                        Transport of Ethernet Frames over
                  Layer 2 Tunneling Protocol Version 3 (L2TPv3)

Status of This Memo

   This document specifies an Internet standards track protocol for the
   Internet community, and requests discussion and suggestions for
   improvements. Please refer to the current edition of the "Internet
   Official Protocol Standards" (STD 1) for the standardization state
   and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Abstract

   This document describes the transport of Ethernet frames over the
   Layer 2 Tunneling Protocol, Version 3 (L2TPv3). This includes the
   transport of Ethernet port-to-port frames as well as the transport of
   Ethernet VLAN frames. The mechanism described in this document can
   be used in the creation of Pseudowires to transport Ethernet frames
   over an IP network.

Table of Contents

1.   Introduction

   The Layer 2 Tunneling Protocol, Version 3 (L2TPv3) can be used as a
   control protocol and for data encapsulation to set up Pseudowires
   (PWs) for transporting layer 2 Packet Data Units across an IP network
   [RFC3931]. This document describes the transport of Ethernet frames
   over L2TPv3 including the PW establishment and data encapsulation.

   The term "Ethernet" in this document is used with the intention to
   include all such protocols that are reasonably similar in their
   packet format to IEEE 802.3 [802.3], including variants or extensions
   that may or may not necessarily be sanctioned by the IEEE (including
   such frames as jumbo frames, etc.). The term "VLAN" in this document
   is used with the intention to include all virtual LAN tagging
   protocols such as IEEE 802.1Q [802.1Q], 802.1ad [802.1ad], etc.

1.1.   Specification of Requirements

   In this document, several words are used to signify the requirements
   of the specification. These words are often capitalized. The key
   words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD",
   "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document
   are to be interpreted as described in [RFC2119].

1.2. Abbreviations

    AC   Attachment Circuit (see [RFC3985])
    CE   Customer Edge (Typically also the L2TPv3 Remote System)
    LCCE   L2TP Control Connection Endpoint (see [RFC3931])
    NSP   Native Service Processing (see [RFC3985])
    PE   Provider Edge (Typically also the LCCE) (see [RFC3985])
    PSN   Packet Switched Network (see [RFC3985])
    PW   Pseudowire (see [RFC3985])
    PWE3   Pseudowire Emulation Edge to Edge (Working Group)

1.3. L2TPv3 Control Message Types

   Relevant L2TPv3 control message types (see [RFC3931]) are listed for
   reference.

    SCCRQ   L2TPv3 Start-Control-Connection-Request control message
    SCCRP   L2TPv3 Start-Control-Connection-Reply control message
    SCCCN   L2TPv3 Start-Control-Connection-Connected control message
    StopCCN L2TPv3 Stop-Control-Connection-Notification control message
    ICRQ   L2TPv3 Incoming-Call-Request control message
    ICRP   L2TPv3 Incoming-Call-Reply control message
    ICCN   L2TPv3 Incoming-Call-Connected control message
    OCRQ   L2TPv3 Outgoing-Call-Request control message
    OCRP   L2TPv3 Outgoing-Call-Reply control message
    OCCN   L2TPv3 Outgoing-Call-Connected control message
    CDN   L2TPv3 Call-Disconnect-Notify control message
    SLI   L2TPv3 Set-Link-Info control message

1.4. Requirements

   An Ethernet PW emulates a single Ethernet link between exactly two
   endpoints. The following figure depicts the PW termination relative
   to the NSP and PSN tunnel within an LCCE [RFC3985]. The Ethernet
   interface may be connected to one or more Remote Systems (an L2TPv3
   Remote System is referred to as Customer Edge (CE) in this and
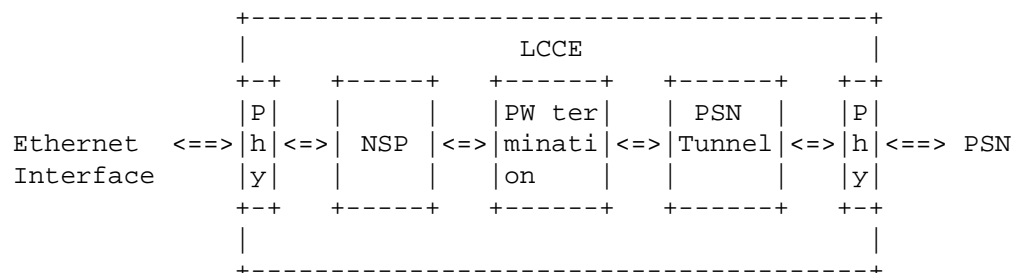   associated PWE3 documents). The LCCE may or may not be a PE.

```
              +---------------------------------------+
              |                   LCCE                 |
              +-+   +-----+   +------+   +------+   +-+
              |P|   |     |   |PW ter|   | PSN  |   |P|
   Ethernet  <==>|h|<=>|  NSP  |<=>|minati|<=>|Tunnel|<=>|h|<==> PSN
   Interface     |y|   |     |   |on    |   |      |   |y|
              +-+   +-----+   +------+   +------+   +-+
              |                                       |
              +---------------------------------------+
                     Figure 1: PW termination
```

The PW termination point receives untagged (also referred to as
'raw') or tagged Ethernet frames and delivers them unaltered to the
PW termination point on the remote LCCE. Hence, it can provide
untagged or tagged Ethernet link emulation service.

The "NSP" function includes packet processing needed to translate the
Ethernet frames that arrive at the CE-LCCE interface to/from the
Ethernet frames that are applied to the PW termination point. Such
functions may include stripping, overwriting, or adding VLAN tags.
The NSP functionality can be used in conjunction with local
provisioning to provide heterogeneous services where the CE-LCCE
encapsulations at the two ends may be different.

The physical layer between the CE and LCCE, and any adaptation (NSP)
functions between it and the PW termination, are outside of the scope
of PWE3 and are not defined here.

2. PW Establishment

With L2TPv3 as the tunneling protocol, Ethernet PWs are L2TPv3
sessions. An L2TP Control Connection has to be set up first between
the two LCCEs. Individual PWs can then be established as L2TP
sessions.

2.1. LCCE-LCCE Control Connection Establishment

The two LCCEs that wish to set up Ethernet PWs MUST establish an L2TP
Control Connection first as described in [RFC3931]. Hence, an
Ethernet PW Type must be included in the Pseudowire Capabilities List
as defined in [RFC3931]. The type of PW can be either "Ethernet
port" or "Ethernet VLAN". This indicates that the Control Connection
can support the establishment of Ethernet PWs. Note that there are
two Ethernet PW Types required. For connecting an Ethernet port to
another Ethernet port, the PW Type MUST be "Ethernet port"; for
connecting an Ethernet VLAN to another Ethernet VLAN, the PW Type
MUST be "Ethernet VLAN".

2.2. PW Session Establishment

The provisioning of an Ethernet port or Ethernet VLAN and its
association with a PW triggers the establishment of an L2TP session
via the standard Incoming Call three-way handshake described in
Section 3.4.1 of [RFC3931].

Note that an L2TP Outgoing Call is essentially a method of
controlling the originating point of a Switched Virtual Circuit
(SVC), allowing it to be established from any reachable L2TP-enabled
device able to perform outgoing calls. The Outgoing Call model and
its corresponding OCRQ, OCRP, and OCCN control messages are mainly
used within the dial arena with L2TPv2 today and has not been found
applicable for PW applications yet.

The following are the signaling elements needed for the Ethernet PW
establishment:

a) Pseudowire Type: The type of a Pseudowire can be either "Ethernet
   port" or "Ethernet VLAN". Each LCCE signals its Pseudowire type
   in the Pseudowire Type AVP [RFC3931]. The assigned values for
   "Ethernet port" and "Ethernet VLAN" Pseudowire types are captured
   in the "IANA Considerations" of this document. The Pseudowire
   Type AVP MUST be present in the ICRQ.

b) Pseudowire ID: Each PW is associated with a Pseudowire ID. The
   two LCCEs of a PW have the same Pseudowire ID for it. The Remote
   End Identifier AVP [RFC3931] is used to convey the Pseudowire ID.
   The Remote End Identifier AVP MUST be present in the ICRQ in order
   for the remote LCCE to determine the PW to associate the L2TP
   session with. An implementation MUST support a Remote End
   Identifier of four octets known to both LCCEs either by manual
   configuration or some other means. Additional Remote End
   Identifier formats that MAY be supported are outside the scope of
   this document.

c) The Circuit Status AVP [RFC3931] MUST be included in ICRQ and ICRP
   to indicate the circuit status of the Ethernet port or Ethernet
   VLAN. For ICRQ and ICRP, the Circuit Status AVP MUST indicate
   that the circuit status is for a new circuit (refer to N bit in
   Section 2.3.3). An implementation MAY send an ICRQ or ICRP before
   an Ethernet interface is ACTIVE, as long as the Circuit Status AVP
   (refer to A bit in Section 2.3.3) in the ICRQ or ICRP reflects the
   correct status of the Ethernet port or Ethernet VLAN link. A
   subsequent circuit status change of the Ethernet port or Ethernet
   VLAN MUST be conveyed in the Circuit Status AVP in ICCN or SLI
   control messages. For ICCN and SLI (refer to Section 2.3.2), the
   Circuit Status AVP MUST indicate that the circuit status is for an
   existing circuit (refer to N bit in Section 2.3.3) and reflect the
   current status of the link (refer to A bit in Section 2.3.3).

2.3. PW Session Monitoring

2.3.1. Control Connection Keep-alive

   The working status of a PW is reflected by the state of the L2TPv3
   session. If the corresponding L2TPv3 session is down, the PW
   associated with it MUST be shut down. The Control Connection keep-
   alive mechanism of L2TPv3 can serve as a link status monitoring
   mechanism for the set of PWs associated with a Control Connection.

2.3.2. SLI Message

   In addition to the Control Connection keep-alive mechanism of L2TPv3,
   Ethernet PW over L2TP makes use of the Set-Link-Info (SLI) control
   message defined in [RFC3931]. The SLI message is used to signal
   Ethernet link status notifications between LCCEs. This can be useful
   to indicate Ethernet interface state changes without bringing down
   the L2TP session. Note that change in the Ethernet interface state
   will trigger an SLI message for each PW associated with that Ethernet
   interface. This may be one Ethernet port PW or more than one
   Ethernet VLAN PW. The SLI message MUST be sent any time there is a
   status change of any values identified in the Circuit Status AVP.
   The only exception to this is the initial ICRQ, ICRP, and CDN
   messages that establish and tear down the L2TP session itself. The
   SLI message may be sent from either LCCE at any time after the first
   ICRQ is sent (and perhaps before an ICRP is received, requiring the
   peer to perform a reverse Session ID lookup).

2.3.3. Use of Circuit Status AVP for Ethernet

   Ethernet PW reports circuit status with the Circuit Status AVP
   defined in [RFC3931]. For reference, this AVP is shown below:

```
    0                   1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |             Reserved      |N|A|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   The Value is a 16-bit mask with the two least significant bits
   defined and the remaining bits reserved for future use. Reserved
   bits MUST be set to 0 when sending and ignored upon receipt.

   The A (Active) bit indicates whether the Ethernet interface is ACTIVE
   (1) or INACTIVE (0).

   The N (New) bit indicates whether the circuit status is for a new (1)
   Ethernet circuit or an existing (0) Ethernet circuit.

3. Packet Processing

3.1. Encapsulation

   The encapsulation described in this section refers to the
   functionality performed by the PW termination point depicted in
   Figure 1, unless otherwise indicated.

   The entire Ethernet frame, without the preamble or frame check
   sequence (FCS), is encapsulated in L2TPv3 and is sent as a single
   packet by the ingress LCCE. This is done regardless of whether or
   not a VLAN tag is present in the Ethernet frame. For Ethernet port-
   to-port mode, the remote LCCE simply decapsulates the L2TP payload
   and sends it out on the appropriate interface without modifying the
   Ethernet header. For Ethernet VLAN-to-VLAN mode, the remote LCCE MAY
   rewrite the VLAN tag. As described in Section 1, the VLAN tag
   modification is an NSP function.

   The Ethernet PW over L2TP is homogeneous with respect to packet
   encapsulation, i.e., both ends of the PW are either untagged or
   tagged. The Ethernet PW can still be used to provide heterogeneous
   services using NSP functionality at the ingress and/or egress LCCE.
   The definition of such NSP functionality is outside the scope of this
   document.

   The maximum length of the Ethernet frame carried as the PW payload is
   irrelevant as far as the PW is concerned. If anything, that value
   would only be relevant when quantifying the faithfulness of the
   emulation.

3.2. Sequencing

   Data packet sequencing MAY be enabled for Ethernet PWs. The
   sequencing mechanisms described in [RFC3931] MUST be used for
   signaling sequencing support.

3.3. MTU Handling

   With L2TPv3 as the tunneling protocol, the IP packet resulting from
   the encapsulation is M + N bytes longer than the Ethernet frame
   without the preamble or FCS. Here M is the length of the IP header
   along with associated options and extension headers, and the value of
   N depends on the following fields:

```
   L2TP Session Header:
      Flags, Ver, Res - 4 octets (L2TPv3 over UDP only)
      Session ID      - 4 octets
      Cookie Size     - 0, 4, or 8 octets
      L2-Specific Sublayer - 0 or 4 octets (i.e., using sequencing)

   Hence the range for N in octets is:
      N = 4-16,  for L2TPv3 data messages over IP;
      N = 16-28, for L2TPv3 data messages over UDP;
      (N does not include the IP header).
```

Fragmentation in the PSN can occur when using Ethernet over L2TP,
unless proper configuration and management of MTU sizes are in place
between the Customer Edge (CE) router and Provider Edge (PE) router,
and across the PSN. This is not specific only to Ethernet over
L2TPv3, and the base L2TPv3 specification [RFC3931] provides general
recommendations with respect to fragmentation and reassembly in
Section 4.1.4. "PWE3 Fragmentation and Reassembly" [RFC4623]
expounds on this topic, including a fragmentation and reassembly
mechanism within L2TP itself in the event that no other option is
available. Implementations MUST follow these guidelines with respect
to fragmentation and reassembly.

4. Applicability Statement

   The Ethernet PW emulation allows a service provider to offer a
   "port-to-port"-based Ethernet service across an IP Packet Switched
   Network (PSN), while the Ethernet VLAN PW emulation allows an "VLAN-
   to-VLAN"-based Ethernet service across an IP Packet Switched Network
   (PSN).

   The Ethernet or Ethernet VLAN PW emulation has the following
   characteristics in relationship to the respective native service:

   o Ethernet PW connects two Ethernet port ACs, and Ethernet VLAN PW
      connects two Ethernet VLAN ACs, which both support bi-directional
      transport of variable-length Ethernet frames. The ingress LCCE
      strips the preamble and FCS from the Ethernet frame and transports
      the frame in its entirety across the PW. This is done regardless
      of the presence of the VLAN tag in the frame. The egress LCCE
      receives the Ethernet frame from the PW and regenerates the
      preamble and FCS before forwarding the frame to the attached
      Remote System (see Section 3.1). Since FCS is not being
      transported across either Ethernet or Ethernet VLAN PWs, payload
      integrity transparency may be lost. To achieve payload integrity
      transparency on Ethernet or Ethernet VLAN PWs using L2TP over IP
      or L2TP over UDP/IP, the L2TPv3 session can utilize IPsec as
      specified in Section 4.1.3 of [RFC3931].

o While architecturally [RFC3985] outside the scope of the L2TPv3 PW
   itself, if VLAN tags are present, the NSP may rewrite VLAN tags on
   ingress or egress from the PW (see Section 3.1).

o The Ethernet or Ethernet VLAN PW only supports homogeneous
   Ethernet frame type across the PW; both ends of the PW must be
   either tagged or untagged. Heterogeneous frame type support
   achieved with NSP functionality is outside the scope of this
   document (see Section 3.1).

o Ethernet port or Ethernet VLAN status notification is provided
   using the Circuit Status AVP in the SLI message (see Sections
   2.3.2 and 2.3.3). Loss of connectivity between LCCEs can be
   detected by the L2TPv3 keep-alive mechanism (see Section 2.3.1 of
   this document and Section 4.4 of [RFC3931]). The LCCE can convey
   these indications back to its attached Remote System.

o The maximum frame size that can be supported is limited by the PSN
   MTU minus the L2TPv3 header size, unless fragmentation and
   reassembly is used (see Section 3.3 of this document and Section
   4.1.4 of [RFC3931]).

o The Packet Switched Network may reorder, duplicate, or silently
   drop packets. Sequencing may be enabled in the Ethernet or
   Ethernet VLAN PW for some or all packets to detect lost,
   duplicate, or out-of-order packets on a per-session basis (see
   Section 3.2).

o The faithfulness of an Ethernet or Ethernet VLAN PW may be
   increased by leveraging Quality-of-Service (QoS) features of the
   LCCEs and the underlying PSN. For example, for Ethernet 802.1Q
   [802.1Q] VLAN transport, the ingress LCCE MAY consider the user
   priority field (i.e., 802.1p) of the VLAN tag for traffic
   classification and QoS treatments, such as determining the
   Differentiated Services (DS) field [RFC2474] of the encapsulating
   IP header. Similarly, the egress LCCE MAY consider the DS field
   of the encapsulating IP header when rewriting the user priority
   field of the VLAN tag or queuing the Ethernet frame before
   forwarding the frame to the Remote System. The mapping between
   the user priority field and the IP header DS field as well as the
   Quality-of-Service model deployed are application specific and are
   outside the scope of this document.

5. Congestion Control

   As explained in [RFC3985], the PSN carrying the PW may be subject to
   congestion, with congestion characteristics depending on PSN type,
   network architecture, configuration, and loading. During congestion,
   the PSN may exhibit packet loss that will impact the service carried
   by the Ethernet or Ethernet VLAN PW. In addition, since Ethernet or
   Ethernet VLAN PWs carry a variety of services across the PSN,
   including but not restricted to TCP/IP, they may or may not behave in
   a TCP-friendly manner prescribed by [RFC2914] and thus consume more
   than their fair share.

   Whenever possible, Ethernet or Ethernet VLAN PWs should be run over
   traffic-engineered PSNs providing bandwidth allocation and admission
   control mechanisms. IntServ-enabled domains providing the Guaranteed
   Service (GS) or DiffServ-enabled domains using EF (expedited
   forwarding) are examples of traffic-engineered PSNs. Such PSNs will
   minimize loss and delay while providing some degree of isolation of
   the Ethernet or Ethernet VLAN PW's effects from neighboring streams.

   LCCEs SHOULD monitor for congestion (by using explicit congestion
   notification or by measuring packet loss) in order to ensure that the
   service using the Ethernet or Ethernet VLAN PW may be maintained.
   When severe congestion is detected (for example, when enabling
   sequencing and detecting that the packet loss is higher than a
   threshold), the Ethernet or Ethernet VLAN PW SHOULD be halted by
   tearing down the L2TP session via a CDN message. The PW may be
   restarted by manual intervention or by automatic means after an
   appropriate waiting time. Note that the thresholds and time periods
   for shutdown and possible automatic recovery need to be carefully
   configured. This is necessary to avoid loss of service due to
   temporary congestion and to prevent oscillation between the congested
   and halted states.

   This specification offers no congestion control and is not TCP
   friendly [TFRC]. Future works for PW congestion control (being
   studied by the PWE3 Working Group) will provide congestion control
   for all PW types including Ethernet and Ethernet VLAN PWs.

6. Security Considerations

   Ethernet over L2TPv3 is subject to all of the general security
   considerations outlined in [RFC3931].

7. IANA Considerations

   The signaling mechanisms defined in this document rely upon the
   following Ethernet Pseudowire Types (see Pseudowire Capabilities List
   as defined in 5.4.3 of [RFC3931] and L2TPv3 Pseudowire Types in 10.6
   of [RFC3931]), which were allocated by the IANA (number space created
   as part of publication of [RFC3931]):

      Pseudowire Types

      0x0004 Ethernet VLAN Pseudowire Type
      0x0005 Ethernet Pseudowire Type

8. Contributors

   The following is the complete list of contributors to this document.

   Rahul Aggarwal
   Juniper Networks

   Xipeng Xiao
   Riverstone Networks

   W. Mark Townsley
   Stewart Bryant
   Maria Alice Dos Santos
   Cisco Systems

   Cheng-Yin Lee
   Alcatel

   Tissa Senevirathne
   Consultant

   Mitsuru Higashiyama
   Anritsu Corporation

9. Acknowledgements

   This RFC evolved from the document, "Ethernet Pseudo Wire Emulation
   Edge-to-Edge". We would like to thank its authors, T.So, X.Xiao, L.
   Anderson, C. Flores, N. Tingle, S. Khandekar, D. Zelig and G. Heron
   for their contribution. We would also like to thank S. Nanji, the
   author of "Ethernet Service for Layer Two Tunneling Protocol", for
   writing the first Ethernet over L2TP document.

   Thanks to Carlos Pignataro for providing a thorough review and
   helpful input.

10. References

10.1. Normative References

   [RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling
              Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.

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

   [RFC4623] Malis, A. and M. Townsley, "Pseudowire Emulation Edge-to-
              Edge (PWE3) Fragmentation and Reassembly", RFC 4623,
              August 2006.

10.2. Informative References

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

   [RFC2914] Floyd, S., "Congestion Control Principles", BCP 41, RFC
              2914, September 2000.

   [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black,
              "Definition of the Differentiated Services Field (DS
              Field) in the IPv4 and IPv6 Headers", RFC 2474, December
              1998.

   [802.3]   IEEE, "IEEE std 802.3 -2005/Cor 1-2006 IEEE Standard for
              Information Technology - Telecommuincations and
              Information Exchange Between Systems - Local and
              Metropolitan Area Networks", IEEE Std 802.3-2005/Cor
              1-2006 (Corrigendum to IEEE Std 802.3-2005)

   [802.1Q]   IEEE, "IEEE standard for local and metropolitan area
              networks virtual bridged local area networks", IEEE Std
              802.1Q-2005 (Incorporates IEEE Std 802.1Q1998, IEEE Std
              802.1u-2001, IEEE Std 802.1v-2001, and IEEE Std 802.1s-
              2002)

   [802.1ad] IEEE, "IEEE Std 802.1ad - 2005 IEEE Standard for Local and
              metropolitan area networks - virtual Bridged Local Area
              Networks, Amendment 4: Provider Bridges", IEEE Std
              802.1ad-2005 (Amendment to IEEE Std 8021Q-2005)

   [TFRC]   Handley, M., Floyd, S., Padhye, J., and J. Widmer, "TCP
              Friendly Rate Control (TFRC): Protocol Specification", RFC
              3448, January 2003.

Author Information

    Rahul Aggarwal
    Juniper Networks
    1194 North Mathilda Avenue
    Sunnyvale, CA 94089

    EMail: rahul@juniper.net

    W. Mark Townsley
    Cisco Systems
    7025 Kit Creek Road
    PO Box 14987
    Research Triangle Park, NC 27709

    EMail: mark@townsley.net

    Maria Alice Dos Santos
    Cisco Systems
    170 W Tasman Dr
    San Jose, CA 95134

    EMail: mariados@cisco.com

Full Copyright Statement

Intellectual Property

Acknowledgement