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

Vasile Radoaca
Nortel Networks

Daniel Tappan
Eric C. Rosen
Cisco Systems, Inc.

Kireeti Kompella
Juniper Networks

Andrew G. Malis
Vinai Sirkay
Vivace Networks, Inc.

April 2002

Encapsulation Methods for Transport of Ethernet Frames Over IP and MPLS Networks

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

Status of this Memo

Abstract

This document describes methods for encapsulating the Protocol Data Units (PDUs) of Ethernet for transport across an MPLS or IP network.

Table of Contents

1 Specification of Requirements 2
2 Introduction 2
3 General encapsulation method 3
3.1 The Control Word 4
3.1.1 Setting the sequence number 5
3.1.2 Processing the sequence number 5
3.2 MTU Requirements 6
4 Protocol-Specific Details 6
4.1 Ethernet VLAN 6
4.2 Ethernet 7
5 Using an MPLS Label as the Demultiplexer Field 7
5.1 MPLS Shim EXP Bit Values 7
5.2 MPLS Shim S Bit Value 7
5.3 MPLS Shim TTL Values 8
6 Security Considerations 8
7 Intellectual Property Disclaimer 8
8 References 8
9 Author Information 9

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 the ethernet protocols. 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. [4] [5] [6]

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

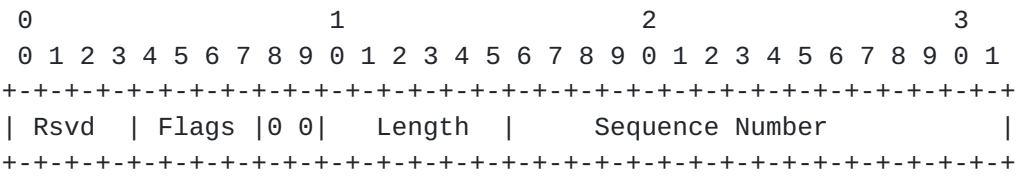
There are three requirements that may need to be satisfied when transporting layer 2 protocols over an IP or MPLS backbone:

-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:



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. Protocol-Specific Details

4.1. Ethernet VLAN

For an Ethernet 802.1q VLAN the entire Ethernet frame without the preamble or FCS is transported as a single packet. 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 4 byte VLAN tag is transported as is, and MAY be overwritten by the egress router.

The ingress router MAY consider the user priority field [3] of the VLAN tag header when determining the value to be placed in the Quality of Service field of the encapsulating protocol (e.g., the EXP fields of the MPLS label stack). In a similar way, the egress router MAY consider the Quality of Service field of the encapsulating protocol when queuing the packet for egress. Ethernet packets

containing hardware level CRC errors, framing errors, or runt packets MUST be discarded on input.

4.2. Ethernet

For simple Ethernet port to port transport, the entire Ethernet frame without the preamble or FCS is transported as a single packet. 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. As in the Ethernet VLAN case, Ethernet packets with hardware level CRC errors, framing errors, and runt packets MUST be discarded on input.

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

[4] "Encapsulation Methods for Transport of ATM Cells/Frame Over IP and MPLS Networks", draft-martini-atm-encap-mpls-00.txt. (work in progress)

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

[6] "Encapsulation Methods for Transport of PPP/HDLC Frames Over IP and MPLS Networks", draft-martini-ppp-hdlc-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

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

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

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