Network Working Group
Internet Draft
Expiration Date: October 2002

Luca Martini
Nasser El-Aawar
Level 3 Communications, LLC.

Giles Heron
PacketExchange Ltd.

Steve Vogelsang
Laurel Networks, Inc.

Chris Liljenstolpe
Cable & Wireless

Andrew G. Malis
Vinai Sirkay
Vivace Networks, Inc.

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

Kireeti Kompella
Juniper Networks

Dimitri Stratton Vlachos
Mazu Networks, Inc.

April 2002

Encapsulation Methods for Transport of ATM Cells/Frame Over IP and MPLS Networks

draft-martini-atm-encap-mpls-00.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

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

Abstract

This document describes methods for encapsulating the Protocol Data Units (PDUs) of layer 2 protocols such as ATM fortransport across an MPLS or IP 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 or IP network, it is possible to use control protocols such as those specified in [1] to set up "emulated virtual circuits" that carry the the Protocol Data Units of layer 2 protocols across the network. A number of these emulated virtual circuits may be carried in a single tunnel. This requires of course that the layer 2 PDUs be encapsulated. We can distinguish three layers of this encapsulation:

- the "tunnel header", which contains the information needed to transport the PDU across the IP or MPLS network; this is header belongs to the tunneling protocol, e.g., MPLS, GRE, L2TP.

- the "demultiplexer field", which is used to distinguish individual emulated virtual circuits within a single tunnel; this field must be understood by the tunneling protocol as well; it may be, e.g., an MPLS label or a GRE key field.

- the "emulated VC encapsulation", which contains the information about the enclosed layer 2 PDU which is necessary in order to properly emulate the corresponding layer 2 protocol.

This document specifies the emulated VC encapsulation for ATM cells and ATM AAL5 SDUs. Although different layer 2 protocols require different information to be carried in this encapsulation, an attempt has been made to make the encapsulation as common as possible for all layer 2 protocols. Other layer 2 protocols are described in separate documents. [5] [6] [7]

This document also specifies the way in which the demultiplexer field is added to the emulated VC encapsulation when an MPLS label is used as the demultiplexer field.

QoS related issues are not discussed in this draft

For the purpose of this document R1 will be defined as the ingress router, and R2 as the egress router. A layer 2 PDU will be received at R1, encapsulated at R1, transported, decapsulated at R2, and transmitted out of R2.

3. General encapsulation method

In most cases, it is not necessary to transport the layer 2 encapsulation across the network; rather, the layer 2 header can be stripped at R1, and reproduced at R2. This is done using information carried in the control word (see below), as well as information that may already have been signaled from R1 to R2.

3.1. The Control Word

There are three requirements that may need to be satisfied when transporting layer 2 protocols over an IP or MPLS backbone:

- -i. Sequentiality may need to be preserved.
- -ii. Small packets may need to be padded in order to be transmitted on a medium where the minimum transport unit is larger than the actual packet size.
- -iii. Control bits carried in the header of the layer 2 frame may need to be transported.

The control word defined here addresses all three of these requirements. For some protocols this word is REQUIRED, and for others OPTIONAL. For protocols where the control word is OPTIONAL implementations MUST support sending no control word, and MAY support sending a control word.

In all cases the egress router must be aware of whether the ingress router will send a control word over a specific virtual circuit.

This may be achieved by configuration of the routers, or by signaling, for example as defined in [1].

The control word 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Rsvd  | Flags |0 0|   Length  |     Sequence Number           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

In the above diagram the first 4 bits are reserved for future use. They MUST be set to 0 when transmitting, and MUST be ignored upon receipt.

The next 4 bits provide space for carrying protocol specific flags. These are defined in the protocol-specific details below.

The next 2 bits MUST be set to 0 when transmitting.

The next 6 bits provide a length field, which is used as follows: If the packet's length (defined as the length of the layer 2 payload plus the length of the control word) is less than 64 bytes, the length field MUST be set to the packet's length. Otherwise the length field MUST be set to zero. The value of the length field, if non-zero, can be used to remove any padding. When the packet reaches the service provider's egress router, it may be desirable to remove the padding before forwarding the packet.

The next 16 bits provide a sequence number that can be used to guarantee ordered packet delivery. The processing of the sequence number field is OPTIONAL.

The sequence number space is a 16 bit, unsigned circular space. The sequence number value 0 is used to indicate an unsequenced packet.

3.1.1. Setting the sequence number

For a given emulated VC, and a pair of routers R1 and R2, if R1 supports packet sequencing then the following procedures should be used:

- the initial packet transmitted on the emulated VC MUST use sequence number 1
- subsequent packets MUST increment the sequence number by one for each packet
- when the transmit sequence number reaches the maximum 16 bit value (65535) the sequence number MUST wrap to 1

If the transmitting router R1 does not support sequence number processing, then the sequence number field in the control word MUST be set to 0.

3.1.2. Processing the sequence number

If a router R2 supports receive sequence number processing, then the following procedures should be used:

When an emulated VC is initially set up, the "expected sequence number" associated with it MUST be initialized to 1.

When a packet is received on that emulated VC, the sequence number should be processed as follows:

- if the sequence number on the packet is 0, then the packet passes the sequence number check

- otherwise if the packet sequence number >= the expected sequence number and the packet sequence number - the expected sequence number < 32768, then the packet is in order.

- otherwise if the packet sequence number < the expected sequence number and the expected sequence number - the packet sequence number >= 32768, then the packet is in order.

- otherwise the packet is out of order.

If a packet passes the sequence number check, or is in order then, it can be delivered immediately. If the packet is in order, then the

expected sequence number should be set using the algorithm:

```
expected_sequence_number := packet_sequence_number + 1 mod 2**16
if (expected_sequence_number = 0) then expected_sequence_number := 1;
```

Packets which are received out of order MAY be dropped or reordered at the discretion of the receiver.

If a router R2 does not support receive sequence number processing, then the sequence number field MAY be ignored.

3.2. MTU Requirements

The network MUST be configured with an MTU that is sufficient to transport the largest encapsulation frames. If MPLS is used as the tunneling protocol, for example, this is likely to be 12 or more bytes greater than the largest frame size. Other tunneling protocols may have longer headers and require larger MTUs. If the ingress router determines that an encapsulated layer 2 PDU exceeds the MTU of the tunnel through which it must be sent, the PDU MUST be dropped. If an egress router receives an encapsulated layer 2 PDU whose payload length (i.e., the length of the PDU itself without any of the encapsulation headers), exceeds the MTU of the destination layer 2 interface, the PDU MUST be dropped.

4. ATM

Two encapsulations are supported for ATM transport: one for ATM AAL5 and another for ATM cells.

The AAL5 CPCS-SDU encapsulation consists of the REQUIRED control word, and the AAL5 CPCS-SDU. The ATM cell encapsulation consists of an OPTIONAL control word, a 4 byte ATM cell header, and the ATM cell

payload.

4.1. ATM AAL5 CPCS-SDU Mode

In ATM AAL5 mode the ingress router is required to reassemble AAL5 CPCS-SDUs from the incoming VC and transport each CPCS-SDU as a single packet. No AAL5 trailer is transported. The control word is REQUIRED; its use, however, is optional, although desirable. Use of the control word means that the ingress and egress LSRs follow the procedures below. If an ingress LSR chooses not to use the control word, it MUST set the flags in the control word to 0; if an egress LSR chooses to ignore the control word, it MUST set the ATM control bits to 0.

The EFCI and CLP bits are carried across the network in the control word. The edge routers that implement this document MAY, when either adding or removing the encapsulation described herein, change the EFCI bit from zero to one in order to reflect congestion in the network that is known to the edge routers, and the CLP bit from zero to one to reflect marking from edge policing of the ATM Sustained Cell Rate. The EFCI and CLP bits MUST NOT be changed from one to zero.

The AAL5 CPCS-SDU is prepended by the following header:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Rsvd  |T|E|L|C|     Length    |         Sequence Number       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      ATM  AAL5 CPCS-SDU                       |
|                             "                                 |
|                             "                                 |
|                             "                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```


* T (transport type) bit

Bit (T) of the control word indicates whether the packet contains an ATM cell or an AAL5 CPCS-SDU. If set the packet contains an ATM cell, encapsulated according to the ATM cell mode section below, otherwise it contains an AAL5 CPCS-SDU. The ability to transport an ATM cell in the AAL5 mode is intended to provide a means of enabling OAM functionality over the AAL5 VC.

* E (EFCI) Bit

  The ingress router, R1, SHOULD set this bit to 1 if the EFCI bit of the final cell of those that transported the AAL5 CPCS-SDU is set to 1, or if the EFCI bit of the single ATM cell to be transported in the packet is set to 1. Otherwise this bit SHOULD be set to 0. The egress router, R2, SHOULD set the EFCI bit of all cells that transport the AAL5 CPCS-SDU to the value contained in this field.

* L (CLP) Bit

  The ingress router, R1, SHOULD set this bit to 1 if the CLP bit of any of the ATM cells that transported the AAL5 CPCS-SDU is set to 1, or if the CLP bit of the single ATM cell to be transported in the packet is set to 1. Otherwise this bit SHOULD be set to 0. The egress router, R2, SHOULD set the CLP bit of all cells that transport the AAL5 CPCS-SDU to the value contained in this field.

* C (Command / Response Field) Bit

  When FRF.8.1 Frame Relay / ATM PVC Service Interworking [3] traffic is being transported, the CPCS-UU Least Significant Bit (LSB) of the AAL5 CPCS-SDU may contain the Frame Relay C/R bit. The ingress router, R1, SHOULD copy this bit to the C bit of the control word. The egress router, R2, SHOULD copy the C bit to the CPCS-UU Least Significant Bit (LSB) of the AAL5 CPCS PDU.

4.2. ATM Cell Mode

In this encapsulation mode ATM cells are transported individually without a SAR process. The ATM cell encapsulation consists of an OPTIONAL control word, and one or more ATM cells - each consisting of a 4 byte ATM cell header and the 48 byte ATM cell payload. This ATM cell header is defined as in the FAST encapsulation [4] section 3.1.1, but without the trailer byte. The length of each frame, without the encapsulation headers, is a multiple of 52 bytes long.

The maximum number of ATM cells that can be fitted in a frame, in this fashion, is limited only by the network MTU and by the ability of the egress router to process them. The ingress router MUST NOT send more cells than the egress router is willing to receive. The number of cells that the egress router is willing to receive may either be configured in the ingress router or may be signaled, for example using the methods described in [1]. The number of cells encapsulated in a particular frame can be inferred by the frame length. The control word is OPTIONAL. If the control word is used then the flag bits in the control word are not used, and MUST be set to 0 when transmitting, and MUST be ignored upon receipt.

The EFCI and CLP bits are carried across the network in the ATM cell header. The edge routers that implement this document MAY, when either adding or removing the encapsulation described herein, change the EFCI bit from zero to one in order to reflect congestion in the network that is known to the edge router, and the CLP bit from zero to one to reflect marking from edge policing of the ATM Sustained Cell Rate. The EFCI and CLP bits SHOULD NOT be changed from one to zero.

This diagram illustrates an encapsulation of two ATM cells:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Control word ( Optional )                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          VPI          |              VCI              | PTI |C|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  ATM Payload ( 48 bytes )                     |
|                           "                                   |
|                           "                                   |
|                           "                                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          VPI          |              VCI              | PTI |C|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  ATM Payload ( 48 bytes )                     |
|                           "                                   |
|                           "                                   |
|                           "                                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

* VPI

The ingress router MUST copy the VPI field from the incoming cell into this field. For particular emulated VCs, the egress router

MAY generate a new VPI and ignore the VPI contained in this field.

* VCI

The ingress router MUST copy the VCI field from the incoming ATM cell header into this field. For particular emulated VCs, the egress router MAY generate a new VCI.

* PTI & CLP (C bit)

The PTI and CLP fields are the PTI and CLP fields of the incoming ATM cells. The cell headers of the cells within the packet are the ATM headers (without HEC) of the incoming cell.

4.2.1. OAM Cell Support

OAM cells MAY be transported on the VC LSP. An egress router that does not support transport of OAM cells MUST discard frames that contain an ATM cell with the high-order bit of the PTI field set to 1. A router that supports transport of OAM cells MUST follow the procedures outlined in [4] section 8 for mode 0 only, in addition to the applicable procedures specified in [1].

4.2.2. CLP bit to Quality of Service mapping

The ingress router MAY consider the CLP bit when determining the value to be placed in the Quality of Service fields (e.g. the EXP fields of the MPLS label stack) of the encapsulating protocol. This gives the network visibility of the CLP bit. Note however that cells from the same VC MUST NOT be reordered.

5. Using an MPLS Label as the Demultiplexer Field

To use an MPLS label as the demultiplexer field, a 32-bit label stack entry [2] is simply prepended to the emulated VC encapsulation, and

hence will appear as the bottom label of an MPLS label stack. This label may be called the "VC label". The particular emulated VC identified by a particular label value must be agreed by the ingress and egress LSRs, either by signaling (e.g, via the methods of [1]) or by configuration. Other fields of the label stack entry are set as follows.

5.1. MPLS Shim EXP Bit Values

If it is desired to carry Quality of Service information, the Quality of Service information SHOULD be represented in the EXP field of the VC label. If more than one MPLS label is imposed by the ingress LSR, the EXP field of any labels higher in the stack SHOULD also carry the same value.

5.2. MPLS Shim S Bit Value

The ingress LSR, R1, MUST set the S bit of the VC label to a value of 1 to denote that the VC label is at the bottom of the stack.

5.3. MPLS Shim TTL Values

The ingress LSR, R1, SHOULD set the TTL field of the VC label to a value of 2.

6. Security Considerations

This document specifies only encapsulations, and not the protocols used to carry the encapsulated packets across the network. Each such protocol may have its own set of security issues, but those issues are not affected by the encapsulations specified herein.

7. Intellectual Property Disclaimer

This document is being submitted for use in IETF standards discussions.

8. References

[1] "Transport of Layer 2 Frames Over MPLS", draft-martini-l2circuit-trans-mpls-09.txt. (work in progress)

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

[3] "Frame Relay / ATM PVC Service Interworking Implementation Agreement", Frame Relay Forum 2000.

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

[5] "Encapsulation Methods for Transport of PPP/HDLC Frames Over IP and MPLS Networks", draft-martini-ppp-hdlc-encap-mpls-00.txt. (work in progress)

[6] "Encapsulation Methods for Transport of Ethernet Frames Over IP and MPLS Networks", draft-martini-ethernet-encap-mpls-00.txt. (work in progress)

[7] "Encapsulation Methods for Transport of Frame-Relay Over IP and MPLS Networks", draft-martini-frame-encap-mpls-00.txt. (work in progress)

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
PacketExchange Ltd.
The Truman Brewery
91 Brick Lane
LONDON E1 6QL
United Kingdom
e-mail: giles@packetexchange.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

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
Pittsburgh, PA 15205
e-mail: sjv@laurelnetworks.com

Andrew G. Malis
Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134
e-mail: Andy.Malis@vivacenetworks.com

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

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

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave

Sunnyvale, CA 94089
e-mail: kireeti@juniper.net