   When using the signalling procedures in [RFC4447], a new Pseudowire
   Interface Parameter Sub-TLV, the Flow Label Sub-TLV (FL Sub-TLV), is
   used to synchronise the flow label states between the ingress and
   egress PEs.

   The absence of an FL Sub-TLV indicates that the PE is unable to
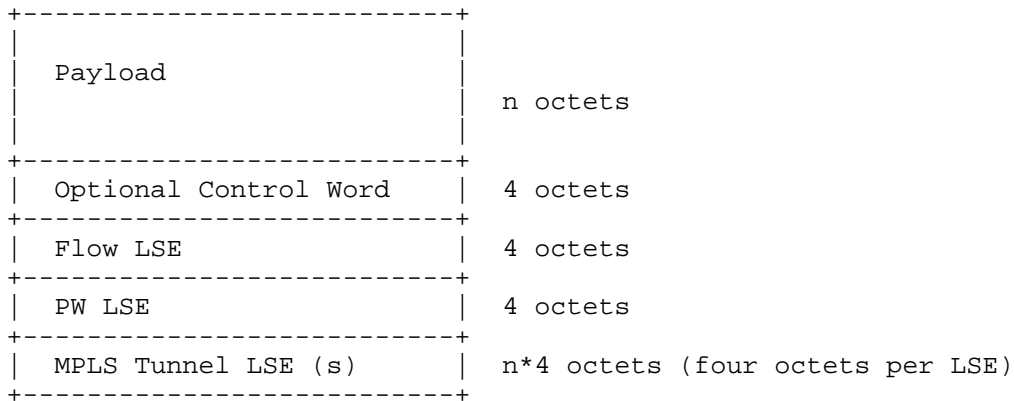   process flow labels. An ingress PE that is using PW signalling and
   that does not send an FL Sub-TLV MUST NOT include a flow label in the
   PW packet. An ingress PE that is using PW signalling and that does
   not receive an FL Sub-TLV from its egress peer MUST NOT include a
   flow label in the PW packet. This preserves backwards compatibility
   with existing PW specifications.

   A PE that wishes to send a flow label in a PW packet MUST include in
   its label mapping message an FL Sub-TLV with T = 1 (see Section 4.1).

   A PE that is willing to receive a flow label MUST include in its
   label mapping message an FL Sub-TLV with R = 1 (see Section 4.1).

   A PE that receives a label mapping message containing an FL Sub-TLV
   with R = 0 MUST NOT include a flow label in the PW packet.

   Thus, a PE sending an FL Sub-TLV with T = 1 and receiving an FL
   Sub-TLV with R = 1 MUST include a flow label in the PW packet. Under
   all other combinations of FL Sub-TLV signalling, a PE MUST NOT
   include a flow label in the PW packet.

   The signalling procedures in [RFC4447] state that "Processing of the
   interface parameters should continue when unknown interface
   parameters are encountered, and they MUST be silently ignored". The
   signalling procedure described here is therefore backwards compatible
   with existing implementations.

   Note that what is signalled is the desire to include the flow LSE in
   the label stack. The value of the flow label is a local matter for
   the ingress PE, and the label value itself is not signalled.

4.1. Structure of Flow Label Sub-TLV

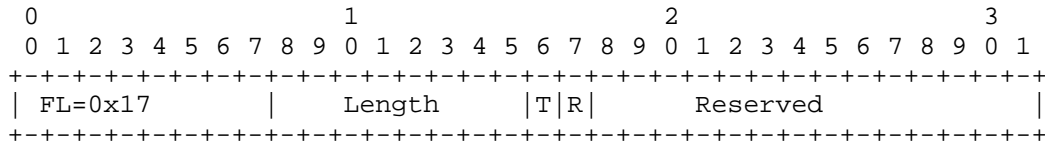   The structure of the Flow Label Sub-TLV is shown in Figure 3.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| FL=0x17       |   Length      |T|R|       Reserved            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                    Figure 3: Flow Label Sub-TLV

   Where:

   o FL (value 0x17) is the Flow Label Sub-TLV identifier assigned by
      IANA (see Section 11).

   o Length is the length of the Sub-TLV in octets and is 4.

   o When T = 1, the PE is requesting the ability to send a PW packet
      that includes a flow label. When T = 0, the PE is indicating that
      it will not send a PW packet containing a flow label.

   o When R = 1, the PE is able to receive a PW packet with a flow
      label present. When R = 0, the PE is unable to receive a PW
      packet with the flow label present.

   o Reserved bits MUST be zero on transmit and MUST be ignored on
      receive.

5. Static Pseudowires

   If PWE3 signalling [RFC4447] is not in use for a PW, then whether the
   flow label is used MUST be identically provisioned in both PEs at the
   PW endpoints. If there is no provisioning support for this option,
   the default behaviour is not to include the flow label.

6. Multi-Segment Pseudowires

   The flow label mechanism described in this document works on
   multi-segment PWs without requiring modification to the Switching PEs
   (S-PEs). This is because the flow LSE is transparent to the label
   swap operation, and because interface parameter Sub-TLV signalling is
   transitive.

7. Operations, Administration, and Maintenance (OAM)

    The following OAM considerations apply to this method of load
    balancing.

    Where the OAM is only to be used to perform a basic test to verify
    that the PWs have been configured at the PEs, Virtual Circuit
    Connectivity Verification (VCCV) [RFC5085] messages may be sent using
    any load balance PW path, i.e., using any value for the flow label.

    Where it is required to verify that a pseudowire is fully functional
    for all flows, a VCCV [RFC5085] connectivity verification message
    MUST be sent over each ECMP path to the pseudowire egress PE. This
    solution may be difficult to achieve and scales poorly. Under these
    circumstances, it may be sufficient to send VCCV messages using any
    load balance pseudowire path, because if a failure occurs within the
    PSN, the failure will normally be detected and repaired by the PSN.
    That is, the PSN's Interior Gateway Protocol (IGP) link/node failure
    detection mechanism (loss of light, bidirectional forwarding
    detection [RFC5880], or IGP hello detection) and the IGP convergence
    will naturally modify the ECMP set of network paths between the
    ingress and egress PEs. Hence, the PW is only impacted during the
    normal IGP convergence time. Note that this period may be reduced if
    a fast re-route or fast convergence technology is deployed in the
    network [RFC4090] [RFC5286].

    If the failure is related to the individual corruption of a Label
    Forwarding Information Base (LFIB) entry in a router, then only the
    network path using that specific entry is impacted. If the PW is
    load-balanced over multiple network paths, then this failure can only
    be detected if, by chance, the transported OAM flow is mapped onto
    the impacted network path, or if all paths are tested. Since testing
    all paths may present problems as noted above, other mechanisms to
    detect this type of error may need to be developed, such as a Label
    Switched Path (LSP) self-test technology.

    To troubleshoot the MPLS PSN, including multiple paths, the
    techniques described in [RFC4378] and [RFC4379] can be used.

    Where the PW OAM is carried out of band (VCCV Type 2) [RFC5085], it
    is necessary to insert an "MPLS Router Alert Label" in the label
    stack. The resultant label stack is as follows:

```
+-----------------------------+
|                             |
|        VCCV Message         |  n octets
|                             |
+-----------------------------+
|   Optional Control Word     |  4 octets
+-----------------------------+
|        Flow LSE             |  4 octets
+-----------------------------+
|         PW LSE              |  4 octets
+-----------------------------+
|     Router Alert LSE        |  4 octets
+-----------------------------+
|     MPLS Tunnel LSE(s)      |  n*4 octets (four octets per label)
+-----------------------------+
```

                    Figure 4: Use of Router Alert Label

   Note that, depending on the number of labels hashed by the LSR, the
   inclusion of the Router Alert label may cause the OAM packet to be
   load-balanced to a different path from that taken by the data packets
   with identical flow and PW labels.

8. Applicability of PWs Using Flow Labels

   A node within the PSN is not able to perform deep packet inspection
   (DPI) of the PW, as the PW technology is not self-describing: the
   structure of the PW payload is only known to the ingress and egress
   PE devices. The method proposed in this document provides a
   statistical mitigation of the problem of load balance in those cases
   where a PE is able to discern flows embedded in the traffic received
   on the attachment circuit.

   The methods described in this document are transparent to the PSN and
   as such do not require any new capability from the PSN.

   The requirement to load-balance over multiple PSN paths occurs when
   the ratio between the PW access speed and the PSN's core link
   bandwidth is large (e.g., >= 10%). ATM and Frame Relay are unlikely
   to meet this property. Ethernet may have this property, and for that
   reason this document focuses on Ethernet. Applications for other
   high-access-bandwidth PWs may be defined in the future.

   This design applies to MPLS PWs where it is meaningful to
   de-construct the packets presented to the ingress PE into flows. The
   mechanism described in this document promotes the distribution of
   flows within the PW over different network paths. In turn, this
   means that whilst packets within a flow are delivered in order

(subject to normal IP delivery perturbations due to topology
variation), order is no longer maintained for all packets sent over
the PW. It is not proposed to associate a different sequence number
with each flow. If sequence number support is required, the flow
label mechanism MUST NOT be used.

Where it is known that the traffic carried by the Ethernet PW is IP,
the flows can be identified and mapped to an ECMP. Such methods
typically include hashing on the source and destination addresses,
the protocol ID and higher-layer flow-dependent fields such as
TCP/UDP ports, Layer 2 Tunneling Protocol version 3 (L2TPv3) Session
IDs, etc.

Where it is known that the traffic carried by the Ethernet PW is
non-IP, techniques used for link bundling between Ethernet switches
may be reused. In this case, however, the latency distribution would
be larger than is found in the link bundle case. The acceptability
of the increased latency is for further study. Of particular
importance, the Ethernet control frames SHOULD always be mapped to
the same PSN path to ensure in-order delivery.

8.1. Equal Cost Multiple Paths

ECMP in packet switched networks is statistical in nature. The
mapping of flows to a particular path does not take into account the
bandwidth of the flow being mapped or the current bandwidth usage of
the members of the ECMP set. This simplification works well when the
distribution of flows is evenly spread over the ECMP set and there
are a large number of flows that have low bandwidth relative to the
paths. The random allocation of a flow to a path provides a good
approximation to an even spread of flows, provided that polarisation
effects are avoided. The method defined in this document has the
same statistical properties as an IP PSN.

ECMP is a load-sharing mechanism that is based on sharing the load
over a number of layer 3 paths through the PSN. Often, however,
multiple links exist between a pair of LSRs that are considered by
the IGP to be a single link. These are known as link bundles. The
mechanism described in this document can also be used to distribute
the flows within a PW over the members of the link bundle by using
the flow label value to identify candidate flows. How that mapping
takes place is outside the scope of this specification. Similar
considerations apply to Link Aggregation Groups.

There is no mechanism currently defined to indicate the bandwidths in
use by specific flows using the fields of the MPLS shim header.
Furthermore, since the semantics of the MPLS shim header are fully
defined in [RFC3032] and [RFC5462], those fields cannot be assigned

semantics to carry this information. This document does not define
any semantic for use in the TTL or TC fields of the label entry that
carries the flow label, but requires that the flow label itself be
selected with a high degree of entropy suggesting that the label
value should not be overloaded with additional meaning in any
subsequent specification.

A different type of load balancing is the desire to carry a PW over a
set of PSN links in which the bandwidth of members of the link set is
less than the bandwidth of the PW. Proposals to address this problem
have been made in the past [PWBONDING]. Such a mechanism can be
considered complementary to this mechanism.

8.2. Link Aggregation Groups

A Link Aggregation Group (LAG) is used to bond together several
physical circuits between two adjacent nodes so they appear to
higher-layer protocols as a single, higher-bandwidth "virtual" pipe.
These may coexist in various parts of a given network. An advantage
of LAGs is that they reduce the number of routing and signalling
protocol adjacencies between devices, reducing control plane
processing overhead. As with ECMP, the key problem related to LAGs
is that due to inefficiencies in LAG load-distribution algorithms, a
particular component of a LAG may experience congestion. The
mechanism proposed here may be able to assist in producing a more
uniform flow distribution.

The same considerations requiring a flow to go over a single member
of an ECMP set apply to a member of a LAG.

8.3. Multiple RSVP-TE Paths

In some networks, it is desirable for a Label Edge Router (LER) to be
able to load-balance a PW across multiple Resource Reservation
Protocol - Traffic Engineering (RSVP-TE) tunnels. The flow label
mechanism described in this document may be used to provide the LER
with the required flow information and necessary entropy to provide
this type of load balancing. An example of such a case is the use of
the flow label mechanism in networks using a link bundle with the all
ones component [RFC4201].

Methods by which the LER is configured to apply this type of ECMP are
outside the scope of this document.

8.4. The Single Large Flow Case

   Clearly, the operator should make sure that the service offered using
   PW technology and the method described in this document do not exceed
   the maximum planned link capacity, unless it can be guaranteed that
   they conform to the Internet traffic profile of a very large number
   of small flows.

   If the NSP cannot access sufficient information to distinguish flows,
   perhaps because the protocol stack required parsing further into the
   packet than it is able, then the functionality described in this
   document does not give any benefits. The most common case where a
   single flow dominates the traffic on a PW is when it is used to
   transport enterprise traffic. Enterprise traffic may well consist of
   a single, large TCP flow, or encrypted flows that cannot be handled
   by the methods described in this document.

   An operator has four options under these circumstances:

   1. The operator can choose to do nothing, and the system will work
       as it does without the flow label.

   2. The operator can make the customer aware that the service
       offering has a restriction on flow bandwidth and police flows to
       that restriction. This would allow customers offering multiple
       flows to use a larger fraction of their access bandwidth, whilst
       preventing a single flow from consuming a fraction of internal
       link bandwidth that the operator considered excessive.

   3. The operator could configure the ingress PE to assign a constant
       flow label to all high-bandwidth flows so that only one path was
       affected by these flows.

   4. The operator could configure the ingress PE to assign a random
       flow label to all high-bandwidth flows so as to minimise the
       disruption to the network at the cost of out-of-order traffic to
       the user.

   The issues described above are mitigated by the following two
   factors:

   o Firstly, the customer of a high-bandwidth PW service has an
      incentive to get the best transport service, because an
      inefficient use of the PSN leads to jitter and eventually to loss
      to the PW's payload.

   o Secondly, the customer is usually able to tailor their
      applications to generate many flows in the PSN. A well-known
      example is massive data transport between servers that use many
      parallel TCP sessions. This same technique can be used by any
      transport protocol: multiple UDP ports, multiple L2TPv3 Session
      IDs, or multiple Generic Routing Encapsulation (GRE) keys may be
      used to decompose a large flow into smaller components. This
      approach may be applied to IPsec [RFC4301] where multiple Security
      Parameter Indexes (SPIs) may be allocated to the same security
      association.

8.5. Applicability to MPLS-TP

   The MPLS Transport Profile (MPLS-TP) [RFC5654] Requirement 44 states
   that "MPLS-TP MUST support mechanisms that ensure the integrity of
   the transported customer's service traffic as required by its
   associated Service Level Agreement (SLA). Loss of integrity may be
   defined as packet corruption, reordering, or loss during normal
   network conditions". In addition, MPLS-TP makes extensive use of the
   fate sharing between OAM and data packets, which is defeated by the
   flow LSE. The flow-aware transport of a PW reorders packets and
   therefore MUST NOT be deployed in a network conforming to MPLS-TP,
   unless these integrity requirements specified in the SLA can be
   satisfied.

8.6. Asymmetric Operation

   The protocol defined in this document supports the asymmetric
   inclusion of the flow LSE. Asymmetric operation can be expected when
   there is asymmetry in the bandwidth requirements making it
   unprofitable for one PE to perform the flow classification, or when
   that PE is otherwise unable to perform the classification but is able
   to receive flow labeled packets from its peer. Asymmetric operation
   of the PW may also be required when one PE has a high transmission
   bandwidth requirement, but has a need to receive the entire PW on a
   single interface in order to perform a processing operation that
   requires the context of the complete PW (for example, policing of the
   egress traffic).

9. Applicability to MPLS LSPs

   An extension of this technique is to create a basis for hash
   diversity without having to peek below the label stack for IP traffic
   carried over Label Distribution Protocol (LDP) LSPs. The
   generalisation of this extension to MPLS has been described in
   [MPLS-ENTROPY]. This generalisation can be regarded as a

complementary, but distinct, approach from the technique described in
this document. While similar consideration may apply to the
identification of flows and the allocation of flow label values, the
flow labels are imposed by different network components, and the
associated signalling mechanisms are different.

10. Security Considerations

The PW generic security considerations described in [RFC3985] and the
security considerations applicable to a specific PW type (for
example, in the case of an Ethernet PW [RFC4448]) apply. The
security considerations in [RFC5920] also apply.

Section 1.3 describes considerations that apply to the TTL value used
in the flow LSE. The use of a TTL value of one prevents the
accidental forwarding of a packet based on the label value in the
flow LSE.

11. IANA Considerations

IANA maintains the registry "Pseudowire Name Spaces (PWE3)" with
sub-registry "Pseudowire Interface Parameters Sub-TLV type Registry".
IANA has registered the Flow Label Sub-TLV type in this registry.

Parameter	ID Length	Description	Reference
0x17	4	Flow Label	RFC 6391

12. Congestion Considerations

The congestion considerations applicable to PWs as described in
[RFC3985] apply to this design.

The ability to explicitly configure a PW to leverage the availability
of multiple ECMPs is beneficial to capacity planning as, all other
parameters being constant, the statistical multiplexing of a larger
number of smaller flows is more efficient than with a smaller number
of larger flows.

Note that if the classification into flows is only performed on IP
packets, the behaviour of those flows in the face of congestion will
be as already defined by the IETF for packets of that type, and no
additional congestion processing is required.

Where flows that are not IP are classified, PW congestion avoidance
must be applied to each non-IP load balance group.

13. Acknowledgements

   The authors wish to thank Mary Barnes, Eric Grey, Kireeti Kompella,
   Joerg Kuechemann, Wilfried Maas, Luca Martini, Mark Townsley, Rolf
   Winter, and Lucy Yong for valuable comments on this document.

14. References

14.1. 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.

   [RFC4379]   Kompella, K. and G. Swallow, "Detecting Multi-Protocol
               Label Switched (MPLS) Data Plane Failures", RFC 4379,
               February 2006.

   [RFC4385]   Bryant, S., Swallow, G., Martini, L., and D. McPherson,
               "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word
               for Use over an MPLS PSN", RFC 4385, February 2006.

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

   [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.

   [RFC4553]   Vainshtein, A., Ed., and YJ. Stein, Ed., "Structure-
               Agnostic Time Division Multiplexing (TDM) over Packet
               (SAToP)", RFC 4553, June 2006.

   [RFC4928]   Swallow, G., Bryant, S., and L. Andersson, "Avoiding
               Equal Cost Multipath Treatment in MPLS Networks",
               BCP 128, RFC 4928, June 2007.

   [RFC5085]   Nadeau, T., Ed., and C. Pignataro, Ed., "Pseudowire
               Virtual Circuit Connectivity Verification (VCCV): A
               Control Channel for Pseudowires", RFC 5085,
               December 2007.

14.2. Informative References

   [MPLS-ENTROPY]
              Kompella, K., Drake, J., Amante, S., Henderickx, W., and
              L. Yong, "The Use of Entropy Labels in MPLS Forwarding",
              Work in Progress, October 2011.

   [PWBONDING] Stein, Y(J)., Mendelsohn, I., and R. Insler, "PW
              Bonding", Work in Progress, November 2008.

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

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

   [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling
              in MPLS Traffic Engineering (TE)", RFC 4201,
              October 2005.

   [RFC4301] Kent, S. and K. Seo, "Security Architecture for the
              Internet Protocol", RFC 4301, December 2005.

   [RFC4378] Allan, D., Ed., and T. Nadeau, Ed., "A Framework for
              Multi-Protocol Label Switching (MPLS) Operations and
              Management (OAM)", RFC 4378, February 2006.

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

   [RFC5654] Niven-Jenkins, B., Ed., Brungard, D., Ed., Betts, M.,
              Ed., Sprecher, N., and S. Ueno, "Requirements of an MPLS
              Transport Profile", RFC 5654, September 2009.

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

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

Authors' Addresses

   Stewart Bryant (editor)
   Cisco Systems
   250 Longwater Ave.
   Reading RG2 6GB
   United Kingdom

   Phone: +44-208-824-8828
   EMail: stbryant@cisco.com

   Clarence Filsfils
   Cisco Systems
   Brussels
   Belgium

   EMail: cfilsfil@cisco.com

   Ulrich Drafz
   Deutsche Telekom
   Muenster
   Germany

   EMail: Ulrich.Drafz@telekom.de

   Vach Kompella
   Alcatel-Lucent

   EMail: vach.kompella@alcatel-lucent.com

   Joe Regan
   Alcatel-Lucent

   EMail: joe.regan@alcatel-lucent.com

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

   EMail: shane@level3.net