INTERNET DRAFT
draft-hufferd-ips-iser-sctp-ib-00.txt

John Hufferd
Mike Ko
IBM Corporation

Yaron Haviv
Voltaire Ltd

December, 2004

Expires: June, 2005

Generalization of iSER for SCTP, Infiniband and other Network Protocols

Status of this Memo

Abstract

The iSCSI Extensions for RDMA document [iSER] currently specifies the RDMA data transfer capability for [iSCSI] over iWARP/TCP. This document generalizes the iSER document to permit it to be used with other RDMA capable protocols such as iWARP/SCTP, InfiniBand, etc. It also describes what should be defined in the InfiniBand Trade Association and what are appropriate for IETF.

Table of Contents

1 Motivation.....5
2 Overall generalizations needed within the iSER specification6
2.1 Generalization of Definitions.....6
2.1.1 The iWARP term.....6
2.1.2 The RNIC term.....7
2.1.3 Steering Tag (STag).....7
2.1.4 Inbound RDMA Read Queue Depth (IRD) & Outbound RDMA Read Queue Depth (ORD).....7
2.1.5 RDMA Protocol (RDMAP).....8
2.1.6 RDMAP Stream.....8
2.1.7 RDMAP Message.....8
2.2 The following should be placed/updated in Acronym Section...9
2.3 Connection Establishment, Login, and Transition to iSER.....9
2.4 Security considerations.....9
2.5 Adjustments to Appendix.....10
2.6 Add Appendix B.....10
3 Architectural discussion of iSER over InfiniBand.....11
3.1 The Host side of the InfiniBand iSCSI & iSER connections...11
3.2 The Storage side of iSCSI & iSER mixed network environment.13
3.3 Discovery processes for an InfiniBand Host.....14
3.4 IBTA Connection specifications.....15
4 Preferred changes to iSCSI Discovery Data.....16
5 The Appropriate Standardization Organizations for iSER/IB..17
5.1 The IETF ips Working Group Standardization.....17
5.2 The InfiniBand Trade Association (IBTA) Standardization....17
5.3 IBTA and optional InfiniBand Features.....17
5.4 ServiceID and IP Port number.....17
6 IANA Considerations.....18
7 References.....19
7.1 Informative References.....19
8 Appendix.....20
8.1 Additional detailed [iSER] document modification.....20
8.1.1 Adjustment to Section 2.1 Motivation.....20
8.1.2 Adjustment to Section 2.2 Architectural Goals.....20
8.1.3 Adjustment to Section 2.3 Protocol Overview.....20
8.1.4 Adjustment to Section 2.4 RDMA services and iSER.....21
8.1.5 Adjustment to Section 2.7 iSCSI/iSER Layering.....21
8.1.6 Generalization of Other iSER Sections.....22
8.1.7 Adjustments to 13.2 Informational References.....25
9 Author’s Address.....27
10 Acknowledgments.....28
11 Full Copyright Statement.....29

Table of Figures

Figure 2 iSCSI, and iSER on IB.....12
Figure 3 Storage Controller with iSCSI, iWARP/TCP, iWARP/SCTP, and IB connections.....14
Figure 1 Generic example of iSCSI/iSER Layering in Full Feature Mode 22

1 Motivation

Currently the work to define iSCSI extensions for RDMA [iSER] only considers using the iWARP protocol suite over the TCP layer. While this objective meets the short term requirement since iSCSI is defined only for TCP, there is a huge benefit to generalize a standardized [iSER] so that it can be used with other types of RDMA capable layers now and in the future, including the following:

- iWARP over SCTP
- InfiniBand (with reliable connections, RC)

The interest in using [iSER] for InfiniBand is based on exploiting the iSCSI protocol features and its discovery and management protocol instead of using the SCSI RDMA Protocol (SRP) which lacks the management and discovery support. Furthermore, with an iSCSI based protocol, the Storage Professional and/or Administrator only needs to understand and support a single basic protocol, which has similar implementations across a suite of different network types (iWARP/TCP, iWARP/SCTP, InfiniBand, etc.).

It was to enable this vision and desire for a single Storage Protocol that the proposed generalizations to [iSER] were created.

2 Overall generalizations needed within the iSER specification

This section will specify changes/adjustments that should be considered to the wordage in the iSER document to make it more general. The goal of these changes is not to modify the basic operation of iSCSI/iSER when operating on the TCP version of iWARP, but to change/adjust the wordage in such a way that iSCSI/iSER can be layered over a different RDMA-capable protocol layer, such as an SCTP version of iWARP, or InfiniBand.

The details of many of the suggested changes can be found in the Appendix of this document.

2.1 Generalization of Definitions

It is required that some of the terminology be clarified as to applicability of the terms to the actual LLP used.

2.1.1 The iWARP term

As currently defined, the iWARP term has a strong TCP centric bias. We will introduce a new, more generic term, known as RDMA-Capable Protocol (RCP) to denote the protocol layer that provides the RDMA functionality for iSER. The following term will be added to the Definition section:

RDMA-Capable Protocol – The protocol or protocol suite that provides the RDMA functionality, e.g., iWARP, Infiniband, etc.

With these new definitions, the "iWARP" term will be generalized as follows:

1. Whenever the term "iWARP protocol suite" occurs in the iSER draft, it will be replaced by "RDMA-Capable Protocol". In addition, the phrase "such as the iWARP protocol suite" will be added where necessary.

2. Whenever the term "iWARP layer" occurs in the iSER draft, it will be replaced by "RDMA-capable protocol layer". In addition, the phrase "such as the iWARP Layer" will be added where necessary.

3. Whenever the term "iWARP" is used as an adjective in other context, it will be replaced with just RDMA, or "RDMA-Capable", whichever is appropriate. E.g., "iWARP functionality" will be replaced with "RDMA functionality".

4. Whenever the term "iWARP" is used as a shorthand for the iWARP protocol suite, it will be replaced by "RDMA-capable protocol".

5. Whenever iWARP is used as a specific implementation example intended for TCP only, such as "iWARP Message Format" in the appendix, it will be changed to iWARP/TCP.

2.1.2 The RNIC term

The term "RNIC" has been generally accepted by the industry to mean an RDMA-enabled Network Interface Controller for the IP world. So to generalize iSER for any RDMA-capable protocol layer, we will introduce a new term known as RDMA-Capable Controller, defined as follows:

. RDMA-Capable Controller – A network I/O adapter or embedded controller with RDMA functionality. E.g., for TCP/IP, this can be an RNIC, and for Infiniband, this could be a HCA (Host Channel Adapter) or TCA (Target Channel Adapter).

Within the body of the iSER document the term RDMA-Capable Controller will be used whenever the intention is to talk about a general controller that provides RDMA functionality. In addition, the clause "such as an RNIC" will be added as necessary.

Within the body of the iSER document, the term RNIC is left unchanged if it specifically or implicitly refers to TCP/IP.

2.1.3 Steering Tag (STag)

The Steering Tag (STag) term needs to have its definition extended so that it applies to both a Tag for a Remote Buffer, and the Tag for a Local Buffer. The following should be considered as a replacement for the existing one in the definition section.

Steering Tag (STag) - An identifier of a Tagged Buffer on a Node (Local or Remote) as defined in [RDMAP] and [DDP]. For other RDMA-Capable protocol layer, the Steering Tag may be known by different names. For example, for Infiniband, a Remote STag is known as an R-Key, and a Local STag is known as an L-Key.

2.1.4 Inbound RDMA Read Queue Depth (IRD) & Outbound RDMA Read Queue Depth (ORD)

To generalize on the terms Inbound RDMA Read Queue Depth (IRD) and the Outbound RDMA Read Queue Depth (ORD) for other RDMA-Capable

protocol layers, the following should be added to the definition for IRD: "For other RDMA-Capable protocol layer, the term "IRD" may be known by a different name. For example, for Infiniband, the equivalent for IRD is the Responder Resources". For ORD, the following should be added: "For other RDMA-Capable Protocol Layer, the term "ORD" may be known by a different name. For example, for Infiniband, the equivalent for ORD is the Initiator Depth."

2.1.5 RDMA Protocol (RDMAP)

In the body of the document the term "RDMA-Capable Protocol", or "RCP" should be used whenever any RDMA wire protocol or RDMA protocol stack is applicable. Only when the document intends to explicitly address a specific wire protocol would the term [RDMAP] be used.

2.1.6 RDMAP Stream

The following should be considered for inclusion in the definition section to replace "RDMAP Stream":

- RCP Stream - A single bidirectional association between the peer RDMA-capable protocol layers on two Nodes over a single transport-level stream. For TCP or SCTP, an RCP Stream is also known as an RDMAP Stream. For iSER/TCP, the association is created when the connection transitions to iSER-assisted mode following a successful Login Phase during which iSER support is negotiated.

In the body of the document, the term "RCP Stream" will be used in place of "RDMAP Stream".

2.1.7 RDMAP Message

The following should be considered for inclusion in the definition section to replace "RDMAP Message":

- RCP Message – The sequence of packets of the RDMA-capable protocol which represent a single RDMA operation or a part of RDMA Read Operation. For TCP or SCTP, an RCP Message is also known as an RDMAP Message.

In the body of the document, the term "RCP Message" will be used in place of "RDMAP Message". The exception is when the term "RDMAP Message" is used to describe the iSER Hello and HelloReply Messages. Here "RDMAP Message" will be replaced by "iSER Message" in order to accommodate LLPs that have message delivery capability, such as SCTP

or [IB]. The iSCSI layer may use that messaging capability immediately after connection establishment before enabling iSER-assisted mode. In this case the iSER Hello and HelloReply Messages are not the first RCP Messages, but they are the first iSER Messages.

2.2 The following should be placed/updated in Acronym Section

HCA	Host Channel Adapter
IB	InfiniBand
IPoIB	IP over InfiniBand
LLP	Lower Layer Protocol
SCTP	Stream Control Transmission Protocol
TCA	Target Channel Adapter

2.3 Connection Establishment, Login, and Transition to iSER

The discussion of connection establishment and the use of a messaging protocol for exchanging Login Request and Login Response Messages should be discussed for SCTP and IB, along with the transitioning of an IB connection to iSER mode, the suggested detail changes can be found (along with others) in the Appendix of this document.

2.4 Security considerations

The discussion of Security should specify that all non IP protocols will define their own requirements for IPsec. However the iSCSI requirements for IPsec are still required wherever an iSER Message enters an IP environment from a non IP one (such as IB). Further the iSCSI/iSER requirement for IPsec on IP based protocols such as TCP and SCTP will continue to require IPsec as a must implement, but optional to use feature. The suggested changes can be found (along with others) in the Appendix of this document.

2.5 Adjustments to Appendix

Rename the current appendix “Appendix A”

Modify the term “iWARP” to “iWARP/TCP” in every section heading in this Appendix.

2.6 Add Appendix B

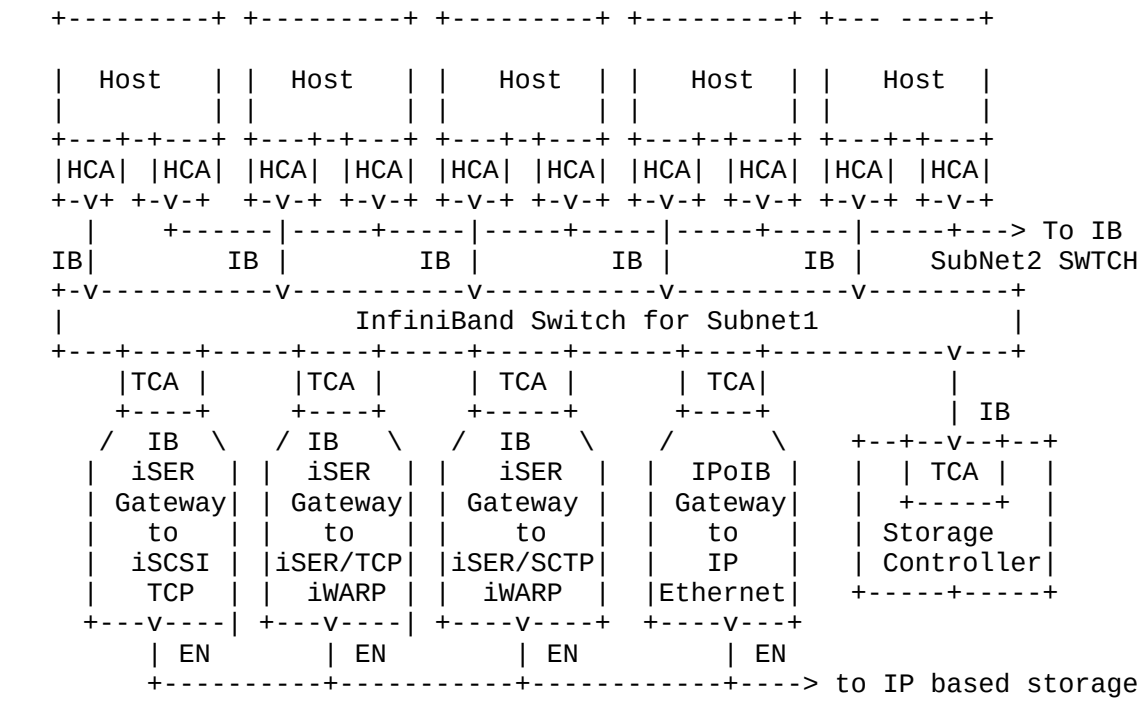
An appendix (Appendix B) should be considered for the iSER draft to explain how an InfiniBand RC connection can be used to carry iSER protocols. The content of any such appendix should be similar in nature to that presented in section 3 of this document.

3 Architectural discussion of iSER over InfiniBand

The following is an explanation of how an InfiniBand network (with Gateways) would be structured. It is intended to provide an insight on how iSER is used in an InfiniBand environment and be generally informational. This information is NOT being proposed as text for the main body of the iSER document. It may be considered for a NON Normative informational Appendix within the iSER document, but its primary purpose is to help put the idea of an iSER operating on InfiniBand into perspective for the readers of this document.

3.1 The Host side of the InfiniBand iSCSI & iSER connections

The following figure (2) defines the topologies in which iSCSI and iSER will be able to operate on an InfiniBand Network.



Ethernet links that carry iSCSI or iWARP

Figure 1 iSCSI, and iSER on IB

In Figure 2, the Host systems are connected via the InfiniBand Host Channel Adapters (HCAs) to the InfiniBand links. With the use of IB switch(es), the InfiniBand links connect the HCA to InfiniBand Target Channel Adapters (TCAs) located in gateways or Storage Controllers. An iSER-capable IB-IP Gateway converts the iSER Messages encapsulated in IB protocols to either standard iSCSI, or iSER Messages for iWARP. An IPoIB Gateway converts InfiniBand IPoIB protocols to IP protocols, and in the iSCSI case, permits iSCSI to be operated on an IB Network between the Hosts and the IPoIB Gateway.

3.2 The Storage side of iSCSI & iSER mixed network environment

Figure 3 shows a storage controller that has four different portal groups: one supporting only iSCSI (TPG-4), one supporting iSER/iWARP/TCP or iSCSI (TPG-2), one supporting iSER/iWARP/SCTP (TPG-3), and one supporting iSER/IB (TPG-1).

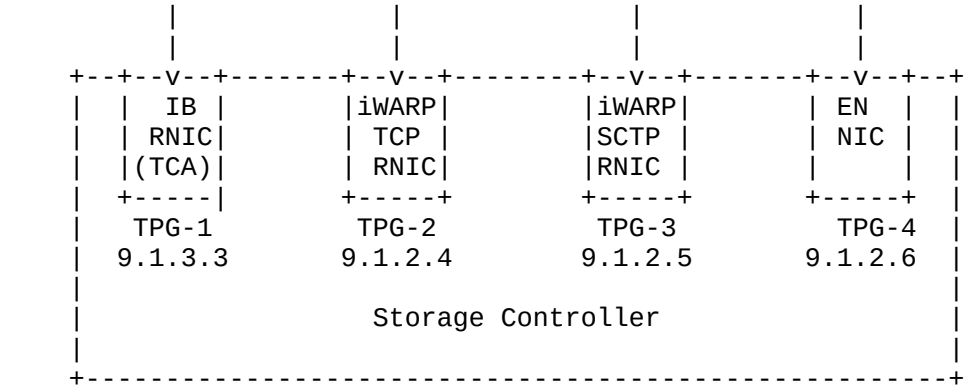


Figure 2 Storage Controller with iSCSI, iWARP/TCP, iWARP/SCTP, and IB connections

The normal iSCSI portal group advertising processes (for SLP, iSNS, or SendTargets commands) are available to a Storage Controller.

3.3 Discovery processes for an InfiniBand Host

An InfiniBand Host system can gather portal group IP address from SLP, iSNS, or the SendTargets discovery processes by using TCP/IP via IPoIB. After obtaining one or more remote portal IP addresses, the Initiator uses the standard IP mechanisms to resolve the IP address to a local outgoing interface and the destination hardware address (Ethernet MAC or IB GID of the target or a gateway leading to the target). If the resolved interface is an IPoIB network interface, then the target portal can be reached through an InfiniBand fabric. In this case the Initiator can establish an iSCSI/TCP or iSCSI/iSER session with the Target over that InfiniBand interface, using the Hardware Address (InfiniBand GID) obtained through the standard Address Resolution (ARP) processes.

If more than one IP address are obtained through the discovery process, the Initiator should select a Target IP address that is on

the same IP subnet as the Initiator if one exists. This will avoid a potential overhead of going through a gateway when a direct path exist.

In addition a user can configure manual static IP route entries if a particular path to the target is preferred.

3.4 IBTA Connection specifications

It is expected that the InfiniBand Trade Association (IBTA) only needs to define:

- Means for permitting a Host to establish an iSCSI/iSER connection with a peer InfiniBand end-node, and indicating when that end-node does not support iSER, so the Host would be able to fall back to iSCSI/TCP over IPoIB.

- Means for permitting the Host to establish connections with IB iSER connections on Storage Controllers or IB iSER connected Gateways in preference to IPoIB connected Gateways/Bridges or connections to Target Storage Controllers that also accept iSCSI via IPoIB.

The IBTA may also decide to specify how iSER may be implemented on connections dealing features that are optional for IB implementations. (See section 5.3)

Everything else that is needed is defined in the generalizations made to the iSER specification in this document.

The following are implementation issues and need not be standardized by IBTA:

- How implementations determine which iSCSI/iSER portal group to use. (Basing the decision on new information that may be placed in the iSCSI discovery information is not required, and simple trial selection is acceptable.)

- How implementations determine how to best handle the concept of multiple connections per session as it deals with multiple IB Addresses:Ports per Portal Group.

4 Preferred changes to iSCSI Discovery Data

The iSER/iWARP/SCTP initiator has similar problems in locating the appropriate Portal Group similar to an InfiniBand initiator. That is, it must check each discovered Portal Group IP address to see if it will accept an iSER/iWARP/SCTP connection.

Even though it is possible for the InfiniBand Host initiator or an iSER/iWARP/SCTP initiator to pick an appropriate connection, the approach may not be optimal, and it takes up needless resources and time by attempting to connect to each portal Group.

Therefore, it is useful to have a connection type associated with the Portal Group Tag that will permit the most appropriate connections to be made without needless connection tries and failures.

For these reasons we will also submit updates to the [iSCSI] (SendTargets), [SLP], and [iSNS] Drafts/RFCs that will add an additional tag following the Portal Group Number. The types will be documented by IANA, with the following definitions.

Portal Group Type	IANA Portal Group Value
iSCSI	0, or blank
iSER/iWARP/TCP	1
iSER/iWARP/SCTP	2
iSER/iWARP/IB	3

If other connection types are defined later they will also need to obtain an IANA Portal Group type value.

The syntax of the parameter will be IP Address:Port[, PG#[, Type#]]

If there is no Type# specified it will be assumed to be an iSCSI capable Portal, or one that must be tested via the connection and Login Process.

5 The Appropriate Standardization Organizations for iSER/IB

5.1 The IETF ips Working Group Standardization

iSER document will be updated to generalize the use of iSER for other RDMA-capable protocol layers such as iSER over IB or iSER over iWARP/SCTP. (See the suggested words in section 2 and the Appendix of this document.) There may also be an informational appendix to explain how the InfiniBand Path connection will be done and how the Login parameters will be handled in an InfiniBand environment.

5.2 The InfiniBand Trade Association (IBTA) Standardization

The InfiniBand Trade Association (IBTA) will be asked to standardize the iSER Connection process that permits the selection of the best path from the Host. (See section 3.4 “IBTA Connection specifications” above.)

5.3 IBTA and optional InfiniBand Features

Since the IBTA has made key capabilities optional that are currently required by iSER, such as ZBTO (a.k.a. ZBVA in InfiniBand) and SendInvSE, it is required that IBTA documents how it will support an iSER connection process (perhaps including the Hello/HelloReply messages) that permits the iSER implementation to understand the limitations of the remote peer. It is also required that IBTA documents how a version of iSER, which understands the limitations of its peer, should operate in that environment.

It should be noted that this places a requirement on the iSER/IB to iSER/iWARP (or iSCSI) Gateway to support the ZBTOs (ZBVAs) on the IP side of the Gateway. This means that it must keep the Virtual Addresses associated with every outstanding STag, so that it can convert the VA to and from the ZBTO required by the iSER/iWARP peer.

5.4 ServiceID and IP Port number

The IBTA will be asked to standardize the iSER ServiceID, and how the ServiceID can be added to the IP port number during the connection process.

6 IANA Considerations

The following additional items will require registration with IANA before the resulting draft can be approved to become an RFC:

None are known at this time.

7 References

7.1 Informative References

[DA] M. Chadalapaka et al., "Datamover Architecture for iSCSI", IETF Internet-draft, draft-ietf-ips-da-00.txt (work in progress), September 2004

[DDP] H. Shah et al., "Direct Data Placement over Reliable Transports", IETF Internet-draft draft-ietf-rddp-ddp-03.txt (work in progress), August 2004

[IPSEC] S. Kent et al., "Security Architecture for the Internet Protocol", RFC 2401, November 1998

[iSCSI] J. Satran et al., "iSCSI", RFC 3720, April 2004

[iSER] M. Ko et. al., ”iSCSI Extensions for RDMA Specification”, IETF Internet-draft draft-ko-iwarp-iser-00.txt

[iSNS] Josh Tseng et. al., Internet Storage Name Service (iSNS), IETF Internet-draft, draft-ietf-ips-isns-22.txt

[MPA] P. Culley et al., "Marker PDU Aligned Framing for TCP Specification", IETF Internet-draft draft-ietf-rddp-mpa-01.txt (work in progress), July 2004

[RDMAP] R. Recio et al., "An RDMA Protocol Specification", IETF Internet-draft draft-ietf-rddp-rdmap-01.txt (work in progress), October 2003

[SAM2] T10/1157D, SCSI Architecture Model - 2 (SAM-2)

[SLP] M. Bakke et. al., Finding iSCSI Targets and Name Servers Using SLP, IETF Internet-draft, draft-ietf-ips-iscsi-slp-09.txt

[TCP] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981

[VERBS] J. Hilland et al., "RDMA Protocol Verbs Specification", RDMAC Consortium Draft Specification draft-hilland-iwarp-verbs-v1.0-RDMAC, April 2003

8 Appendix

8.1 Additional detailed [iSER] document modification

The new terms introduced in the subsections under section 2.1 will replace the existing ones in the [iSER] document where appropriate. In addition, the following changes and clarifications are needed.

8.1.1 Adjustment to Section 2.1 Motivation

The fourth paragraph should be adjusted such as:

Supporting direct data placement is the main function of an RDMA-capable protocol. An RDMA-Capable Controller (such as a NIC enhanced with the RDMAP/DDP functions layered on top of MPA/TCP or SCTP, or an Infiniband Host Channel Adapter or Target Channel Adapter) can be used by any application that has been extended to support RDMA.

8.1.2 Adjustment to Section 2.2 Architectural Goals

The following are changes that should be considered for the numbered paragraphs:

1. Provide an RDMA data transfer model for iSCSI that enables direct in order or out of order data placement of SCSI data into pre-allocated SCSI buffers while maintaining in order data delivery.

5. Allow initiator and target implementations that utilize generic RDMA-capable controllers such as RNICs and implement iSCSI and iSER in software (not require iSCSI or iSER specific assists in the RDMA-capable protocol or RDMA-capable controller).

6. Require full and only generic RDMA-capable protocol functionality at both the initiator and the target.

8.1.3 Adjustment to Section 2.3 Protocol Overview

The following is a suggestion for changes that should be considered for the paragraph number 6:

6. The RDMA-capable protocol guarantees data integrity. (For example, for TCP, iWARP includes a CRC-enhanced framing layer (called MPA) on top of TCP; for SCTP, the CRCs are included in the SCTP protocol; and for Infiniband, the CRCs are included in the

Reliable Connection mode.) For this reason, iSCSI header and data digests are negotiated to "None" for iSCSI/iSER sessions.

8.1.4 Adjustment to Section 2.4 RDMA services and iSER

Additional generalization wordage is needed. The following is a change that should be considered for the first paragraph:

iSER is designed to work with software and/or hardware protocol stacks providing the protocol services defined in RDMA-capable protocol documents such as [RDMAP], [IB], etc.

8.1.5 Adjustment to Section 2.7 iSCSI/iSER Layering

The layering wordage needs additional generalization and the example needs to be made more general. The following is suggested wordage and a suggested replacement for figure 1:

“iSCSI Extensions for RDMA (iSER) is layered between the iSCSI layer and the RDMA-capable protocol layer. Figure 1 shows an example of the relationship between SCSI, iSCSI, iSER, RDMA-capable protocol layers such as iWARP and [IB], and the underlying transports such as TCP, SCTP, or [IB].”

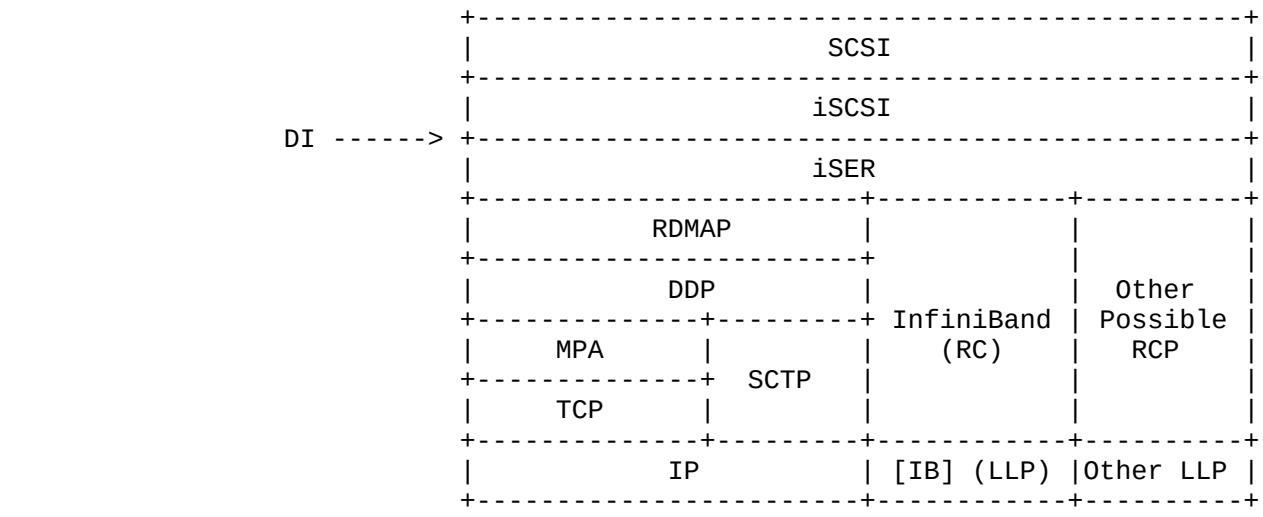


Figure 3 Generic example of iSCSI/iSER Layering in Full Feature Mode

8.1.6 Generalization of Other iSER Sections

The following are changes that should be considered for section 4.1 Interactions with the iWARP Layer (which should change to Interaction with the RDMA-Capable Protocol)

. In the next to last * paragraph consider replacing that paragraph with the following:

o “* For LLPs operating in the stream mode, such as TCP, the RDMA-capable protocol implementation supports the enabling

of the RDMA mode after Connection establishment and the exchange of Login parameters in stream mode. For LLPs that have message delivery capability, such as SCTP or [IB], the iSCSI Layer may use that messaging capability immediately after connection establishment before enabling iSER-assisted mode. The native messaging facility of such an LLP may be used for the Login parameter exchanges."

The following are changes that should be considered for section 4.2 Interactions with the Transport Layer

. "The iSER Layer does not interface with the transport layer (e.g., TCP, SCTP or [IB]) directly. During Connection setup, the iSCSI Layer is responsible for setting up the Connection. If the login is successful, the iSCSI Layer invokes the Enable_Datamover Operational Primitive to request the iSER Layer to transition to the iSER- assisted mode for that iSCSI connection. See section 5.1 on iSCSI/iSER Connection setup. After transitioning to iSER-assisted mode, the RDMA-capable protocol layer and the underlying LLP are responsible for maintaining the Connection and reporting to the iSER Layer of any Connection failures. "

8.1.6.1 Adjustments to 5.1 iSCSI/iSER Connection Setup

The following is a new Section 5.1 paragraph (insert after paragraph 1) that should be considered

. "When a messaging capability is supported by the underlying transport (e.g. SCTP or InfiniBand), the messaging capability may be used by both the initiator and the target to exchange the iSCSI Login Request and Login Response PDUs. The method for establishing the actual connection is protocol specific and outside the scope of this specification."

The following are changes that should be considered for the last paragraph in 5.1

. "When the RDMAExtensions key is negotiated to "Yes", the HeaderDigest and the DataDigest keys MUST be negotiated to "None" on all iSCSI/iSER connections participating in that iSCSI session. This is because, for an iSCSI/iSER connection, the RDMA-capable protocol provides a CRC based error detection for all iSER Messages."

8.1.6.2 Adjustment to Section 5.1.1 Initiator Behavior

The following are changes that should be considered for the 11th paragraph of section 5.1.1 Initiator Behavior.

- “3. If necessary, the iSER Layer MUST enable the RDMA-capable protocol and transition the connection to iSER-assisted mode. (Some RDMA-capable protocol, such as [IB], does not require special enablement for RDMA support.)”

8.1.6.3 Adjustment to Section 5.1.2 Target Behavior

In section 5.1.2 all the references to “iWARP” should be replaced with “the RDMA-capable protocol”.

Also in Section 5.1.2, the paragraph numbered as “4.” The following should be considered as a replacement

- “4. After sending the final SCSI Login Response PDU, the iSER Layer MUST enable the RDMA-capable protocol if necessary and transition the connection to iSER-assisted mode. (Some RDMA-capable protocol, such as [IB], does not require special enablement for RDMA support.) Note that for TCP, the final SCSI Login Response PDU is sent in byte stream mode.”

And the last paragraph in Section 5.1.2 should consider the following for a replacement

- “Note: In the above sequence, the operations as described in bullets 3 and 4 must be performed atomically for iWARP connections. Failure to do this may result in race conditions.”

The following are changes that should be considered for the second paragraph of 5.1.3 iSER Hello Exchange. (It tolerates connections that might already be in RDMA mode when the Hello Exchanges were sent.)

- “In response to the iSER Hello Message, the iSER Layer at the target MUST return the iSER HelloReply Message as the first RCP Message sent by the target after the connection transitions into iSER-assisted mode. The iSER HelloReply Message is used by the iSER Layer at the target to declare iSER parameters to the initiator. See section 9.4 on iSER Header Format for iSER HelloReply Message.”

8.1.6.4 Adjustments to Section 11 Security Considerations

The following paragraphs should be considered as the replacement paragraphs for Section 11 Security Considerations.

“When iSER is layered on top of an RDMA-capable protocol layer and provides the RMDA extension to the iSCSI protocol, the security considerations of iSER are similar to that of the underlying RDMA-capable protocol layer. For iWARP, this is described in [RDMAP].

If iSER is layered on top of the iWARP protocol Stack (TCP or SCTP), all the Security protocol mechanisms described in [iSCSI] may be deployed for an iSCSI/iSER connection. If the IPsec mechanism is used, then it MUST be established before the connection transitions to iSER-assisted mode.

If iSER is layered on top of a non-IP based RDMA-capable protocol layer, the non-IPsec security protocol mechanisms described in [iSCSI] MAY be deployed for an iSCSI/iSER connection. The authorized standards organization for that network’s protocols (such as InfiniBand Trade Associations) is responsible for determining the capability and requirement of that environment on the implementation of IPsec.

If iSER is layered on top of a non-IP protocol, the IPsec protocols and features, as specified in [iSCSI] MUST be implemented at any point where the iSER protocol enters the IP network (e.g., via gateways).

8.1.7 Adjustments to 13.2 Informational References

Add the following references:

[SCTP] R. Stewart, et. al., “Stream Control Transmission Protocol”,

RFC 2960 October 2000 (Updated by RFC3309)

[IB] InfiniBand Architecture Specification Volume 1 Release 1.2,

October 2004

[IPoIB] H.K. Chu et al, "Transmission of IP over InfiniBand", IETF Internet-draft draft-ietf-ipoib-ip-over-infiniband-07.txt (work in progress), August, 2004

9 Author's Address

John Hufferd
IBM Corp.
5600 Cottle Rd.
San Jose, CA 95120, USA
Phone: +1-408-256-0403
Email: hufferd@us.ibm.com

Mike Ko
IBM Corp.
650 Harry Rd.
San Jose, CA 95120, USA
Phone: +1-408-927-2085
Email: mako@us.ibm.com

Yaron Haviv
Voltaire Ltd.
9 Hamanofim St.
Herzelia 46725, Israel
Phone: +972.9.9717655
Email: yaronh@voltaire.com

10 Acknowledgments

David Black
EMC Corporation
176 South St.
Hopkinton, MA 01748, USA
Phone: +1-508-293-7953
Email: black_david@emc.com

Mallikarjun Chadalapaka
Hewlett-Packard Company
8000 Foothills Blvd.
Roseville, CA 95747-5668, USA
Phone: +1-916-785-5621
Email: cbm@rose.hp.com

Mike Krause
Hewlett-Packard Company
43LN
19410 Homestead Road
Cupertino, CA 95014, USA
Phone: +1-408-447-3191
Email: krause@cup.hp.com

Alex Nezhinsky
Voltaire Ltd.
9 Hamanofim St.
Herzelia 46725, Israel
Phone: +972.9.9717637
Email: alexn@voltaire.com

Renato J. Recio
IBM Corp.
11501 Burnett Road
Austin, TX 78758, USA
Phone: +1-512-838-3685
Email: recio@us.ibm.com

Tom Talpey
Network Appliance
375 Totten Pond Road
Waltham, MA 02451, USA
Phone: +1-781-768-5329
EMail: thomas.talpey@netapp.com

11 Full Copyright Statement