PWE3
Internet-Draft
Intended status: Standards Track
Expires: July 15, 2009

Moran Roth (Ed.)
Ronen Solomon
Corrigent Systems
Munefumi Tsurusawa
KDDI

January 15, 2009

Reliable Fibre Channel Transport Over MPLS Networks
draft-ietf-pwe3-fc-flow-00.txt

Status of this Memo

Copyright Notice

Abstract

A Fibre Channel pseudowire (PW) is used to carry Fibre Channel frames over an MPLS network. This enables service providers to offer "emulated" Fibre Channel services over existing MPLS networks. This document specifies the mechanisms controlling the reliable transport of Fibre Channel PW over MPLS networks. The encapsulation of Fibre Channel PDUs within a pseudowire and the procedures for using a PW to provide a Fibre Channel service are specified in [FC-encap].

Requirements Language

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 [1].

Table of Contents

1. Introduction

As metro transport networks migrate towards a packet-oriented network infrastructure, the PSN is being extended in order to allow all services to be transported over a common network infrastructure. This has been accomplished for services such as Ethernet [RFC4448], Frame Relay [RFC4619], ATM [RFC4717] and SONET/SDH [RFC4842] services. Another such service, which has yet to be addressed, is the transport of Fibre Channel (FC) frames over the PSN. This will allow network service providers to transparently carry FC services over the packet-oriented network, along with the aforementioned data and TDM services.

FC frames encapsulation within PW PDUs and the procedures for using a PW to provide a Fibre Channel service are specified in a companion document [FC-encap]. However, complementary mechanisms are required to provide reliable FC transport between FC entities.

This document specifies the mechanisms to provide reliable transport for FC traffic over MPLS networks, specifically, two mechanisms are specified:

(1) Congestion avoidance: this mechanism is intended to alleviate congestion conditions. In order to provide TCP-friendly operation the transmission rate is controlled by the throughput equation specified in [RFC3448].

(2) Selective Retransmission: this mechanism enable lossless transmission of FC frames by retransmission of PW PDUs that were discarded in transit in order to provide lossless transmission for the FC service. A selective retransmission mechanism is specified.

2. Congestion Control

FC PW traffic can be transmitted over networks that may experience congestion due to statistical multiplexing. When congestion conditions are experienced frames may be discarded within the MPLS PSN. Congestion control mechanism is required to prevent congestion collapse and provide fairness among the different connections. Fairness is usually defined with respect to TCP flow control [RFC2914]. The FC PW relies on a congestion control mechanism that provides TCP-friendly behavior by controlling the transmission rate into the PSN by a rate shaper, whose output rate is a function of network congestion.

Frame loss within the MPLS PSN also requires a reliable transmission mechanism in the PE to support faithful emulation of FC service, providing in-order, no-loss transport of FC traffic between CE1 and CE2. Reliable transmission is provided by a sliding-window selective retransmission (SR) mechanism to allow efficient retransmission of lost frames. This was standardized for FC transport in [FC-BB]. The SR mechanism also provides congestion indication (i.e. Frame loss events) to the rate control mechanism.

2.1. Rate Control

The rate control mechanism provides adaptive shaper control in response to network congestion indications. The rate shaper is configured with BW attributes, such as CIR and EIR, assigned to the FC PW service. The rate control operation is based on [RFC3448]. In the following sections the applicability of [RFC3448] to FC PW is analyzed, and rate control operation is detailed.

[RFC3448] is a receiver-based congestion control mechanism, where the congestion control information (i.e., the loss event rate) is calculated by the receiver. In FC PW, on the other hand, the congestion control information is calculated by the sender. This approach is more appropriate for the point-to-point nature of FC PW. This sender-based approach is also mentioned in [RFC3448] as a possible variant of the protocol.

2.1.1. Protocol Mechanism

In accordance with [RFC3448] the actual allowed sending rate is directly computed by a throughput equation, as a function of lost frames and round trip time. In general, the congestion control mechanism works as follows:

- The receiver detects lost frames and feeds this information back to the sender as part of the SR mechanism.
- The sender calculates the frame loss probability and measures the round-trip time (RTT) as defined in [FC-BB].
- The lost frame probability and RTT are then fed into the throughput equation, calculating the acceptable transmission rate.
- The sender then adjusts its transmission rate to match the calculated rate in accordance with the service BW attributes (CIR, EIR).

As the CIR is guaranteed, the throughput equation controls only the excess transmission rate. The parameters of the throughput equation are set as follows:

- The retransmission timeout (t_RTO) is replaced by the T1 timer of the SR mechanism as defined in section 6.3.
- The number of frames acknowledged by a single SR acknowledgment frame (b) is set to b = 1, as recommended in [RFC3448]. Different implementation MAY use delayed acknowledgement by increasing the value of b.

Frame loss probability (p) and RTT (R) are calculated as specified in Section 6.1.2.

2.1.2. Data Sender Protocol

The data sender sends a stream of data frames to the data receiver at a controlled rate. When a feedback frame is received from the data receiver, the data sender calculates the frame loss probability and changes its sending rate accordingly. If the sender does not receive a feedback frame during a timeout period, it reduces its sending rate. This is achieved by the SR T1 timer.

We specify the sender-side protocol in the following steps:

- Sender initialization.
- The sender behavior when a feedback frame is received.
- The sender calculation of the frame loss probability.
- The sender behavior when a feedback frame is not received for a timeout period.

The sender rate shaper is initialized to transmit at the CIR. The SR mechanism is also initialized by resetting the sequence numbers.

The sender calculates RTT (R), based on delay measurement frames transmitted by the NSP (as defined in [FC-BB]). These frames MUST be sent at least every 100 milliseconds, and are used to measure round trip samples that are averaged to obtain RTT (refer to [RFC3448] section 4.3 for details). If an RTT measurement is missed (either due to a loss of a delay measurement frame, or to an RTT larger than the measurement period), PE1 SHOULD shut the PW down, as specified in [FC-BB].

The sender calculates the frame loss probability based on feedback frames generated by the receiver. A feedback frame with accordance to the SR mechanism defined in [FC-BB] is one of the following:

- Receiver Ready (RR) - a frame that includes the N(R) counter to acknowledge the sender frames up to frame N(R).
- Receiver Not Ready (RNR) - a frame that includes the N(R) counter to acknowledge the sender frames up to frame N(R), and pause the sender from sending additional frames.
- Selective Reject (SREJ) - a frame that includes lost frames indication (sequence numbers).

When the sender receives a feedback frame it re-calculates the frame loss probability. RR and RNR will effectively decrease the frame loss probability due to no frame loss. On the other hand, reception of a SREJ frame tends to increase the frame loss probability. An implementation MAY consider sending feedback frames, in a controlled network environment, with expedite forwarding (EF) CoS to assure delivery.

After the frame loss probability is updated, the sender calculates a new transmission rate for the rate shaper. The transmission rate is calculated as: Rate = MIN(PIR, MAX(CIR,X)), where CIR is the Committed Information Rate, PIR is the Peak Information Rate (PIR = CIR+EIR), X is the outcome of the throughput equation as specified in [RFC3448], and MIN/MAX are functions returning the smallest/largest value among their operands, respectively. Note that the transmission rate as controlled by the above function, is bounded in the range [CIR,PIR].

No feedback in accordance with [RFC3448] is defined by the timer T1. When the sender does not receive a feedback for such an interval it halves its transmission rate as defined in [RFC3448]. The transmission rate equation as specified above MUST still be applied to guarantee that the CIR is the lower limit for the throughput. The procedure controlling timer T1 (refer to Section 6.3) guarantees that transmission rate is not halved during idle periods, as the timer is not activated during these periods.

The maximum burst size allowed MUST be limited to a round-trip time's worth of packets, to achieve efficient transmission while conforming to [RFC3448].

In case the transmission rate is equal to CIR for a period greater than RTT, and transmitted frames are still lost in transit, as

indicated to the sender by receiving SREJ frames, the sender MUST signal PW status of "Unable to maintain minimum transmission rate" (refer to Section 9 - "IANA Considerations" for details) as specified in [RFC4447], and MUST stop transmission over the PW for a duration of 10 seconds (this period allows a transient network problem to resolve itself, and guarantees that no more than one HELLO message [FC-SW] is lost, and the link between the two FC devices is not affected). The sender and receiver MUST discard all frames residing in the buffers associated with the congested PW. The sender MUST also discard all frames received from the attached FC device. If within 10 seconds after transmission was restarted severe congestion conditions are encountered, as described above (i.e., CIR cannot be maintained), the sender MUST shut the PW down, as described in Section 6.5 of [RFC3985]. If the PW has been set up using the PWE3 control protocol [RFC4447], the regular PW teardown procedures SHOULD be used. The PW MUST NOT be automatically restarted, and administrative intervention is required. Upon PW shutdown the sender and receiver MUST discard all frames associated with the PW. Note that congestion may be avoided by employing connection admission control (CAC) mechanism, which assures that congestion conditions will not be reached when a PW is transmitting at its configured CIR.

2.1.3. Data Receiver Protocol

The data receiver receives a stream of data frames from the data sender, generates SR feedback frames (SR-RR, SR-RNR and SR-SREJ), and sends them to the data sender. The details of feedback frames generation and transmission are specified in section 6.3.

2.2. Selective Retransmission overview

The selective retransmission mechanism provides efficient retransmission of lost frames to enable faithful emulation of FC service, with no frame loss experienced by the CE. The proposed selective retransmission mechanism was standardized for FC transport in [FC-BB], and is specified in details in this standard.

The SR protocol is an efficient sliding window full-duplex protocol that supports both the flow control and error recovery functions. SR has been adopted from ITU's Link Access Protocol B (LAPB) that was derived from ISO/IEC's High-level Data Link Control (HDLC) balanced classes. Use of LAPB in SR is limited to a subset of the synchronous modulo 32768 super sequence numbering service option.

SR works between two PE devices (see figure 7). SR flow control works by streaming multiple messages within an allowed window, bounded by

the system parameter k, and awaits acknowledgements before sending more messages. Acknowledgements indicate which messages were correctly received and there is a provision for requesting retransmission of selected messages in the current window. Fibre Channel Sequences and Exchanges are not visible to the SR flow control protocol which sees the PW packets constructed from the FC frames.

Some benefits of the SR protocol are summarized below:

a) it is used for reliable transport of all Class 2, 3, 4, and F frames between two PE devices;

b) it optimizes buffer management at the PE devices;

c) it acts as a congestion avoidance technique to match the capacity of the sender to the capacity of the network that carries the payload;

d) it ensures correct delivery of messages (i.e., an error control and recovery function); and

e) it provides a continuous stream of traffic across the MPLS PSN thus leading to a higher throughput (i.e., optimizes bandwidth utilization at each BBW device).

Note that the synchronization of the Sender PE and the Receiver PE at the PW message level, which is required for correct SR operation is performed through PW signaling.

The four different SR messages described in section 6.2.1 have a correspondence to the LAPB frame types. Note that only the information transfer SR-I message is flow-controlled while all other messages are control messages of the protocol.
The SR protocol specifies the maximum number (k) of outstanding messages at any given time. k is a system parameter that is not negotiated and is fixed in a given implementation. The value of this system parameter depends on the MPLS PSN delay characteristics and the number of buffers available. Typically, the value of k is expected to be far below the maximum number of 32767.

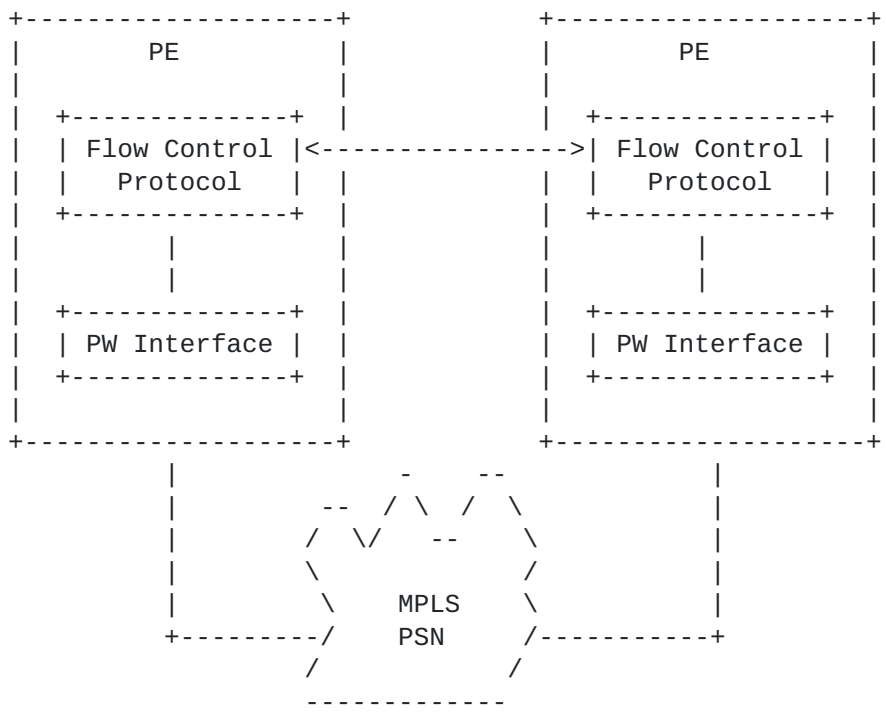

Figure 7 - SR flow control protocol between two PEs

2.2.1. FC Encapsulation Header

The FC Encapsulation Header defines two types of field formats that are used to perform information transfer (i.e., I-format frames), and supervisory functions (i.e., S-format frames). SR makes use of four different types of messages:

a) I-format (1): SR-I frame.

This frame is used to perform an information transfer. The Encapsulation Header of an I-format frame is shown in figure 8. The I-frame Encapsulation Header contains the following fields:
(1) N(S): Transmitter send sequence number.
(2) N(R): Transmitter receive sequence number.
(3) P: Poll bit (1 = Poll).

A detailed description of the different fields and explanation of The functionality involved is provided in section 6.2.2.

An SR-I frame is a command message (i.e., the A-bit in the Control Word is set to 1), and carries an encapsulated FC frame.

b) S-format (3): SR-RR, SR-RNR, SR-SREJ frames.

These frames are used to perform supervisory control functions of the Selective Retransmission mechanism, such as acknowledge SR-I messages, request retransmission of SR-I messages, and to request a temporary suspension of transmission of SR-I messages. The Encapsulation Header of an S-format frame is shown in figure 9.

The S-frame Encapsulation Header contains the following fields:

(1) N(R): Transmitter receive sequence number.

(2) S: Supervisory function bits to define the frame type.
S = 00: SR-RR.
S = 01: Reserved.
S = 10: SR-RNR.
S = 11: SR-SREJ.

(3) P: Poll/Final bit (refer to section 6.2.2 for detailed description).

(4) Reserved: MUST be set to 0 by the ingress PE, and MUST be ignored by the egress PE.

A detailed description of the different fields and explanation of The functionality involved is provided in section 6.2.2.

An SR-RR frame carries no payload, and may be either a command or response message (the A-bit in the Control Word is set to 1 for a Command, and to 0 for a Response). It indicates Ready to Receive SR-I messages (negates busy condition) and acknowledges previous SR-I messages.

An SR-RNR frame carries no payload, and may be either a command or response message. It indicates Receiver not Ready to accept more SR-I messages (busy condition) and acknowledges previous SR-I messages.

An SR-SREJ frame may be either a command or response message, and carries a payload that indicate SR-I frames in need of selective retransmission.

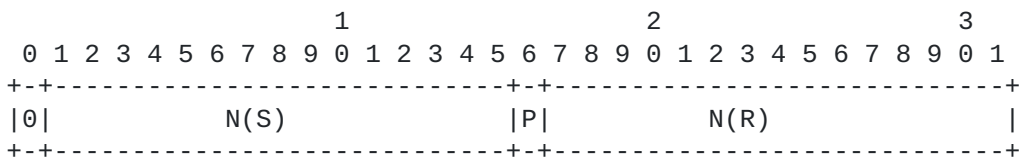

Figure 8 - FC Encapsulation Header format for I-frame

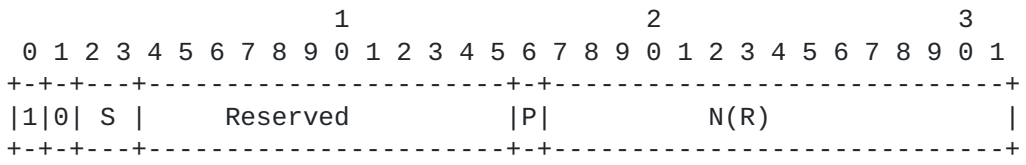

Figure 9 - FC Encapsulation Header format for S-frame

2.2.2. Encapsulation Header field parameters

The following describes the different fields of the Encapsulation Header and details how these fields are handled.

a) Modulus of SR - Each SR-I message is sequentially numbered and may have the value 0 through modulus minus 1, where "modulus" is equal to 32768 (i.e., the maximum value of the sequence numbers). The sequence numbers cycle through the entire range.

b) Send state variable V(S) - The send state variable V(S) denotes the sequence number of the next-in-sequence SR-I message to be transmitted. V(S) may take on the values 0 through modulus minus 1. The value of V(S) is incremented by 1 with each successive SR-I message transmission, but cannot exceed the N(R) of the last received SR-I or supervisory message by more than the maximum number of outstanding SR-I messages k. The value of k is defined in section 6.3.5.

c) Send sequence number N(S) - Only SR-I messages contain N(S), the send sequence number of the transmitted SR-I message. At the time that an in-sequence SR_I message is designated for transmission, the value of N(S) is set to the value of the send state variable V(S).

d) Receive state variable V(R) - The receive state variable V(R) denotes the sequence number of the next-in-sequence SR-I message expected to be received. V(R) may take on the values 0 through modulus minus 1. The value of V(R) is incremented by 1 by the receipt of an error-free, in-sequence SR-I message whose send sequence number N(S) equals the receive state variable V(R).

e) Rceive sequence number N(R) - All SR-I messages and supervisory messages, except SR-SREJ messages with the F bit set to 0, SHALL contain N(R), the expected send sequence number of the next received SR-I message. At the time that a message of the above types is designated for transmission, the value of N(R) is set to the current value of the receive state variable V(R). N(R) indicates that the PE transmitting the N(R) has correctly received all SR_I messages numbered up to and including N(R)-1.

f) Functions of the Poll/Final bit (P-bit) - All messages contain P-bit, the Poll/Final bit. In command messages, the P-bit is referred to as the Poll bit. In response messages it is referred to as the Final bit.

   The Poll bit set to 1 is used by the PE to solicit (i.e., poll) a response from the remote PE.

   The Final bit set to 1 is used by the PE to indicate the response message transmitted by the remote PE, as a result of the soliciting (i.e., poll) command.
   The use of the P/F bit is further described in section 6.3.3.

2.2.3. Selective reject (SR-SREJ) frame

The SR-SREJ supervisory message is used by a PE to request retransmission of one or more, not necessarily contiguous, SR-I messages. The N(R) field SHALL contain the sequence number of the earliest SR-I message to be retransmitted and the information field (see figure 9) SHALL contain, in ascending order (i.e., 32767 is higher than 32766 and 0 is higher than 32767 for modulo 32768), the sequence numbers of additional SR-I message(s), if any, that needs to be retransmitted.

The payload field SHALL be encoded such that there is a 2-byte field for each standalone SR-I message in need of retransmission, and a 4-byte span list for each sequence of two or more contiguously numbered SR-I messages in need of retransmission, as depicted in figure 9. Standalone SR-I messages are identified in the payload field by the appropriate N(R) value preceded by a 0 bit in the 2-byte field used. Span lists are identified in the payload field by the N(R) value of the first SR-I message in the span list preceded by a 1 bit in the 2-byte field used, followed by the N(R) value of the last message in the span list preceded by a 1 bit in the 2-byte field used.

The maximum payload size of a SR-SREJ message is 2148 bytes corresponding to a maximum possible encoding of 1074 standalone SR-I messages or a maximum possible encoding of 537 span list sets.
If the P-bit in an SR-SREJ message is set to 1, then SR-I messages numbered up to N(R)-1 inclusive, N(R) being the value in the Encapsulation Header field, shall be considered as acknowledged. If the P-bit in an SR-SREJ message is set to 0, then the N(R) in the Encapsulation Header field of the SR-SREJ message does not indicate acknowledgement of SR-I messages.

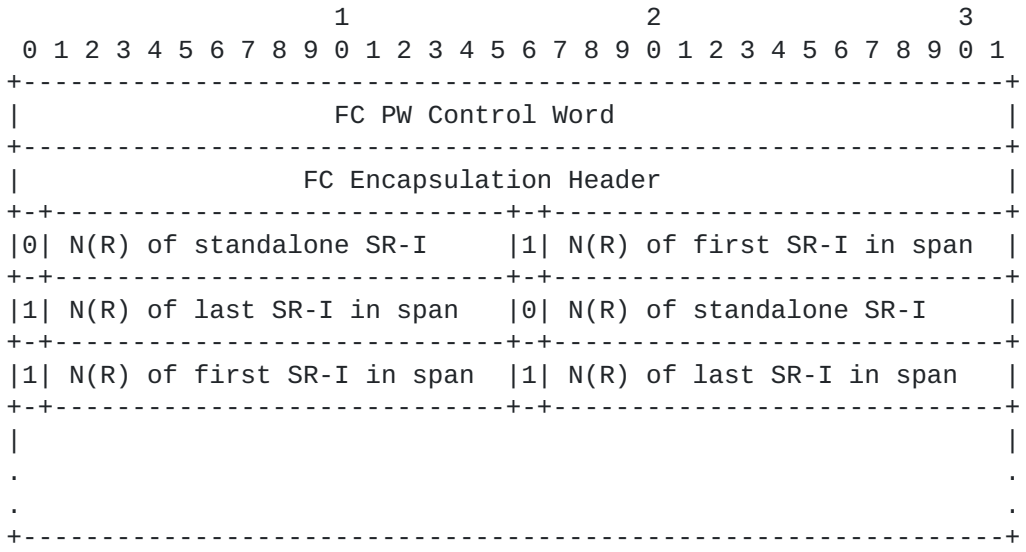

Figure 10 - SR-SREJ frame format example

2.2.4. Exception condition reporting and recovery

The error recovery procedures that are available to effect recovery following the detection/occurrence of an exception condition are described in this section. Exception conditions described are those situations that may occur as the result of transmission errors, PE device malfunction, or operational situations.

a) Busy condition - The busy condition results when the PE is temporarily unable to continue to receive SR-I messages due to internal constraints (e.g., receive buffering limitations). Upon entering the busy condition, a PE transmits an SR-RNR message. SR-I messages pending transmission may be transmitted from the busy PE prior to or following the SR-RNR message.
An indication that the busy condition has cleared is communicated by the transmission of SR-RR or SR-SREJ.

b) N(S) sequence error condition - The information field of all received SR-I messages whose N(S) is not in the range V(R) and V(R)+k-1 inclusive, SHALL be discarded. The information field of all SR-I messages received by the PE whose N(S) is in the range V(R) and V(R) + k -1 inclusive, SHALL be saved in the receive buffer.

An N(S) sequence error exception condition occurs in the receiver when a received SR-I message contains an N(S) that is not equal to the receive state variable V(R) at the receiver. The receiver SHALL not acknowledge (i.e., increment its receive state variable) the SR-I message causing the sequence error, or any SR-I message that may follow, until an SR-I message with the correct N(S) is received.

A PE that receives one or more valid SR-I messages having sequence errors or subsequent supervisory messages (i.e., SR-RR, SR-RNR, or SR-SREJ) shall accept and handle the N(R) field and the P-bit.

The means specified below shall be available for initiating the retransmission of lost or errored SR-I messages following the occurrence of an N(S) sequence error condition.

(1) SR-SREJ recovery - The SR-SREJ message shall be used to initiate more efficient error recovery by selectively requesting the retransmission of one or more, not necessarily contiguous, lost or errored SR-I message(s) following the detection of sequence errors, rather than requesting the retransmission of all SR-I messages.

When a PE receives an out-of-sequence message, the SR-I message shall be saved in a receive buffer. The SR-I message shall be delivered to the upper layer only when all SR-I messages numbered below N(S) are correctly received. If message number N(S)-1 has not been received previously, then an SR-SREJ response message with the P-bit set to 0 shall be transmitted containing the sequence numbers of the block of consecutive missing SR-I messages ending at N(S)-1. On receiving such an SR-SREJ message the PE shall retransmit all requested SR-I messages. After retransmitting these SR-I messages, the BBW may transmit new SR_I messages, if they become available.

When a PE receives a command message with the P-bit set to 1, if there are out-of-sequence SR-I messages saved in the receive buffer, it shall transmit an SR-SREJ message, with the

F bit set to 1, containing a complete list of missing sequence numbers. The PE that receives the SR-SREJ message shall retransmit all requested SR-I messages, except those that were transmitted subsequent to the last command message with the P bit set to 1.

(2) Time-out recovery - If a PE, due to a transmission error, does not receive, or receives and discards, a single SR-I message or the last SR-I message in a sequence of SR-I messages, it shall not detect a N(S) sequence error condition and, therefore, shall not transmit an SR-SREJ message.

The PE that transmitted the unacknowledged SR-I message(s) shall, following the completion of a system specified time-out period (see section 6.3.4 items b) and j) below), send a supervisory command message (i.e., SR-RR or SR-RNR) with the P-bit set to 1. SR-I messages shall be retransmitted on the receipt of an SR-RR response message with the F bit set to 1 or an SR-SREJ message.

c) Invalid message condition - Any message that is invalid shall be discarded, and no action is taken as the result of that message. An invalid message is defined as one that contains:

(1) the Control Word with an invalid encoding; or

(2) the Encapsulation Header with an invalid encoding.

2.3. Selective Retransmission procedures

2.3.1. SR mode of operation

The SR protocol shall be limited to a subset of the synchronous modulo 32768 super sequence numbering service option operation of the LAPB protocol. The SR protocol is initialized upon PW set-up following a successful signaling session.

2.3.2. SR procedure for addressing

The Address Bit field in the Control Word (see figure 3) identifies a message as either a command or a response. This field is used in conjunction with the P-bit (Poll/Final).

2.3.3. SR procedure for the use of the Poll/Final bit

The PE receiving a supervisory command (i.e., SR-RR, SR-RNR, SR-SREJ), or SR-I message with the P bit set to 1 shall set the F bit to 1 in the next response message it transmits.

The response message returned by the PE to an SR-I message with the P bit set to 1, shall be an SR-RR, SR-SREJ, or SR-RNR response with the F bit set to 1.

The response message returned by the PE to a supervisory command with the P bit set to 1, shall be an SR-RR, SR-RNR, or SR-SREJ response with the F bit set to 1.

The P bit may be used by the PE in conjunction with the timer recovery condition (see section 6.3.4. item j) below).

2.3.4. Procedures for information transfer

a) Procedures for SR-I messages

The procedures that apply to the transmission of SR-I messages in each direction using multi-selective reject are described below.

b) Sending new SR-I messages

When the PE has a new SR-I message to transmit (i.e., an SR-I message not already transmitted), it shall transmit it with a N(S) equal to its current send state variable V(S), and a N(R) equal to its current receive state variable V(R). At the end of the transmission of the SR-I message, it shall increment its send state variable V(S) by 1.

If the SR timer T1 is not running at the time of transmission of the SR-I message, it shall be started.

If the SR send state variable V(S) is equal to the last value N(R) received plus k, where k is the maximum number of outstanding SR-I frames (see section 6.3.5), the PE shall not transmit any new SR-I frames.

If the remote PE is busy, the PE shall not transmit any new SR-I messages.

When the PE is in the busy condition, it may still transmit SR-I messages, provided that the remote PE is not busy.

c) Receiving an in-sequence SR-I message

When the PE is not in a busy condition and receives a valid SR-I message whose send sequence number N(S) is equal to its receive state variable V(R), the PE shall accept the information field of this message and increment by one the receive state variable V(R). If the SR-I message, whose N(S) is equal to the incremented value of V(R), is present in the receive buffer, then the PE shall remove it from the receive buffer, deliver it to the upper layer and increment V(R) by one. The PE shall repeat this procedure until V(R) reaches a value such that the SR-I message whose N(S) is equal to V(R) is not present in the receive buffer. The PE shall then take one of the following actions:

(1) if the PE is now in the busy condition, it shall transmit an SR-RNR message with N(R) equal to the value of the SR receive variable V(R) (see item i) below); or

(2) if the PE is still not in a busy condition:

- if the P bit is set to 1, then the PE shall transmit a response message with the F bit set to 1, as specified in item l) below;

- if an SR-I message is available for transmission the PE shall act as described in item b) above, sending new SR-I messages and acknowledging the received SR-I message by setting N(R) in the Encapsulation Header field of the next transmitted SR-I message to the value of the SR receive state variable V(R), or the PE shall acknowledge the received SR-I message by transmitting an SR-RR message with the N(R) equal to the value of the SR receive state variable V(R); or

- the PE shall transmit an SR-RR message with N(R) equal to the value of the SR receive state variable V(R).

When the PE is in a busy condition, it may ignore the information field contained in any received SR-I message.

d) Reception of invalid messages

When the PE receives an invalid message (see 6.2.4. item c), it shall discard the message.

e) Reception of out-of-sequence SR-I messages

When the PE is not in a busy condition and it receives a valid SR-I message whose send sequence number N(S) is out-of-sequence, (i.e.,

not equal to the receive state variable V(R)), then it shall perform one of the following actions:

1) if N(S) is less than V(R) or greater than or equal to V(R) + k, then it shall discard the information field of the SR-I message. If the P bit of the SR-I message is set to 1, then the PE shall transmit a response message with the F bit set to 1, as specified in item l) below; or

2) if N(S) is greater than V(R) and less than V(R) + k, then it shall save the SR-I message in the receive buffer. It shall then perform one of the following actions:

   - if the P bit of the SR-I message is set to 1, then the PE shall transmit a response message with the F bit set to 1, as specified in item l) below;

   - if the PE is now in a busy condition, it shall transmit an SR-RNR message with N(R) equal to the value of the receive variable V(R), as specified in item i) below; or

   - if the SR-I message numbered N(S)-1 has not yet been received, then the PE shall transmit an SR-SREJ response message with the F bit set to 0. The PE shall create a list of contiguous sequence numbers N(X), N(X)+1, N(X)+2,..., N(S)-1, where N(X) is greater than or equal to V(R) and none of the SR-I messages N(X) to N(S)-1 have been received. The N(R) field of the SR-SREJ message shall be set to N(X) and the information field set to the list N(X)+1,...,N(S)-1. If the list of sequence numbers is too large to fit into the information field of the SR-SREJ message, then the list shall be truncated to fit in one SR-SREJ message, by including only the earliest sequence numbers.

When the PE is in the busy condition, it may ignore the information field contained in any received SR-I message.

f) Receiving acknowledgement

When correctly receiving an SR-I message or a supervisory message (i.e., SR-RR, SR-RNR, or SR-SREJ with the F bit set to 1), even in the busy condition, the PE shall consider the N(R) contained in this message as an acknowledgement for all the SR-I messages it has transmitted with a N(S) up to and including the received N(R)-1. The PE shall stop the timer T1 if the received supervisory message has the F bit set to 1 or if there is no outstanding poll condition and

the N(R) is higher than the last received N(R), actually acknowledging some SR-I messages.

If timer T1 has been stopped by the receipt of an SR-I message, an SR-RR command message, an SR-RR response message with the F bit set to 0, or an SR-RNR message, and if there are outstanding SR-I messages still unacknowledged, the PE shall restart timer T1. If timer T1 has been stopped by the receipt of an SR-SREJ message with the F bit set to 1, the PE shall follow the retransmission procedure specified in item g.2) below. If timer T1 has been stopped by the receipt of an SR-RR message with the F bit set to 1, the PE shall follow the retransmission procedure specified in item k) below.

g) Receiving an SR-SREJ response message

1) Receiving an SR-SREJ response message with the F bit set to 0

When receiving an SR-SREJ response message with the F bit set to 0, the PE shall retransmit all SR-I messages, whose sequence numbers are indicated in the N(R) field and the information field of the SR-SREJ message, in the order specified in the SR-SREJ message. Retransmission shall conform to the following:

- if the PE is transmitting a supervisory or SR-I message when it receives the SR-SREJ message, it shall complete that transmission before commencing transmission of the requested SR-I messages; or

- if the PE is not transmitting any message when it receives the SR-SREJ message, it shall commence transmission of the requested SR-I messages immediately.

If there is no outstanding poll condition, then a poll shall be sent, either by transmitting an SR-RR command, or SR-RNR command if the PE is in the busy condition, with the P bit set to 1 or by setting the P bit in the last retransmitted SR-I message and timer T1 shall be restarted.

If there is an outstanding poll condition, then timer T1 shall not be restarted.

2) Receiving an SR-SREJ response message with the F bit set to 1

When receiving an SR-SREJ response message with the F bit set to 1, the PE shall retransmit all SR-I messages, whose sequence numbers are indicated in the N(R) field and the information field of the SR-SREJ message, in the order specified in the

SR-SREJ message, except those messages that were sent after the message with the P bit set to 1 was sent. Retransmission shall conform to the following:

- if the PE is transmitting a supervisory message or SR-I message when it receives the SR-SREJ message, it shall complete that transmission before commencing transmission of the requested SR-I messages; or

- if the PE is not transmitting any message when it receives the SR-SREJ message, it shall commence transmission of the requested SR-I messages immediately.

If any messages are retransmitted, then a poll shall be sent, either by transmitting an SR-RR command, or SR-RNR command if the PE is in the busy condition, with the P bit set to 1 or by setting the P bit in the last retransmitted SR-I message.

Timer T1 shall be restarted.

h) Receiving an SR-RNR message

After receiving an SR-RNR message, the PE shall stop transmission of SR-I messages until an SR-RR or SR-SREJ message is received.

The PE shall start timer T1, if necessary, as specified in section 6.3.5.

When timer T1 runs out before receipt of a busy clearance indication, the PE shall transmit a supervisory message (i.e., SR-RR, SR-RNR), with the P bit set to 1 and shall restart timer T1, in order to determine if there is any change in the receive status of the remote PE. The remote PE shall respond to the P bit set to 1 with a supervisory response message (i.e., SR-RR, SR-RNR, SR-SREJ) with the F bit set to 1 indicating continuation of the busy condition (i.e., SR-RNR message) or clearance of the busy condition (i.e., SR-RR, SR-SREJ). Upon receipt of the remote PE response, timer T1 shall be stopped. The PE shall process the supervisory response message as follows:

1) if the response is an SR-RR message, the busy condition shall be assumed to be cleared and the PE may retransmit messages as specified in item k) below. New SR-I messages may be transmitted as specified in item b) above;

2) if the response is an SR-SREJ message, the busy condition shall be assumed to be cleared and the PE may retransmit messages as

specified in item g.2) above. New SR-I messages may be transmitted as specified in item b) above; or

3) if the response is an SR-RNR message, the busy condition shall be assumed to still exist and the PE, after a period of time (e.g., the duration of timer T1), shall repeat the enquiry of the remote PE receive status.

If timer T1 runs out before a status response is received, the enquiry process above shall be repeated. If N2 attempts to get a status response fail, the PE MAY declare the PW as down.

If, at any time during the enquiry process, an unsolicited SR-RR or SR-SREJ message is received from the remote PE, it shall be considered to be an indication of clearance of the busy condition. Should the unsolicited SR-RR message be a command message with the P bit set to 1, the appropriate response message with the F bit set to 1 shall be transmitted (see item l) below) before the PE may resume transmission of SR-I messages. The PE shall not clear the outstanding poll condition. The PE shall not stop timer T1. If an unsolicited SR-SREJ message is received, then the PE shall perform retransmissions as specified in item g.1) above.

i) BBW busy condition

When the PE enters a busy condition, it shall transmit an SR-RNR message at the earliest opportunity. The SR-RNR message shall be a command frame with the P bit set to 1 if an acknowledged transfer of the busy condition indication is required, otherwise the SR-RNR message may be a command or response message. While in the busy condition, the PE shall accept and process supervisory messages, accept and process the N(R) field of SR-I, SR-RR, and SR-SREJ messages with the F bit set to 1, and return an SR-RNR response with the F bit set to 1 if it receives a supervisory command or SR-I command message with the P bit set to 1. Received SR-I messages may be discarded or saved as specified in items c) and e) above, however, SR-RR or SR-SREJ messages shall not be transmitted. To clear the busy condition, the PE shall transmit an SR-RR message, with the N(R) field set to the current receive state variable V(R). The SR-RR message shall be a command message with the P bit set to 1 if an acknowledged transfer of the busy-to-non-busy transition is required, otherwise the SR-RR message may be either a command or response message.

j) Awaiting acknowledgement

If the timer T1 runs out while waiting for the acknowledgement of an SR-I message from the remote PE, the PE shall restart timer T1 and transmit an appropriate supervisory command message (i.e., SR-RR, SR-RNR) with the P bit set to 1. The PE may transmit new SR-I messages after sending this enquiry message.

If the PE receives an SR-SREJ response message with the F bit set to 1, the PE shall restart timer T1 and retransmit SR-I messages as specified in item g.2) above.

If the PE receives an SR-SREJ response message with the F bit set to 0, the PE shall retransmit SR-I messages as specified in item g.2) above.

If the PE receives an SR-RR response message with the F bit set to 1, the PE shall restart timer T1 and retransmit SR-I messages as specified in item k) below.

If the PE receives an SR-RR response message with the F bit set to 0, or an SR-RR command message or SR-I message with the P bit set to 0 or 1, the PE shall not restart timer T1, but shall use the received N(R) as an indication of acknowledgement of transmitted SR-I messages up to and including SR-I message numbered N(R)-1.

If timer T1 runs out before a supervisory response message with the F bit set to 1 is received, the PE shall retransmit an appropriate supervisory command message (i.e., SR-RR, SR-RNR) with the P bit set to 1. After N2 such attempts, the PE MAY declare the PW as down.

k) Receiving an SR-RR response messages with the F bit set to 1

When receiving an SR-RR response message with the F bit set to 1, the PE shall process the N(R) field as specified in item f) above. If there are outstanding SR-I messages that are unacknowledged and no new SR-I messages have been transmitted subsequent to the last message with the P bit set to 1, then the PE shall retransmit all outstanding SR-I messages except those that were sent after the message with the P bit set to 1 was sent. Retransmission shall conform to the following:

1) if the PE is transmitting a supervisory or SR-I message when it receives the SR-RR message, it shall complete that transmission before commencing transmission of the requested SR-I messages;

2) if the PE is not transmitting any message when it receives the SR-RR message, it shall commence transmission of the requested SR-I messages immediately.

If any messages are retransmitted, then a poll shall be sent, either by transmitting an SR-RR command, or SR-RNR command if the PE is in the busy condition, with the P bit set to 1 or by setting the P bit in the last retransmitted SR-I message.

The timer T1 shall be stopped. If any SR-I messages are outstanding, then timer T1 shall be started.

l) Responding to command messages with the P bit set to 1

When receiving an SR-RR, SR-RNR, or-SR_I command message with the P bit set to 1, the PE shall generate an appropriate response message as follows:

1) if the PE is in the busy condition, it shall transmit an SR-RNR response message with the F bit set to 1;

2) if there are some out-of-sequence messages in the receive buffer, then it shall transmit an SR-SREJ message with the F bit set to 1; N(R) shall be set to the receive state variable V(R) and the information field set to the sequence numbers of all missing SR-I messages, except V(R). If the list of sequence numbers is too large to fit in the information field of the SR-SREJ message, then the list shall be truncated by including only the earliest sequence numbers; or

3) if there are no out-of-sequence messages in the receive buffer, then an SR-RR response message with the F bit set to 1 shall be sent.

2.3.5. List of SR system parameters

a) SR Poll Timeout (Timer T1)

The same value of the timer T1 shall be made known and agreed to by the two PEs.

The period of timer T1, at the end of which retransmission of a message may be initiated (see 6.3.4), shall take into account whether T1 is started at the beginning or the end of the transmission of a message.

The proper operation of the procedure requires that the transmitter's timer T1 be greater than the maximum time between transmission of a message (i.e., SR_I, or supervisory command) and the reception of the corresponding message returned as an answer to that message (i.e.,

acknowledging message). Therefore, the receiver should not delay the response or acknowledging message returned to one of the above messages by more than a value T2, where T2 is a system parameter (see item b) below).

The PE shall not delay the response or acknowledging message returned to one of the above remote PE messages by more than a period T2.

b) SR Response Timeout (Timer T2)

The same value of the parameter T2 shall be made known and agreed to by the two PEs. The period of parameter T2 shall indicate the amount of time available at the PE before the acknowledging message shall be initiated in order to ensure its receipt by the remote PE, prior to timer T1 running out at the PEs (parameter T2 < timer T1).

The period of parameter T2 shall take into account the following timing factors:

1) the transmission time of the acknowledging message;

2) the propagation time over the access link;

3) the stated processing times at the PEs; and

4) the time to complete the transmission of the message(s) in the PE transmit queue that are neither displaceable nor modifiable in an orderly manner.

Given a value for timer T1 for the PEs, the value of parameter T2 shall be no larger than T1 minus 2 times the propagation time over the access data link, minus the message processing time at the PE, minus the message processing time at the remote PE, and minus the transmission time of the acknowledging message by the PE.

c) SR Poll Retries (N2)

The same value of the N2 system parameter shall be made known and agreed to by the two PEs.

The value of N2 shall indicate the maximum number of attempts made by the PE to complete the successful transmission of a message to the remote PE.

d) SR Window Size (k)

The same value of the k system parameter shall be made known and agreed to by the two PEs.

The value of k shall indicate the maximum number of sequentially numbered SR-I messages that the PEs may have outstanding (i.e., unacknowledged) at any given time. The value of k shall never exceed 32767 for modulo 32768 operation.

3. Timing Consideration

Correct Fibre Channel information exchange requires that the inherent latency between CE1 and CE2 (refer to Figure 1) be:

a) no more than one-half of the R_T_TOV (Receiver Transmitter Timeout Value, default value: 100 milliseconds, defined in [FC-FS]) of the attached devices for Primitive Sequences;

b) no more than one-half of the E_D_TOV (Error Detect Timeout Value, default value: 2 seconds, defined in [FC-FS]) of the attached devices for frames; and

c) within the R_A_TOV (Resource Allocation Timeout Value, default value: 10 seconds, defined in [FC-FS]) of the attached fabric(s), if any.

Requirement a) above, controlling the latency associated with FC Primitive Sequence transport is addressed by [FC-BB], stating that in case Primitive Sequences are received from the CE or the remote PE, while the device is unable to forward these Primitive Sequence due to backpressure indication, it shall flush its respective buffer (PSN-facing if the Primitive Sequences were received from the CE, CE-facing if they were received from the remote PE), and shall forward the Primitive Sequences.

Another case is when there is no specific backpressure indication, rather Primitive Sequences are being delayed due to retransmission of dropped frames. In this case also, the PE shall flush its PSN-facing buffer and shall forward the Primitive Sequences.

Requirements b) and c) above apply to the latency associated with transporting FC frames, and the system MUST comply with the lower of the two timeouts. The mechanism controlling this latency is described below for the PE1 --> PE2 direction. A duplicate mechanism MUST be used to control the PE2 --> PE1 direction.

PE2 (the receiver) maintain a timer Td, set to the minimum timeout set by requirements b) and c), with a safety margin to allow a deviation of the estimated RTT from the actual RTT. PE2 SHOULD set Td = 0.8 x (0.5 x MIN(E_D_TOV,R_A_TOV) - 2 x RTT), where RTT is an average value calculated as specified in Section 6.1.2, and the factor 0.8 provides safety margins for RTT fluctuations. In case the calculated Td < 2 x RTT for two consecutive calculation periods, the FC PW MUST be shut down (this avoids getting into conditions where Td expiration is too frequent).

To guarantee correct operation of this mechanism FC PW SHOULD NOT be used in environments where the RTT may have high variability, i.e., environments where the estimated RTT may not be off by more than 5% of MIN(E_D_TOV,R_A_TOV). If the FC PW is used in an environment where this limit is exceeded, the safety margins MUST be increased to encompass twice the expected maximum variability in the RTT.

Upon Td expiration PE2 declares WAN Down as defined in [FC-BB], send the Not Operational (NOS) Primitive Sequence to CE2, and flush its buffers.

The timer Td is started when either of the following two conditions occurs:

a) SR-RNR is sent by PE2 toward PE1, i.e., PE2 indicates backpressure, and stop PE1 from transmitting;

b) SR-SREJ is sent by PE2 toward PE1, requiring retransmission.

PE2 notes the sequence number (Nx) of the first frame to be received following the timer initiation (for SR_RNR this will be the next frame to be transmitted by PE1, i.e., the sequence number of the last frame received by PE2 before starting the timer plus 1. For SR_SREJ this frame will be the first sequence number in the retransmission list).

The timer continues counting and is initialized in one of two cases:

a) PE2 receives the frame with sequence number (Nx + k), where k is the Selective Retransmission window size;

b) PE2 identifies idle period (all sent frames were acknowledged, and the CE-facing buffer is empty).

4. Applicability Statement

The methods specified in this document MUST be applied for the transport of FC PW over uncontrolled (i.e., providing excess information rate for the service) networks, in order to provide reliable Fibre Channel transport and TCP-friendly behavior under network congestion.

5. Normative References

[FC-encap] Roth, M., et al, "Encapsulation Methods for Transport of Fibre Channel frames Over MPLS Networks", RFC TBD, to appear.
RFC Editor: Please contact authors to obtain the correct RFC number and date for the "to appear" in the above reference prior to publication.

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

[RFC3916] Xiao, X., et al, "Requirements for Pseudo Wire Emulation Edge-to-Edge (PWE3)", RFC 3916, September 2004.

[RFC3448] Floyd, S., et al, "TCP Friendly Rate Control (TFRC): Protocol Specification", draft-ietf-dccp-rfc3448bis-06, April 2008.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", RFC 4447, April 2006.

[RFC4447] Martini, L., et al, "Pseudowire Setup and Maintenance using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4385] Bryant, S., et al, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for use over an MPLS PSN", RFC 4385, February 2006.

[RFC4623] Malis, A., Townsley, M., "PWE3 Fragmentation and Reassembly", RFC 4623, August 2006.

[FC-BB] "Fibre Channel Backbone-4" (FC-BB-4), ANSI INCITS 419:2008, to appear.
RFC Editor: Please contact authors to obtain the correct date for the "to appear" in the above reference prior to publication.

[FC-SW] "Fibre Channel - Switch Fabric - 4" (FC-SW-4), ANSI INCITS 418:2006, April 2006.

[FC-FS] "Fibre Channel - Framing and Signaling - 2" (FC-FS-2), ANSI INCITS 424:2007, February 2007.

[BCP14] Bradner, S., "Key words for use in RFCs to Indicate requirement Levels", BCP 14, RFC 2119, March 1997.

6. Informative references

[RFC3668] Bradner, S., "Intellectual Property Rights in IETF Technology", RFC 3668, February 2004.

[RFC3821] M. Rajogopal, E. Rodriguez, "Fibre Channel over TCP/IP (FCIP)", RFC 3821, July 2004.

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

[RFC2581] Allman, M., et al, "TCP Congestion Control", RFC 2581, April 1999.

7. Author's Addresses

Moran Roth
Corrigent Systems
101, Metro Drive
San Jose, CA 95110
Phone: +1-408-392-9292
Email: moranr@corrigent.com

Ronen Solomon
Corrigent Systems
126, Yigal Alon st.
Tel Aviv, ISRAEL
Phone: +972-3-6945316
Email: ronens@corrigent.com

Munefumi Tsurusawa
KDDI R&D Laboratories Inc.
Ohara 2-1-15, Fujimino-shi,

Saitama, Japan
Phone: +81-49-278-7828

8. Contributing Author Information

David Zelig
Corrigent Systems
126, Yigal Alon st.
Tel Aviv, ISRAEL
Phone: +972-3-6945273
Email: davidz@corrigent.com

Leon Bruckman
Corrigent Systems
126, Yigal Alon st.
Tel Aviv, ISRAEL
Phone: +972-3-6945694
Email: leonb@corrigent.com

Luis Aguirre-Torres
Corrigent Systems
101 Metro Drive
San Jose, CA 95110
Phone: +1-408-392-9292
Email: Luis@corrigent.com