Document: T10/02-323r1 Date: September 8, 2002

To: T10 Committee Membership

From: Edward A. Gardner, Basil Networks

Subject: SAS Data Corruption Problem

Consider the situation where a transmission error occurs when an initiator is sending write data to a target. For the sake of argument assume that an SOF is corrupted, so that a frame is completely lost. However the result is substantially similar with other errors, e.g. coding violations or a CRC error.

Since data transfers are not interlocked, the initiator may not become aware of the problem until many frames after the fact. The target will not be aware of the problem until after it times out the data's arrival or receives a TASK ABORT. Meanwhile the data frame in error is dropped from the middle of the data stream received at the target.

Many or all present day disk drives stream data from SCSI or Fibre Channel onto the media. If SAS drives were to use similar designs, the corrupted data stream will typically have already been written to the wrong location on the media by the time the error is discovered. This creates a substantial time window where a power failure, system crash or similar problem will leave the corrupted data on the media permanently. That is unacceptable for most enterprise applications.

Fibre Channel avoids this problem by defining two header fields for error checking. The one most commonly implemented is SEQ_CNT, a sequence count used to check that all data frames arrived. Less often used is RELATIVE OFFSET, which allows the recipient to check that all data bytes arrived.

Parallel SCSI avoids this problem by providing a low overhead method to insert a CRC into the data stream. The target can choose the location of the CRCs, e.g. a CRC for every disk block. The overhead for inserting a CRC is negligible.

In contrast, the only way a SAS target can validate a write data transfer is to break it up into many smaller transfers, each with its own XFER_RDY. The bus round-trip required to issue a new XFER_RDY essentially guarantees that the connection will be closed and a new connection will have to be opened. That is not a low overhead operation.

The requirement is for a low overhead way for a target to validate a write data transfer at granularity comparable to the disk block size or frame size. One approach would be to define some way to stream multiple XFER_RDYs for a single command without interlocking the connection. That was rejected as being too complex and disruptive to SAS as it has been defined. Instead this proposal suggests a per connection counter that counts frames or information units.

The basic concept is to have a counter that is initialized whenever a connection is opened. The counter increments for every frame or information unit sent on the connection. The counter is included in the information unit header. If the received value does not match the expected value, it implies that one or more frames have been lost. The recipient transport would discard that information unit and all subsequent information units received on the connection, similar to discarding frames with incorrect address hashes. Note that if a frame is lost, the sender will detect an ACK/NAK timeout when it next interlocks the connection. This guarantees that all frames that are discarded are for the same (non-interlocked) data transfer as the frame that was lost. Note also that the existing limit of at most 255 outstanding R_RDYs ensures that an eight bit counter is sufficient.

In discussing this concept with other vendors, they suggested that the connection frame counter should only be included in DATA information units. This proposal incorporates that suggestion by defining a flag to indicate the presence or absence of the counter.

Clause 9.2.1, SSP frame format, and table 59. Define bit 2 of byte 11 in table 59 to be CFCE. The corresponding bit in Fibre Channel headers is defined as "Reserved for Exchange reassembly", for use in exchanges that span many source or destination ports. Define byte 15 in table 59 to be

CONNECTION FRAME COUNTER. The corresponding byte in Fibre Channel headers contains the low byte of the sequence count. Add the following definitions to the text following table 59:

> The connection frame counter enable (CFCE) bit is set to one to indicate that the CONNECTION FRAME COUNTER field is valid. The CFCE bit is set to zero to indicate that the CONNECTION FRAME COUNTER field is reserved.
>
> The CONNECTION FRAME COUNTER field counts frames with the CFCE bit set to one that are sent on a single connection. SAS devices shall set the CONNECTION FRAME COUNTER field to one in the first frame with the CFCE bit set to one that is sent on a connection. The contents of the CONNECTION FRAME COUNTER field shall increment in subsequent frames with the CFCE bit set to one that are sent on the same connection. The counter value FFh shall wrap around to 00h when incremented.
>
> The recipient may compare the CONNECTION FRAME COUNTER field to the expected value. If it is checked and the CONNECTION FRAME COUNTER field does not match the expected value, that frame and all subsequent frames received on the same connection shall be discarded by the transport layer.