Document:	T10/01-325r0	Date:	November 2, 2001
To:	T10 Committee Membership		
From:	Edward A. Gardner, Ophidian Designs		
Subject:	Ophidian Designs' comments on SRP-r10		

OD 1. Page 13, lines 5-7, multiple RDMA writes on the same channel store data in order.

Some RDMA communication services (e.g. iWARP) are unable to ensure strict ordering of overlapping RDMA Write operations during normal operation. While methods are available to ensure strict ordering, invoking them for all RDMA Writes would severely affect performance.

SAM-2 does not specify the result of multiple commands to overlapping buffers in most cases. It is unclear whether it specifies the result in any situation (see T10/01-309). Overlapping transfers, also called data overlay, within a single command is unusual enough that some SCSI protocols routinely prohibit it.

This requirement should be removed from SRP. It should be replaced with a statement that overlapping transfers may yield unpredictable results unless the RDMA client (SRP) takes special precautions. The nature of said special precautions, if any, are RDMA communication service specific. A section should be added to clause 5 discussing data overlay to specify that SRP target ports shall take said special precautions whenever data overlay occurs within a command.

OD 2. Page 13, line 13, RDMA read operations may complete in any order.

While this states that RDMA Read operations may complete in any order, it is not clear what data they are required to return. See the first example in T10/01-309r0.

If T10/01-309r0 is accepted, this should be clarified to indicate that the data returned by RDMA Read operations need not reflect concurrent RDMA Writes that precede the RDMA Read.

If T10/01-309r0 is not accepted, this should be changed to require that RDMA Reads and RDMA Writes to overlapping locations are strictly ordered for memory access.

OD 3. Page 14, RDMA channel disconnection
Page 15, Multiple independent RDMA channel operation
Page 16, lines 9 and 10 (list items b and c)
Page 27, SRP_LOGIN_RSP response
Page 30, SRP_I_LOGOUT request
Page 31, SRP_T_LOGOUT request

One of the characteristics of a network or fabric communication service is that errors affecting a channel can rarely be reported using that channel. In the context of SRP, many errors that disconnect an RDMA channel will be reported to one consumer but not the other. The consumer receiving the report cannot use the same RDMA channel to notify the other consumer, as the channel is no longer operational.

It is nonetheless useful for both consumers to know that an RDMA channel has failed. When using multiple independent RDMA channels, the consumers could use one of the other channels to report a channel failure. SRP should be extended to support this. This should be mandatory behavior whenever multiple channels are used between the same SRP initiator port and the same SRP target port. The following paragraphs summarize the changes to SRP to accomplish this.

The SRP_LOGIN_RSP response should return a channel handle. The channel handle shall be non-zero and unique among all channels in use on the same I_T nexus. Zero is valid if and only if

the SRP target port only supports one channel per nexus. The channel handle should be a 16-bit field in bytes 28 and 29.

The SRP_I_LOGOUT and SRP_T_LOGOUT requests should specify an optional channel handle. The channel handle should be a 16-bit field in bytes 2 and 3. If the channel handle is zero, it specifies that the channel on which the request was sent is being logged out; no response is generated. This is identical to the behavior currently specified by SRP. If the channel handle is non-zero then the specified channel is being logged out. A response is generated to confirm the logout and to indicate that all outstanding requests on that channel have been discarded. Targets shall not use of a non-zero channel handle that specifies the channel on which the SRP_T_LOGOUT request is sent. Use of a non-zero channel handle that specifies the channel on which the SRP_I_LOGOUT request is sent results in target specific behavior.

Extend the discussion of RDMA channel disconnection (page 14) and multiple independent RDMA channel operation (page 15) to require that targets report disconnection using an alternate channel if one is available.

Amend the list of requests that do not have responses on page 16 to say that SRP_I_LOGOUT and SRP_T_LOGOUT do not have responses when the channel handle is zero, but do have responses when the channel handle is non-zero.

Note that this change cannot be straightforwardly added in an SRP-2. An initiator or target that ignores the channel handle field (because it was reserved in SRP) would logout the wrong channel.

OD 4. Page 56, tables B.2 and B.3

SRP port identifiers for Infiniband are 128-bit identifiers with an embedded GUID (EUI-64). Infiniband GIDs are 128-bit identifiers with an embedded GUID (EUI-64). Unfortunately they are formatted incompatibly. Annex B specifies that the EUI-64 occupies the most significant bytes of an SRP port identifier while the EUI-64 occupies the least significant bytes of an InfiniBand GID or IPv6 formatted address. The bytes not occupied by the EUI-64 are also different.

Having conflicting formats of otherwise equivalent identifiers is guaranteed to lead to interoperability problems. Various people have stated (in SRP working groups) that they expect to identify SRP targets using IPv6 formatted identifiers. SRP should be changed to satisfy this.

A new informative annex should be added recommending that SRP port identifiers adhere to IPv6 address formatting conventions and use one of the three forms listed below. Annex B should require that InfiniBand SRP port identifiers be one of the three forms listed below.

1. The Link-Local prefix (FE80h:0:0:0::/64) concatenated with an EUI-64.
2. The Site-Local prefix (FEC0h:0:0::/48) concatenated with 16-bit locally administered value concatenated with an EUI-64.
3. Any value configured manually or by a system management agent.

OD 5. Pages 4 and 5, glossary terms, and their use throughout the document.
Clause 4

When SRP was proposed and for much of its development no satisfactory glossary of RDMA terms was available. Available external documents used definitions specific to particular implementations. That has recently changed. See the message titled “iWARP Glossary” posted to the yahoo RDMA reflector on September 27, 2001 by Jim Wendt. It would be beneficial if SRP were changed to use the same terms and definitions.

OD 6. Page 11 lines 20-22, normal and solicited message reception.

This feature is described in the RDMA communication service model, yet not used by SRP. Interrupt mitigation is important in high end systems. Therefore this should be supported by SRP information units. A description of how to do so follows.

Define a bit to be included in all SRP information units. Recommend this be bit 0 of byte 1 and called NOTURG (notification urgency or not urgent, take your pick).

In initiator to target requests, NOTURG specifies the notification urgency for the response. The initiator may set it to any value.

In target to initiator responses, NOTURG specifies the notification urgency. The target shall copy it from the request.

In target to initiator requests, NOTURG shall be zero. Specify this individually in each request, not as a general rule, so that it may be changed for future requests.

In initiator to target responses, the target shall ignore NOTURG.

In Annex B, specify that the target shall send information units with solicited event notification enabled if NOTURG is zero. The target shall send information units with solicited event notification disabled if NOTURG is one. The initiator shall ignore NOTURG and send all information units with solicited event notification enabled.

OD 7. RDMA communication service specific opcode

SRP currently requires RDMA Read support for practical operation. However there are RDMA communication services that do not support an RDMA Read. So-called Unreliable Connections on InfiniBand are on example. Note that these have the same reliability characteristics as most existing SCSI protocols (e.g. FCP). Various people have suggested that they would be the most natural service for storage access, except for the lack of RDMA Read. Some VI Architecture implementations also lack RDMA Read.

It is straightforward to emulate an RDMA Read. The target sends a request to the initiator identifying the data to be read. The initiator responds with an RDMA Write supplying the required data, then a response to indicate completion. All that is missing is SRP opcodes that could be used for the request and response.

This is one example of a need for an RDMA communication service specific operation. Others might be required in the future for as yet unanticipated reasons. The purpose of defining this now is to describe proper behavior for an initiator that does not recognize the request.

The following could be defined using a new pair of opcodes or as an extension to the existing SRP_CRED_REQ and SRP_CRED_RSP. I don't particularly care which is used.

Define a target to initiator request. It is formatted identically to SRP_CRED_REQ with the addition of an action code field and action code specific parameters. I recommend a 16-bit action code field. The action code specific parameters may be any length (including zero) provided the total request length is within the limit agreed to during login.

Define the corresponding initiator to target response. It is formatted identically to SRP_CRED_RSP with the addition of an action code, an action response code and action code specific parameters. The action code is an echo of the value in the request (could be omitted). The action response code indicates the outcome of the action. Define value zero to designate the action is not supported, all other values reserved. The action code specific parameters may be any length (including zero) provided the total request length is within the limit agreed to during login. If the response code indicates the action was not supported, the action code specific parameters shall be zero length.

OD 8. Page 18, lines 13-37 and elsewhere, data buffer format code and count values.

The combination of a data buffer format code and a data buffer format count is awkward. Their interpretation is interdependent. We really have a single 12-bit field. It would simplify the description (and probably the implementation) if we had a single encoded data buffer format field. The following is a suggested way to encode an 8-bit data buffer format code:

00h	NO DATA BUFFER DESCRIPTOR PRESENT
01h	DIRECT DATA BUFFER DESCRIPTOR
02h – 0Fh	Reserved
1xh	INDIRECT DATA BUFFER DESCRIPTOR
10h	INDIRECT DATA BUFFER DESCRIPTOR WITH NO PARTIAL MEMORY DESCRIPTOR LIST
11h	INDIRECT DATA BUFFER DESCRIPTOR WITH 1 ENTRY PARTIAL MEMORY DESCRIPTOR LIST
12h	INDIRECT DATA BUFFER DESCRIPTOR WITH 2 ENTRY PARTIAL MEMORY DESCRIPTOR LIST
etc.	
1Fh	INDIRECT DATA BUFFER DESCRIPTOR WITH 15 ENTRY PARTIAL MEMORY DESCRIPTOR LIST
20h to FFh	Reserved

These values would occupy bytes 6 and 7 of SRP_CMD, byte 5 would be reserved.