Network Working Group
Internet Draft
Expiration Date: November 2001

Luca Martini
Nasser El-Aawar
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.

May 2001

Encapsulation Methods for Transport of Layer 2 Frames Over MPLS

draft-martini-l2circuit-encap-mpls-02.txt

Status of this Memo

Abstract

This document describes methods for encapsulating the Protocol Data Units (PDUs) of layer 2 protocols such as Frame Relay, ATM AAL5, or Ethernet for transport 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 encapsulation procedures for accomplishing this. One possible control protocol method is described in [1]. QoS related issues are not discussed in this draft. For the purpose of this document R1 will be defined as the ingress LSR, and R2 as the egress LSR. A layer 2 PDU will be received at R1, encapsulated at R1, transported, decapsulated at R2, and transmitted out of R2. In a similar way, the "VC label" is defined as the label at the bottom of the label stack used to transmit the layer 2 PDU.

3. General encapsulation method

When transporting layer 2 protocols over MPLS it is, in most cases, not necessary to transport the layer 2 encapsulation across the MPLS network. In most cases the layer 2 header can be stripped at R1, and reproduced at R2 with the help of some extra encapsulation information, some of which is a priori signaled, and some of which may be carried in the control word described below.

3.1. The Control Word

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

-i. Sequentiality may need to be preserved.

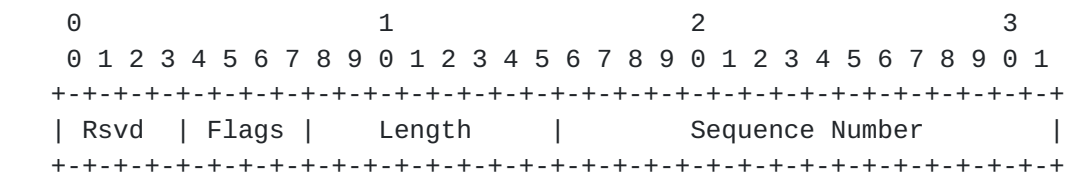
-ii. Small packets may need to be padded in order to be transmitted on a medium where the minimum transport unit is larger than the actual packet size.

-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.

In all cases the the egress LSR must be aware of whether the ingress LSR will send a control word over a specific virtual circuit. This may be achived by configuration of the LSRs, 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 8 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 256 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 LSR, 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

Given a VC label V and a pair of LSRs R1 and R2, where R2 has distributed V to R1. If R1 supports packet sequencing then the following procedures should be used:

- the initial packet transmitted to label V 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 LSR 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 an LSR R2 supports receive sequence number processing, then the following procedures should be used:

When a VC label V is first distributed, the "expected sequence number" associated with V MUST be initialized to 1

When a packet is received with label V 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 an LSR R2 does not support receive sequence number processing, then the sequence number field MAY be ignored.

3.2. MTU Requirements

The MPLS network MUST be configured with an MTU that is sufficient to transport the largest frame size that will be transported in the LSPs. Note that this is likely to be 12 or more bytes greater than the largest frame size. If a packet length, once it has been encapsulated on the ingress LSR, exceeds the LSP MTU, it MUST be dropped. If an egress LSR receives a packet on a VC LSP with a length, once the label stack and control word have been popped, that exceeds the MTU of the destination layer 2 interface, it MUST be dropped.

3.3. MPLS Shim EXP Bit Values

The ingress LSR, R1, SHOULD set the EXP field of the VC label to the same value as the EXP field of the previous label in the stack (if in fact a stack of more than one label is imposed at the ingress.) This will ensure that the EXP field will be visible to the egress LSR, R2, in the event of the packet having been penultimate hop popped.

3.4. 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.

3.5. MPLS Shim TTL Values

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

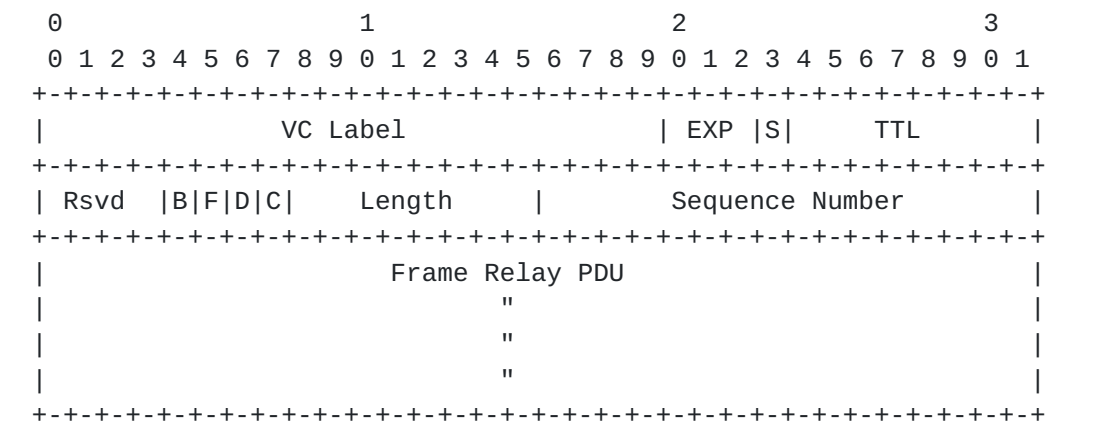
4. Protocol-Specific Details

4.1. Frame Relay

A Frame Relay PDU is transported without the Frame Relay header or the FCS. The control word is REQUIRED.

The BECN, FECN, DE and C/R bits are carried across the network in the control word. The edge LSRs that implement this document MAY, when either adding or removing the encapsulation described herein, change the BECN and/or FECN bits from zero to one in order to reflect congestion in the MPLS network that is known to the edge LSRs, and the D/E bit from zero to one to reflect marking from edge policing of the Frame Relay Committed Information Rate. The BECN, FECN, and D/E bits MUST NOT be changed from one to zero.

The following is an example of a Frame Relay packet:



* B (BECN) Bit

  The ingress LSR, R1, MUST copy the BECN field from the incoming Frame Relay header into this field. The egress LSR, R2, MUST generate a new BECN field based on the value of the B bit.

* F (FECN) Bit

  The ingress LSR, R1, MUST copy the FECN field from the incoming Frame Relay header into this field. The egress LSR, R2, MUST generate a new FECN field based on the value of the F bit.

* D (DE) Bit

  The ingress LSR, R1, MUST copy the DE field from the incoming Frame Relay header into this field. The egress LSR, R2, MUST generate a new DE field based on the value of the D bit.

  The ingress LSR, R1, MAY consider the DE bit of the Frame Relay header when determining the value to be placed in the EXP field of the MPLS label stack. In a similar way, the egress LSR, R2, MAY consider the EXP field of the VC label when queuing the packet for egress. Note however that frames from the same VC MUST NOT be reordered by the MPLS network.

* C (C/R) Bit

  The ingress LSR, R1, MUST copy the C/R bit from the received Frame Relay PDU to the C bit of the control word. The egress LSR, R2, MUST copy the C bit into the output frame.

The Label, EXP, S, and TTL fields are described in [2].

4.2. ATM

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

The AAL5 CPCS-PDU encapsulation consists of the MPLS label stack, a REQUIRED control word, and the AAL5 CPCS-PDU.

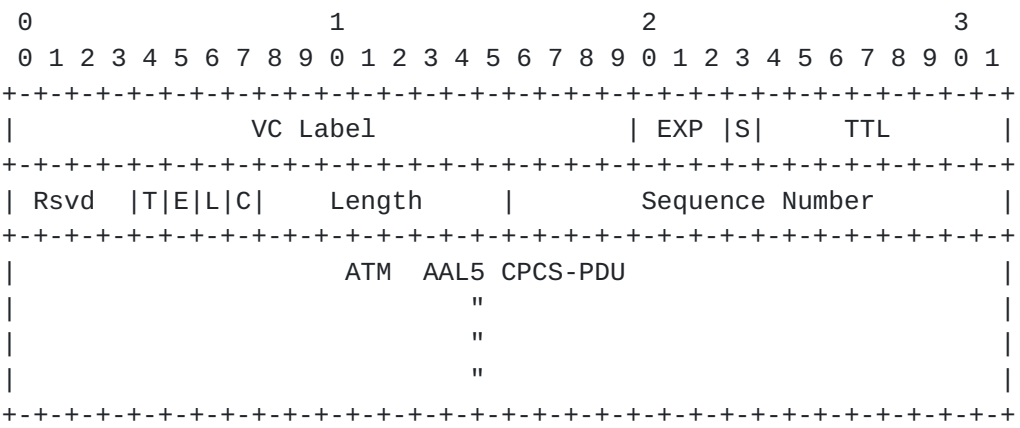
The ATM cell encapsulation consists of an MPLS label stack, an OPTIONAL control word, a 4 byte ATM cell header, and the ATM cell payload.

4.2.1. ATM AAL5 CPCS-PDU Mode

In ATM AAL5 mode the ingress LSR is required to reassemble AAL5 CPCS-PDUs from the incoming VC and transport each CPCS-PDU as a single packet. No AAL5 trailer is transported. The control word is REQUIRED.

The EFCI and CLP bits are carried across the network in the control word. The edge LSRs 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 MPLS network that is known to the edge LSRs, 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-PDU is prepended by the following header:



* T (transport type) bit

  Bit (T) of the control word indicates whether the MPLS packet contains an ATM cell or an AAL5 CPCS-PDU. If set the packet contains an ATM cell, encapsulated according to the ATM cell mode

section below, otherwise it contains an AAL5 CPCS-PDU. 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 LSR, R1, SHOULD set this bit to 1 if the EFCI bit of the final cell of those that transported the AAL5 CPCS-PDU is set to 1, or if the EFCI bit of the single ATM cell to be transported in the MPLS packet is set to 1. Otherwise this bit SHOULD be set to 0. The egress LSR, R2, SHOULD set the EFCI bit of all cells that transport the AAL5 CPCS-PDU to the value contained in this field.

* L (CLP) Bit

  The ingress LSR, R1, SHOULD set this bit to 1 if the CLP bit of any of the ATM cells that transported the AAL5 CPCS-PDU is set to 1, or if the CLP bit of the single ATM cell to be transported in the MPLS packet is set to 1. Otherwise this bit SHOULD be set to 0. The egress LSR, R2, SHOULD set the CLP bit of all cells that transport the AAL5 CPCS-PDU 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-PDU may contain the Frame Relay C/R bit. The ingress LSR, R1, SHOULD copy this bit to the C bit of the control word. The egress LSR, R2, SHOULD copy the C bit to the CPCS-UU Least Significant Bit (LSB) of the AAL5 CPCS PDU.

The Label, EXP, S, and TTL fields are described in [2].

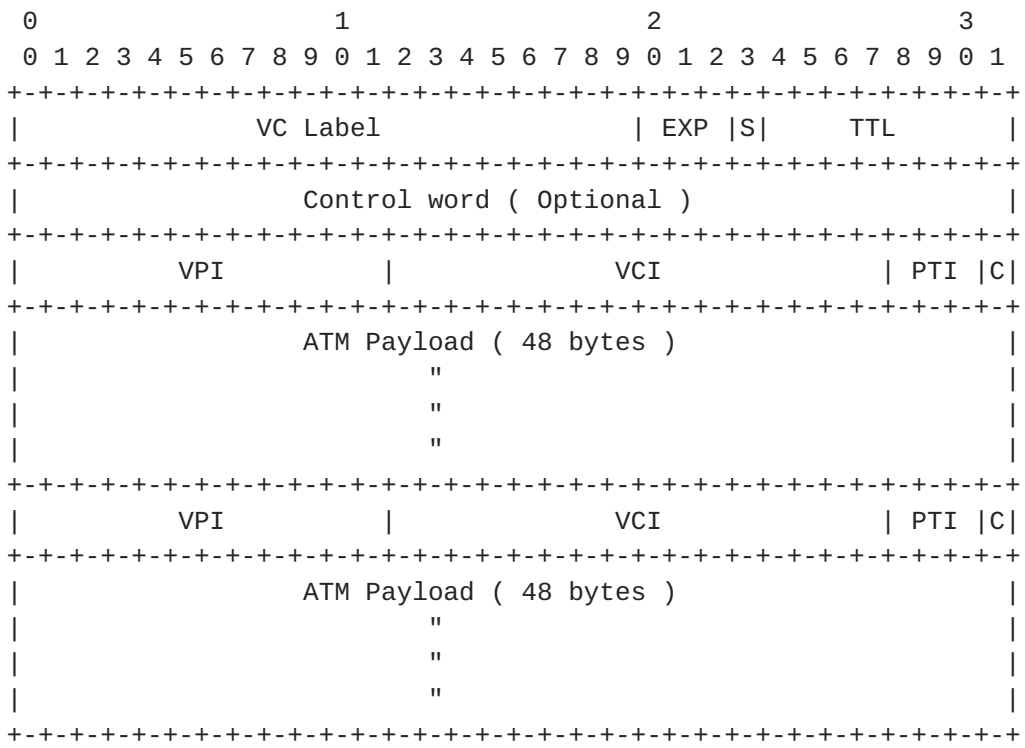
4.2.2. ATM Cell Mode

In this encapsulation mode ATM cells are transported individually without a SAR process. The ATM cell encapsulation consists of an MPLS label stack, 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 MPLS header and the control word, is a multiple of 52 bytes long. The maximum number of ATM cells that can be fitted in an MPLS frame, in this fashion, is limited only by the MPLS network MTU and by the ability of the egress LSR to process them. The ingress

LSR MUST NOT send more cells than the egress LSR is willing to receive. The number of cells that the egress LSR is willing to receive may either be configured in the ingress LSR 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 LSRs 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 MPLS network that is known to the edge LSRs, 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.

This diagram illustrates an encapsulation of two ATM cells:



* VPI

  The ingress router MUST copy the VPI field from the incoming cell into this field. The egress router MAY generate a new VPI based on the value of the VC label 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. The egress router MAY generate a new VCI based on the value of the VC label.

* 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.

4.2.3. OAM Cell Support

OAM cells MAY be transported on the VC LSP. A router that does not support transport of OAM cells MUST discard incoming MPLS frames on an ATM VC LSP 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].

4.2.4. CLP bit to MPLS label stack EXP bit mapping

The ingress LSR MAY consider the CLP bit when determining the value to be placed in the EXP fields of the MPLS label stack. This will give the MPLS network visibility of the CLP bit. Note however that cells from the same VC MUST NOT be reordered by the MPLS network.

4.3. 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 LSR.

The ingress LSR MAY consider the user priority field [5] of the VLAN tag header when determining the value to be placed in the EXP fields of the MPLS label stack. In a similar way, the egress LSR MAY consider the EXP field of the VC label when queuing the packet for egress. Ethernet packets containing hardware level CRC errors, framing errors, or runt packets MUST be discarded on input.

4.4. 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.

4.5. HDLC (Cisco)

HDLC (Cisco) mode provides port to port transport of Cisco HDLC encapsulated traffic. The HDLC PDU is transported in its entirety, including the HDLC address, control and protocol fields, but excluding HDLC flags and the FCS. Bit stuffing is undone. 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.

4.6. PPP

PPP mode provides point to point transport of PPP encapsulated traffic, as specified in [6]. The PPP PDU is transported in its entirety, including the protocol field (whether compressed using PFC or not), but excluding any media-specific framing information, such as HDLC address and control fields or FCS. Since media-specific framing is not carried the following options will not operate correctly if the PPP peers attempt to negotiate them:

Frame Check Sequence (FCS) Alternatives Address-and-Control-Field-Compression (ACFC) Asynchronous-Control-Character-Map (ACCM)

Note also that VC LSP Interface MTU negotiation as specified in [1] is not affected by PPP MRU advertisement. Thus if a PPP peer sends a PDU with a length in excess of that negotiated for the VC LSP that PDU will be discarded by the ingress LSR.

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.

5. Security Considerations

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

6. Intellectual Property Disclaimer

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

7. References

[1] "Transport of Layer 2 Frames Over MPLS", draft-martini-l2circuit-trans-mpls-06.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] "IEEE 802.3ac-1998" IEEE standard specification.

[6] "The Point-to-Point Protocol (PPP)", RFC 1661.

8. 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.
2607 Nicholson Rd.
Sewickley, PA 15143
e-mail: sjv@laurelnetworks.com

John Shirron
Laurel Networks, Inc.
2607 Nicholson Rd.
Sewickley, PA 15143
e-mail: jshirron@laurelnetworks.com

Toby Smith
Laurel Networks, Inc.
2607 Nicholson Rd.
Sewickley, PA 15143
e-mail: tob@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: 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
chris@cw.net