Network Working Group — Luca Martini
Internet Draft — Nasser El-Aawar
Expiration Date: January 2002 — Level 3 Communications, LLC.

Steve Vogelsang
John Shirron
Toby Smith
Laurel Networks, Inc.

Daniel Tappan
Eric C. Rosen
Alex Hamilton
Jayakumar Jayakumar
Cisco Systems, Inc.

Vasile Radoaca
Nortel Networks

Dimitri Stratton Vlachos
Mazu Networks, Inc.

Andrew G. Malis
Vinai Sirkay
Vivace Networks, Inc.

Chris Liljenstolpe
Cable & Wireless

Giles Heron
Gone2 Ltd.

Dave Cooper
Global Crossing

Kireeti Kompella
Juniper Networks

July 2001

Transport of Layer 2 Frames Over MPLS

draft-martini-l2circuit-trans-mpls-07.txt

Status of this Memo

Abstract

This document describes methods for transporting the Protocol Data Units (PDUs) of layer 2 protocols such as Frame Relay, ATM AAL5, Ethernet, and providing a SONET circuit emulation service across an MPLS network.

Table of Contents

1. 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.

2. Introduction

In an MPLS network, it is possible to carry the Protocol Data Units (PDUs) of layer 2 protocols by prepending an MPLS label stack to these PDUs. This document specifies the necessary label distribution procedures for accomplishing this using the encapsulation methods in [7]. We restrict discussion to the case of point-to-point transport. QoS related issues are not discussed in this draft. This document describes methods for transporting a number of protocols; in some cases, transporting a particular protocol may have several modes of operation. Each of these protocols and/or modes may be implemented independently.

An accompanying document [8] also describes a method for transporting time division multiplexed (TDM) digital signals (TDM circuit emulation) over a packet-oriented MPLS network. The transmission system for circuit-oriented TDM signals is the Synchronous Optical Network (SONET)[5]/Synchronous Digital Hierarchy (SDH) [6]. To support TDM traffic, which includes voice, data, and private leased line service, the MPLS network must emulate the circuit characteristics of SONET/SDH payloads. MPLS labels and a new circuit emulation header are used to encapsulate TDM signals and provide the Circuit Emulation Service over MPLS (CEM).

3. Tunnel Labels and VC Labels

Suppose it is desired to transport layer 2 PDUs from ingress LSR R1 to egress LSR R2, across an intervening MPLS network. We assume that there is an LSP from R1 to R2. That is, we assume that R1 can cause a packet to be delivered to R2 by pushing some label onto the packet and sending the result to one of its adjacencies. Call this label the "tunnel label", and the corresponding LSP the "tunnel LSP".

The tunnel LSP merely gets packets from R1 to R2, the corresponding label doesn't tell R2 what to do with the payload, and in fact if penultimate hop popping is used, R2 may never even see the corresponding label. (If R1 itself is the penultimate hop, a tunnel label may not even get pushed on.) Thus if the payload is not an IP packet, there must be a label, which becomes visible to R2, that tells R2 how to treat the received packet. Call this label the "VC label".

So when R1 sends a layer 2 PDU to R2, it first pushes a VC label on its label stack, and then (if R1 is not adjacent to R2) pushes on a tunnel label. The tunnel label gets the MPLS packet from R1 to R2; the VC label is not visible until the MPLS packet reaches R2. R2's

disposition of the packet is based on the VC label.

Note that the tunnel could be a GRE encapsulated MPLS tunnel between R1 and R2. In this case R1 would be adjacent to R2, and only the VC label would be used, and the intervening network need only carry IP packets.

If the payload of the MPLS packet is, for example, an ATM AAL5 PDU, the VC label will generally correspond to a particular ATM VC at R2. That is, R2 needs to be able to infer from the VC label the outgoing interface and the VPI/VCI value for the AAL5 PDU. If the payload is a Frame Relay PDU, then R2 needs to be able to infer from the VC label the outgoing interface and the DLCI value. If the payload is an Ethernet frame, then R2 needs to be able to infer from the VC label the outgoing interface, and perhaps the VLAN identifier. This process is unidirectional, and will be repeated independently for bidirectional operation. It is REQUIRED to assign the same VC ID, and VC type for a given circuit in both directions. The group id MUST NOT be required to match in both directions. The transported frame MAY be modified when it reaches the egress router. If the header of the transported layer 2 frame is modified, this MUST be done at the egress LSR only.

Note that the VC label must always be at the bottom of the label stack, and the tunnel label, if present, must be immediately above the VC label. Of course, as the packet is transported across the MPLS network, additional labels may be pushed on (and then popped off) as needed. Even R1 itself may push on additional labels above the tunnel label. If R1 and R2 are directly adjacent LSRs, then it may not be necessary to use a tunnel label at all.

This document does not specify a method for distributing the tunnel label or any other labels that may appear above the VC label on the stack. Any acceptable method of MPLS label distribution will do.

This document does specify a method for assigning and distributing the VC label. Static label assignment MAY be used, and implementations SHOULD provide support for this. When signaling is used, the VC label MUST be distributed from R2 to R1 using LDP in the downstream unsolicited mode; this requires that an LDP session be created between R1 and R2. It should be noted that this LDP session is not necessarily transported along the same path as the Layer2 PDUs. [1] In addition, when using LDP to distribute the VC label, liberal label retention mode SHOULD be used. However, as required in [1], the label request operation (mainly used by conservative label retention mode) MUST be implemented. VC labels MUST be allocated from the per-platform label space.

Note that this technique allows an unbounded number of layer 2 "VCs" to be carried together in a single "tunnel". Thus it scales quite well in the network backbone.

While this document currently defines the emulation of Frame Relay and ATM PVC services, it specifically does not preclude future enhancements to support switched service (SVC and SPVC) emulation.

4. Protocol-Specific Details

4.1. Frame Relay

The Frame Relay PDUs are encapsulated according to the procedures defined in [7]. The MPLS edge LSR MUST provide Frame Relay PVC status signaling to the Frame Relay network. If the MPLS edge LSR detects a service affecting condition as defined in [2] Q.933 Annex A.5 sited in IA FRF1.1, it MUST withdraw the label that corresponds to the frame relay DLCI. The Egress LSR SHOULD generate the corresponding errors and alarms as defined in [2] on the Frame relay VC.

4.2. ATM

4.2.1. ATM AAL5 VCC Transport

ATM AAL5 CSPS-PDUs are encapsulated according to [7] ATM AAL5 CPCS-PDU mode. This mode allows the transport of ATM AAL5 CSPS-PDUs traveling on a particular ATM PVC across the MPLS network to another ATM PVC.

4.2.2. ATM Transparent Cell Transport

This mode is similar to the Ethernet port mode. Every cell that is received at the ingress ATM port on the ingress LSR, R1, is encapsulated according to [7], ATM cell mode, and sent across the LSP to the egress LSR, R2. This mode allows an ATM port to be connected to only one other ATM port. [7] allows for grouping of multiple cells into a single MPLS frame. Grouping of ATM cells is OPTIONAL for transmission at the ingress LSR, R1. If the Egress LSR R2 supports cell concatenation the ingress LSR, R1, should only concatenate cells up to the "Maximum Number of concatenated ATM cells" parameter received as part of the FEC element.

4.2.3. ATM VCC and VPC Cell Transport

This mode is similar to the ATM AAL5 VCC transport except that only cells are transported. Every cell that is received on a pre-defined ATM PVC, or ATM PVP, at the ingress ATM port on the ingress LSR, R1, is encapsulated according to [7], ATM cell mode, and sent across the LSP to the egress LSR R2. Grouping of ATM cells is OPTIONAL for transmission at the ingress LSR, R1. If the Egress LSR R2 supports cell concatenation the ingress LSR, R1, MUST only concatenate cells up to the "Maximum Number of concatenated ATM cells in a frame" parameter received as part of the FEC element.

4.2.4. OAM Cell Support

OAM cells MAY be transported on the VC LSP. When the LSR is operating in AAL5 CPCS-PDU transport mode if it does not support transport of ATM cells, the LSR MUST discard incoming MPLS frames on an ATM VC LSP that contain a VC label with the T bit set [7]. When operating in AAL5 PDU transport mode an LSR that supports transport of OAM cells using the T bit defined in [7], or an LSR operating in any of the three cell transport modes MUST follow the procedures outlined in [9] section 8 for mode 0 only, in addition to the applicable procedures specified in [6].

4.2.4.1. OAM Cell Emulation Mode

AN LSR that does not support transport of OAM cells across an LSP MAY provide OAM support on ATM PVCs using the following procedures:

A pair of LSRs may emulate a bidrectional ATM VC by two uni-directioal LSPs. If an F5 end-to-end OAM cell is received from a ATM VC, by either LSR that is transporting this ATM VC, with a loopback indication value of 1, and the LSR has a label mapping for the ATM VC, then the LSR MUST decrement the loopback indication value and loop back the cell on the ATM VC. Otherwise the loopback cell MUST be discarded by the LSR.

The ingress LSR, R1, may also optionally be configured to periodically generate F5 end-to-end loopback OAM cells on a VC. If the LSR fails to receive a response to an F5 end-to-end loopback OAM cell for a pre-defined period of time it MUST withdraw the label mapping for the VC.

If an ingress LSR, R1, receives an AIS F5 OAM cell, fails to receive a pre-defined number of the End-to-End loop OAM cells, or a physical interface goes down, it MUST withdraw the label mappings for all VCs

associated with the failure. When a VC label mapping is withdrawn, the egress LSR, R2, MUST generate AIS F5 OAM cells on the VC associated with the withdrawn label mapping. In this mode it is very useful to apply a unique group ID to each interface. In the case where a physical interface goes down, a wild card label withdraw can be sent to all LDP neighbors, greatly reducing the signaling response time.

4.2.5. ILMI Support

An MPLS edge LSR MAY provide an ATM ILMI to the ATM edge switch. If an ingress LSR receives an ILMI message indicating that the ATM edge switch has deleted a VC, or if the physical interface goes down, it MUST withdraw the label mappings for all VCs associated with the failure. When a VC label mapping is withdrawn, the egress LSR SHOULD notify its client of this failure by deleting the VC using ILMI.

4.3. Ethernet VLAN

The Ethernet frame will be encapsulated according to the procedures in [7]. It should be noted that if the VLAN identifier is modified by the egress LSR, according to the procedures outlined above, the Ethernet spanning tree protocol might fail to work properly. If the LSR detects a failure on the Ethernet physical port, or the port is administratively disabled, it MUST withdraw the label mappings for all VCs associated with the port.

4.4. Ethernet

The Ethernet frame will be encapsulated according to the procedures in [7]. If the LSR detects a failure on the Ethernet physical port, or the port is administratively disabled, the corresponding VC label mapping MUST be withdrawn.

4.5. HDLC

HDLC frames are encapsulated according to the procedures in [7]. If the MPLS edge LSR detects that the physical link has failed, or the port is adminstratively disabled, it MUST withdraw the label mapping that corresponds to the HDLC link.

4.6. PPP

PPP frames are encapsulated according to the procedures in [7]. If the MPLS edge LSR detects that the physical link has failed, or the port is adminstratively disabled, it MUST withdraw the label mapping that corresponds to the PPP link.

5. LDP

The VC label bindings are distributed using the LDP downstream unsolicited mode described in [1]. The LSRs will establish an LDP session using the Extended Discovery mechanism described in [1, section 2.4.2 and 2.5], for this purpose a new type of FEC element is defined. The FEC element type is 128. [note1] Note that if the tunnel label is not available, the VC label SHOULD NOT be advertised.

The Virtual Circuit FEC element, is defined as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    VC tlv     |C|         VC Type             |VC info Length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Group ID                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        VC ID                                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Interface parameters                    |
|                              "                                |
|                              "                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- VC Type

  A 15 bit quantity containing a value which represents the type of VC. Assigned Values are:

VC Type	Description
0x0001	Frame Relay DLCI
0x0002	ATM AAL5 VCC transport
0x0003	ATM transparent cell transport
0x0004	Ethernet VLAN
0x0005	Ethernet
0x0006	HDLC
0x0007	PPP

0x8008	CEM [8]
0x0009	ATM VCC cell transport
0x000A	ATM VPC cell transport

- Control word bit (C)

  The highest order bit (C) of the Vc type is used to flag the presence of a control word (defined in [7]) as follows:
  bit 15 = 1 control word present on this VC.
  bit 15 = 0 no control word present on this VC.

- VC information length

  Length of the VC ID field and the interface parameters field in octets. If this value is 0, then it references all VCs using the specified group ID and there is no VC ID present, nor any interface parameters.

- Group ID

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

- VC ID

  A non zero 32-bit connection ID that together with the VC type, identifies a particular VC.

- Interface parameters

  This variable length field is used to provide interface specific parameters, such as interface MTU.

5.1. Interface Parameters Field

This field specifies edge facing interface specific parameters and SHOULD be used to validate that the LSRs, and the ingress and egress ports at the edges of the circuit, have the necessary capabilities to interoperate with each other. The field structure is defined as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Parameter ID |    Length     |    Variable Length Value      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Variable Length Value                 |
|                             "                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The parameter ID is defined as follows:

Parameter	ID Length	Description
0x01	4	Interface MTU in octets.
0x02	4	Maximum Number of concatenated ATM cells.
0x03	up to 82	Optional Interface Description string.
0x04	4	CEM [8] Payload Bytes.
0x05	4	CEM options.

The Length field is defined as the length of the interface parameter including the parameter id and length field itself.

- Interface MTU

  A 2 octet value indicating the MTU in octets. This is the Maximum Transmission Unit, excluding encapsulation overhead, of the egress packet interface that will be transmitting the decapsulated PDU that is received from the MPLS network. This parameter is applicable only to VC types 1, 2, 4, 5, 6, and 7, and is REQUIRED for these VC types. If this parameter does not match in both directions of a specific VC, that VC MUST NOT be enabled.

- Maximum Number of concatenated ATM cells

  A 2 octet value specifying the maximum number of concatenated ATM cells that can be processed as a single PDU by the egress LSR. An ingress LSR transmitting concatenated cells on this VC can concatenate a number of cells up to the value of this parameter, but MUST NOT exceed it. This parameter is applicable only to VC types 3, 9, and 0x0a, and is REQUIRED for these VC types. This parameter does not need to match in both directions of a specific VC.

- Optional Interface Description string

  This arbitrary, OPTIONAL, interface description string can be used to send an administrative description text string to the

remote LSR. This parameter is OPTIONAL, and is applicable to all VC types. The interface description parameter length is variable, and can be up to 80 octets.

- Payload Bytes

  A 2 octet value indicating the the number of TDM payload octets contained in all packets on the CEM stream, from 48 to 1,023 octets. All of the packets in a given CEM stream have the same number of payload bytes. Note that there is a possibility that the packet size may exceed the SPE size in the case of an STS-1 SPE, which could cause two pointers to be needed in the CEM header, since the payload may contain two J1 bytes for consecutive SPEs. For this reason, the number of payload bytes must be less than or equal to 783 for STS-1 SPEs.

- CEM Options. An optional 16 Bit value of CEM Flags. See [8] for the definition of the bit values.

5.2. LDP label Withdrawal procedures

As mentioned above the Group ID field can be used to withdraw all VC labels associated with a particular group ID. This procedure is OPTIONAL, and if it is implemented the LDP label withdraw message should be as follows: the VC information length field is set to 0, the VC ID field is not present, and the interface paramenters field is not present. For the purpose of this document this is called the "wild card withdraw procedure", and all LSRs implementing this design are REQUIRED to accept such a withdraw message, but are not required to send it.

The interface parameters field MUST NOT be present in any LDP VC label withdrawal message or release message. A wildcard release message MUST include only the group ID.

6. IANA Considerations

As specified in this document, a Virtual Circuit FEC element contains the VC Type field. VC Type value 0 is reserved. VC Type values 1 through 10 are defined in this document. VC Type values 11 through 63 are to be assigned by IANA using the "IETF Consensus" policy defined in RFC2434. VC Type values 64 through 127 are to be assigned by IANA, using the "First Come First Served" policy defined in RFC2434. VC Type values 128 through 32767 are vendor-specific, and values in this range are not to be assigned by IANA.

As specified in this document, a Virtual Circuit FEC element contains the Interface Parameters field, which is a list of one or more parameters, and each parameter is identified by the Parameter ID field. Parameter ID value 0 is reserved. Parameter ID values 1 through 6 are defined in this document. Parameter ID values 7 through 63 are to be assigned by IANA using the "IETF Consensus" policy defined in RFC2434. Parameter ID values 64 through 127 are to be assigned by IANA, using the "First Come First Served" policy defined in RFC2434. Parameter ID values 128 through 255 are vendor-specific, and values in this range are not to be assigned by IANA.

7. Security Considerations

This document does not affect the underlying security issues of MPLS.

8. References

[1] "LDP Specification." L. Andersson, P. Doolan, N. Feldman, A. Fredette, B. Thomas. January 2001. RFC3036

[2] ITU-T Recommendation Q.933, and Q.922 Specification for Frame Mode Basic call control, ITU Geneva 1995

[3] "MPLS Label Stack Encoding", E. Rosen, Y. Rekhter, D. Tappan, G. Fedorkow, D. Farinacci, T. Li, A. Conta. RFC3032

[4] "IEEE 802.3ac-1998" IEEE standard specification.

[5] American National Standards Institute, "Synchronous Optical Network Formats," ANSI T1.105-1995.

[6] ITU Recommendation G.707, "Network Node Interface For The Synchronous Digital Hierarchy", 1996.

[7] "Encapsulation Methods for Transport of Layer 2 Frames Over MPLS", draft-martini-l2circuit-encap-mpls-02.txt (Work in progress)

[8] "SONET/SDH Circuit Emulation Service Over MPLS (CEM) Encapsulation", draft-malis-sonet-ces-mpls-04.txt (Work in progress)

[9] "Frame Based ATM over SONET/SDH Transport (FAST)," 2000.

[note1] FEC element type 128 is pending IANA approval.

9. Author Information

Luca Martini
Level 3 Communications, LLC.
1025 Eldorado Blvd.
Broomfield, CO, 80021
e-mail: luca@level3.net

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

Giles Heron
Gone2 Ltd.
c/o MDP
One Curzon Street
London
W1J 5HD
United Kingdom
e-mail: giles@goneto.net

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

Dan Tappan
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824
e-mail: tappan@cisco.com

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

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

Eric Rosen
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824
e-mail: erosen@cisco.com

Steve Vogelsang
Laurel Networks, Inc.
Omega Corporate Center
1300 Omega Drive
Pittsburg, PA 15205
e-mail: sjv@laurelnetworks.com

John Shirron
Omega Corporate Center
1300 Omega Drive
Pittsburg, PA 15205
Laurel Networks, Inc.
e-mail: jshirron@laurelnetworks.com

Andrew G. Malis
Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134
Phone: +1 408 383 7223
Email: Andy.Malis@vivacenetworks.com

Vinai Sirkay
Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134
e-mail: vinai.sirkay@vivacenetworks.com

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

Chris Liljenstolpe
Cable & Wireless
11700 Plaza America Drive
Reston, VA 20190
e-mail: chris@cw.net

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

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