Document: T10/01-177r2 Date: 4 August 2001
To: T10 Committee Membership
From: Edward A. Gardner, Ophidian Designs
Subject: SRP Model for RDMA Communication Services

Attached is the RDMA communication service model. While we've discussed this as replacing subclause 4.1, I cannot rationalize this as being part of clause 4, which is titled "Structure and concepts". I'm planning on making this a separate clause unless someone has a better suggestion.

Revision 0: original version.

Revision 1: changes from June 19-20 SRP working group meeting in Redmond.

Revision 2: changes from July 19-20 SRP working group meeting in Colorado Springs.

Glossary entries to be added:

3.4.1 accept data: Application protocol data communicated from a server consumer to the client consumer when a new RDMA channel is accepted (see 4.2). SRP uses accept data to communicate the SRP_LOGIN_RSP response.

3.4.2 channel attributes: Information provided during RDMA channel establishment that identifies the type and characteristics of the desired RDMA channel (see 4.2). The format and interpretation of channel attributes are RDMA communication service specific.

3.4.3 consumer: An entity that communicates with other consumers using an RDMA communication service (see 4.1). Within SRP, a consumer is either a target port or an initiator port.

3.4.4 login data: Application protocol data communicated from a client consumer to a server agent or consumer during RDMA channel establishment (see 4.2). SRP uses login data to communicate the SRP_LOGIN_REQ request.

3.4.5 message: A communication sent by one consumer to another using an RDMA channel (see 4.3).

3.4.6 RDMA channel: A communication path between two consumers of an RDMA communication service (see 4.1).

3.4.7 RDMA communication service: A transport protocol or service that provides messages and RDMA operations between pairs of consumers (see clause 4).

3.4.8 RDMA operation: Either an RDMA Read operation or an RDMA Write operation.

3.4.9 RDMA Read operation: An operation by which a requesting consumer may fetch data from memory registered by the other consumer associated with an RDMA channel (see 4.4).

3.4.10 RDMA Write operation: An operation by which a requesting consumer may store data into memory registered by the other consumer associated with an RDMA channel (see 4.4).

3.4.11 reject data: Application protocol data communicated from a server agent or consumer to the client consumer when a new RDMA channel is rejected (see 4.2). SRP uses reject data to communicate the SRP_LOGIN_REJ response.

3.4.12 server address: Information provided to an RDMA communication service by a client consumer that identifies a server with which to establish an RDMA channel (see 4.2). The format and interpretation of channel attributes are RDMA communication service specific.

4 RDMA communication service model

4.1 Overview

SRP is designed to operate using an RDMA communication service. An RDMA communication service provides communication between pairs of consumers by means of messages for control information and RDMA operations for data transfers.

Figure 1 shows an example system that uses an RDMA communication service. Communication is provided by RDMA channels, shown as solid lines in the figure. An RDMA channel provides communication between two consumers. A single pair of consumers may communicate using many RDMA channels if sufficient resources are available. Some environments may use multiple special purpose RDMA channels between a single pair of consumers (e.g., a pair of consumers could use certain RDMA channels for messages and other RDMA channels for RDMA operations).

The RDMA communication service in figure 1 is comprised of adapters and other unspecified components (e.g. wires, fabric switches). The actual components of an RDMA communication service are implementation specific. Components such as adapters may or may not be present.

This clause describes various functions that may be provided by an RDMA communication service. A specific implementation of an RDMA communication service may or may not provide these exact functions. Any of these functions may be mapped to a sequence of several functions provided by the RDMA communication service. Annex **TBD** describes the mapping of these functions to those provided by Infiniband™.

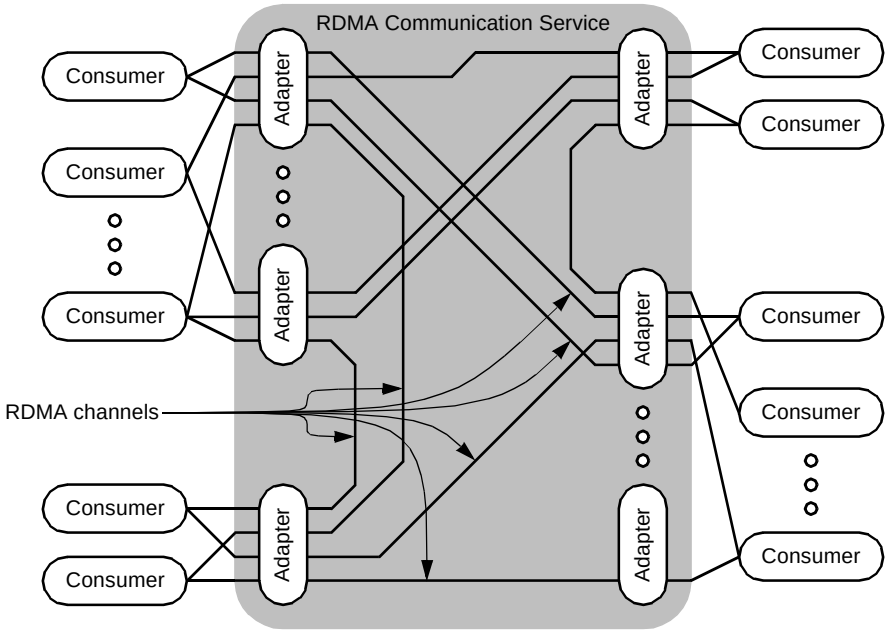

Figure 1 - RDMA communication service example

4.2 RDMA Channels

An RDMA channel provides communication between a pair of consumers using messages, RDMA operations or both. An RDMA channel is a dynamic connection, established and disconnected upon request. Establishing an RDMA channel may require obtaining resources to support the channel, either within the channel's consumers or within the RDMA communication service or both. Multiple channels may be established between the same pair of consumers if sufficient resources are available. The resources associated with a channel may be released after the channel is disconnected.

Figure 2 shows an example of the process by which a new RDMA channel is established. A client consumer requests that the RDMA communication service establish a new RDMA channel. The request is directed to a server and, if successful, resolved to a server consumer. The resulting RDMA channel provides communication between the client consumer and the server consumer.

A client consumer provides a server address to identify the server with which to establish an RDMA channel. The format and interpretation of a server address are specific to the RDMA communication service. A server address may specify an individual server consumer or multiple server consumers. For example, a server address may identify an adapter as shown in figure 1, specifying all consumers that implement a specific application protocol and are accessible through that adapter.

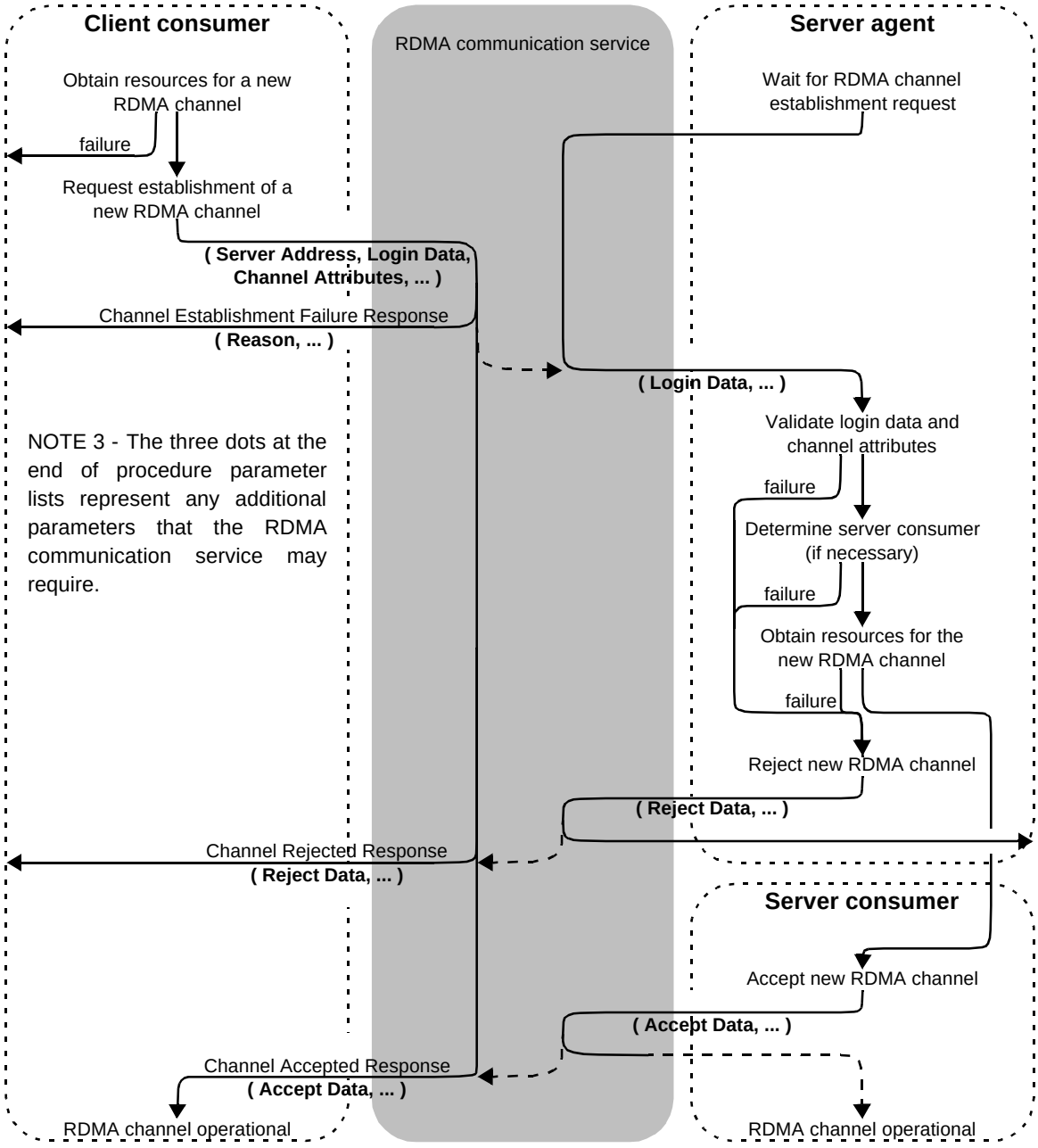

Figure 2 - Example RDMA channel establishment

In the example shown in figure 2 the recipient of an RDMA channel establishment request, identified by a server address, is called a server agent. The server agent may use application protocol and server specific knowledge to determine whether an RDMA channel establishment request may be accepted and the server consumer to which it shall be assigned. The actual actions required of a server agent and their order are specific to the RDMA communication service and server. A server agent may or may not be a distinct entity. Some or all of the actions that figure 2 shows being performed by a server agent may be performed by a server consumer or by the RDMA communication service.

An RDMA communication service may require that the client consumer obtain resources before requesting that a new RDMA channel be established. After obtaining those resources, the client consumer may request that the RDMA communication service establish a new RDMA channel. The request includes a server address, login data, channel attributes, and any other parameters required by the RDMA communication service. With SRP the client consumer shall be an SRP initiator port, the server address shall identify one or more SRP target ports, and the login data shall contain an SRP_LOGIN_REQ request.

The RDMA communication service returns one of three responses to the client consumer for a new RDMA channel establishment request.

The RDMA communication service may return a channel establishment failure response. A channel establishment failure response indicates that the new RDMA channel could not be established for some reason internal to the RDMA communication service. A channel establishment failure response may return an RDMA communication service specific reason code to identify the cause of the failure as well as other RDMA communication service specific data.

The RDMA communication service may return a channel rejected response. A channel rejected response indicates that the request was communicated to the server but rejected by the server agent or consumer. A channel rejected response returns reject data, which is application protocol data provided by the server agent or consumer. Reject data may include a reason for rejecting the request or other application protocol information. A channel rejected response may also return RDMA communication service specific data.

The RDMA communication service may return a channel accepted response. A channel accepted response indicates that the new RDMA channel has been successfully established and is operational. The client consumer may use the channel in accordance with the application protocol. A channel accepted response returns accept data, which is application protocol data provided by the server agent or consumer. Accept data may include application protocol parameters governing how the new RDMA channel should be used. A channel accepted response may also return RDMA communication service specific data.

An RDMA communication service may require that a server agent register itself prior to receiving connection establishment requests. In figure 2 this is shown as a registration request (e.g., subroutine call) that returns control to the server agent when a new RDMA channel establishment request is received. The way that a server agent registers with an RDMA communication service is specific to that service or the server.

New RDMA channel establishment requests that are acceptable to the RDMA communication service shall be passed to the server agent. The server agent is provided the login data from the client consumer's request in addition to RDMA communication service specific data.

The server agent determines whether the new RDMA channel establishment request may be accepted and, if necessary, determines the server consumer to be associated with the new RDMA channel. If the request is not accepted the server agent or consumer instructs the RDMA communication service to reject the new RDMA channel. The server agent or consumer provides reject data, which is application protocol data to be passed to the client consumer, as well as any RDMA communication service or server specific data that may be required. With SRP the reject data shall contain an SRP_LOGIN_REJ response.

If the new RDMA channel establishment request is accepted, the server agent or consumer instructs the RDMA communication service to accept the new RDMA channel. The server agent or consumer provides accept data,

which is application protocol data to be passed to the client consumer, as well as any RDMA communication service or server specific data that may be required. With SRP the accept data shall contain an SRP_LOGIN_RSP response.

An RDMA channel may be disconnected by a request from either of the channel's consumers or from the RDMA communication service itself. The consumers may each be notified that the RDMA channel has been disconnected, allowing them to recover any resources associated with the RDMA channel. The time to deliver such a notification may vary depending upon the RDMA communication service, the consumer being notifed, and the specific circumstances of the disconnection request.

A disconnect request causes an RDMA channel to become non-operational. Operations in progress on an RDMA channel at the time of a disconnect request and operations requested subsequent to a disconnect request may or may not complete.

4.3 Messages

An RDMA channel may allow its consumers to exchange messages. A message is sent by one consumer associated with an RDMA channel (the sending consumer) to the other consumer associated with the RDMA channel (the receiving consumer). A message contains a payload of some number of data bytes. An RDMA communication service may provide normal and solicited message reception notification, which may be used to distinguish between more urgent and less urgent messages.

A sending consumer requests that a message be sent by providing the following to an RDMA communication service:

a) the message's payload length;
b) the message's payload data;
c) the RDMA channel to use; and
d) whether to use normal or solicited message reception notification.

The RDMA communication service attempts to deliver the message to the other consumer associated with the specified RDMA channel (the receiving consumer). If delivery succeeds, the RDMA communication service notifies the receiving consumer that a message has been received, providing the message's length, payload, and the channel on which the message was received. The RDMA communication service may also provide an indication of whether the sending consumer specified normal or solicited message reception notification.

An RDMA communication service may require that receiving consumers provide message receive buffers to RDMA channels before messages are sent to them, and that the provided message receive buffers be large enough to hold any messages that arrive. Sending a message on an RDMA channel when no receive buffer has been provided, or when the provided receive buffer is too small for the message, may result in behavior that is not specified by this standard.

> NOTE 3 - Such behavior may include (but is not limited to) disconnecting the RDMA channel, discarding or truncating the message, or delaying delivery of the message until a suitable message receive buffer becomes available. The RDMA communication service may or may not provide an error indication.

An RDMA communication service may or may not provide a way for a sending consumer to determine whether a message has been delivered to the receiving consumer.

4.4 RDMA operations

An RDMA channel may provide RDMA Write operations, RDMA Read operations, or both between its consumers.

A consumer may allow RDMA access by registering some or all of its memory with an RDMA communication service. The RDMA communication service returns a memory handle to identify the registered memory. The consumer may specify that the memory handle is usable for memory access on only a specified RDMA channel

or on a group of RDMA channels. The consumer may impose other access restrictions allowed by the RDMA communication service as well (e.g. read-only acess).

A consumer that has registered memory and obtained a memory handle may communicate the memory handle to another consumer. This may be done using an application protocol contained in message payloads. The other consumer may then use the memory handle to request RDMA operations that access the memory registered by the first consumer.

The registered memory identified by a memory handle is represented as a memory address space. Accessible locations are identified by addresses. An RDMA communication service is not required to provide a way to determine, from a message handle, which memory locations are accessible, the number of locations that are accessible, or the type of access allowed. Such information may be communicated by an application protocol.

An RDMA Write operation allows a requesting consumer to store data into memory registered by another consumer. A requesting consumer provides the following to an RDMA communication service when it requests an RDMA Write operation:

a) An RDMA channel to use for the operation;
b) A memory handle that is usable for access on that RDMA channel;
c) A range of addresses within the memory address space identified by the memory handle; and
d) Data to be written into the specified range of addresses.

An RDMA communication service is not required to provide a way for a requesting consumer to determine whether the data has been written into the specified range of addresses in registered memory. An RDMA communication service is not required to provide a way for the consumer that registered the memory to determine whether an RDMA Write operation is in progress or has completed.

An RDMA Read operation allows a requesting consumer to fetch data from memory registered by another consumer. A requesting consumer provides the following to an RDMA communication service when it requests an RDMA Read operation:

a) An RDMA channel to use for the operation;
b) A memory handle that is usable for access on that RDMA channel;
c) A range of addresses within the memory address space identified by the memory handle; and
d) A buffer into which to place the data read from the specified range of addresses.

The RDMA communication service notifies the requesting consumer after data has been successfully obtained from the specified range of addresses and placed in the requestor's buffer. An RDMA communication service is not required to provide a way for the consumer that registered the memory to determine whether an RDMA Read operation is in progress or has completed.

4.5 Ordering and Reliability

SRP operates using an RDMA communication service having the characteristics described in this subclause. Use of SRP with an RDMA communication service having different characteristics is outside the scope of this standard.

An RDMA communication service shall deliver each message sent on an RDMA channel to the receiving consumer or else disconnect the RDMA channel. Each delivered message shall be delivered to the receiving consumer exactly once; the RDMA communication service shall discard any duplicates that may result from retransmission or other mechanisms. Each delivered message shall be delivered to the receiving consumer complete and error-free.

Messages sent by the same consumer on the same RDMA channel shall be delivered to the receiving consumer in the order they were sent. The data for all RDMA Write operations requested on an RDMA channel by a consumer prior to that same consumer sending a message on the same RDMA channel shall be available to the receiving consumer (e.g. stored into registered memory) before the message is delivered to the receiving

consumer. If multiple RDMA Write operations requested on an RDMA channel by a consumer store data into the same registered memory location, the location's resulting contents shall be the data stored by the last RDMA Write operation.

Messages sent on different RDMA channels or by different consumers may be delivered in any order. The data for RDMA Write operations may be stored into registered memory in any order relative to the delivery of messages sent on other RDMA channels or by other consumers. RDMA Write operations requested on different RDMA channels may store data into the same registered memory location in any order.

RDMA Read operations may complete in any order.

If an RDMA communication service is unable to satisfy these requirements on an RDMA channel, it shall disconnect the RDMA channel.