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

John Hufferd
Brocade

Mike Ko
IBM Corporation

Yaron Haviv
Voltaire Ltd

July, 2005

Expires: January, 2006

Generalization of iSER for 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. This document generalizes the iSER document to permit it to be used with other RDMA capable protocols such as InfiniBand.

Table of Contents

Table of Figures

1 Motivation

Currently the work to define iSCSI extensions for RDMA [iSER] only considers using the iWARP protocol suite. 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 Protocol layers now and in the future such as 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, 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 are to be made 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 iWARP, but to change/adjust the wording in such a way that iSCSI/iSER can be layered over a different RDMA-capable protocol layer such as InfiniBand.

The details of many of the suggested changes can be found in the Section 3 of this document.

In general the iSER specification is hereby modified to apply to not only iWARP, but to other RDMA-Capable Protocols, such as InfiniBand. Except for the unique features of non iWARP protocols dealing with initial Login and Security, the rest of the iSER document is applicable these other RDMA-Capable Protocols (such as InfiniBand.)

2.1 Generalization of Definitions

It is required that some of the terminology be clarified as to the 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 are introducing 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 is hereby generalized as follows:

1. Whenever the term "iWARP protocol suite" occurs in the iSER draft, it is hereby replaced by "RDMA-Capable Protocol". In addition, the phrase "such as the iWARP protocol suite" is hereby added only where necessary to denote cases that only apply for iWARP.

2. Whenever the term "iWARP layer" occurs in the iSER draft, it is hereby replaced by "RDMA-capable protocol layer". In addition,

the phrase "such as the iWARP Layer" is hereby added only where necessary to denote cases that only apply for iWARP.

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

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

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 is hereby used whenever the intention is to refer to a general controller that provides RDMA functionality. In addition, the clause "such as an RNIC" is hereby added as necessary where the clear intent of the statement is to address an iWARP RDMA-Capable Controller.

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 hereby has its definition extended so that it applies to both a Tag for a Remote Buffer, and the Tag for a Local Buffer. The following is a replacement for the existing Steering Tag definition 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 layers, the Steering Tag may be known by different names but even so they will be herein referred to as STags. For example, for InfiniBand, a Remote

STag is known as an R-Key, and a Local STag is known as an L-Key and they will both be considered STags.

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 is added to the definition for IRD: "For other RDMA-Capable protocol layers, 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 is 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" is hereby used whenever any RDMA wire protocol or RDMA protocol stack is applicable. Only when the document intends to explicitly address a specific iWARP wire protocol is the term [RDMAP] used.

2.1.6 RDMAP Layer

In the body of the document the term "RDMAP Layer" is hereby replaced with the term "RCP Layer".

2.1.7 RDMAP Stream

The following is hereby included in the definition section replacing the term "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, 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 "RDMAP Stream" is hereby replaced by the term "RCP Stream".

2.1.8 RDMAP Message

The following is included 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, an RCP Message is also known as an RDMAP Message.

In the body of the document, the term "RDMAP Message" is hereby replaced by the term "RCP Message". The exception to this is when the term "RDMAP Message" is used to describe the iSER Hello and HelloReply Messages. Here "RDMAP Message" is hereby replaced by "iSER Message" in order to accommodate LLPs that have message delivery capability such as [IB]. The iSCSI layer may use that messaging capability immediately after connection establishment before enabling iSER-assisted mode. In the case the iSER Hello and HelloReply Messages are not the first RCP Messages, but they are the first iSER Messages.

2.2 The following is placed/updated in the Acronym Section

HCA	Host Channel Adapter
IB	InfiniBand
IPoIB	IP over InfiniBand
LLP	Lower Layer 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 PDUs for IB are inserted in this section, along with the extended specification of the transition of an IB connection to iSER mode. The suggested detail changes can be found in section 3.1.6.1 through section 3.1.6.3 of this document.

2.4 Security considerations

The security consideration are updated to include requirements on security for transports other than TCP, the document now states that the security concerns must be addressed appropriately for different transport environments. However the iSCSI implementation 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 will continue to require IPsec as a must implement, but optional to use. There is now a SHOULD implement (optional to use) requirement on non IP networks for a packet by packet security facility that is at least as strong as that required by [iSCSI].

The exact wordage can be found in section 3.1.6.4 of this document.

2.5 Adjustments to the iSER Appendix.

The current iSER appendix will hereby be renamed "Appendix A".

2.6 Add Appendix B

A new informational appendix (Appendix B) is hereby added that explains how an InfiniBand RC connection can be used to carry the iSER protocol. The content of the new appendix B is that which is contained in the appendix (section 6) of this document.

3 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.

3.1.1 Adjustment to Section 2.1 Motivation

The fourth paragraph is hereby adjusted to:

> 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 an InfiniBand Host Channel Adapter or Target Channel Adapter) can be used by any application that has been extended to support RDMA.

3.1.2 Adjustment to Section 2.2 Architectural Goals

The following are changes 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 to utilize generic RDMA-Capable Controllers such as RNICs, or 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.

3.1.3 Adjustment to Section 2.3 Protocol Overview

The following change is hereby made to 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; 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.

3.1.4 Adjustment to Section 2.4 RDMA services and iSER

The following change is hereby made to 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.

3.1.5 Adjustment to Section 2.7 iSCSI/iSER Layering

The layering wordage needed additional generalization and the example needed to be made more general. Therefore, the following is the change in wordage and the 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, or [IB]. Note that the iSCSI layer as described here supports the RDMA Extensions as used in iSER."

```
                    +----------------------------------------------+
                    |                      SCSI                    |
                    +----------------------------------------------+
                    |                     iSCSI                    |
         DI ------> +----------------------------------------------+
                    |                     iSER                     |
                    +--------------+-----------------+-------------+
                    |    RDMAP     |                 |             |
                    +--------------+                 |             |
                    |     DDP      |                 |   Other     |
                    +--------------+ InfiniBand (RC) |  Possible   |
                    |     MPA      |                 |    RCP      |
                    +--------------+                 |             |
                    |     TCP      |                 |             |
                    +--------------+-----------------+-------------+
                    |     IP       |    [IB] (LLP)   |  Other LLP  |
                    +--------------+-----------------+-------------+
```

Figure 1 - Example of iSCSI/iSER Layering in Full Feature Mode

3.1.6 Generalization of Other iSER Sections

The title of section 4.1 -- "Interaction with the iWARP Layer" -- is hereby changed to "Interaction with the RDMA-Capable Protocol Layer".

The first paragraph in Section 4.1 is hereby changed to:

> The iSER protocol layer is layered on top of the RCP Stack (see Figure 1) and the following are the key features that are assumed to be supported by the RDMA-Capable Protocol Layer

The third paragraph in Section 4.1 is hereby changed to:

> * The layers handling the RDMA Capable Protocol provide reliable, in-order message delivery and direct data placement.

And in that same section (4.1) the next to last * paragraph is hereby replaced with the following:

> *   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 [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 is a replacement for section 4.2 Interactions with the Transport Layer

> The iSER Layer does not directly setup the transport layer connection (e.g., TCP, or [IB]). 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 any Connection failures.

3.1.6.1 Adjustments to 5.1 iSCSI/iSER Connection Setup

The following is a new Section 5.1 paragraph which is hereby inserted after paragraph 1:

> When a reliable messaging capability is supported by the underlying transport (e.g. InfiniBand), the reliable 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 text is hereby added after the second paragraph of Section 5.1:

> If the transport layer is not TCP, and if the RDMAExtensions key is not negotiated to Yes, then the connection will need to be re-established in TCP capable mode. (For InfiniBand this will require an [IPoIB] type connection.)

The following text is hereby added after the third paragraph of Section 5.1:

> Discovery sessions are always conducted using the TCP transport layer.

The following is a replacement 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.

3.1.6.2 Adjustment to Section 5.1.1 Initiator Behavior

The following are changes 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 Protocols, such as [IB], do not require special enablement for RDMA support.)

3.1.6.3 Adjustment to Section 5.1.2 Target Behavior

In section 5.1.2 all the references to "iWARP" are hereby replaced with "the RDMA-Capable Protocol".

Also in Section 5.1.2, the paragraph numbered as "3." & "4." are hereby replaced with the following:

> 3. If the underlying transport is TCP, then the iSER Layer MUST send the final SCSI Login Response PDU in byte stream mode to conclude the iSCSI Login Phase. If the underlying transport has reliable messaging capability (e.g. IB RC) then the iSER layer MUST send the final SCSI Login Response PDU in the reliable message mode to conclude the iSCSI login phase.

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 Protocols, such as [IB], do not require special enablement for RDMA support.)

And the last paragraph in Section 5.1.2 is hereby replaced with:

> 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 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.

3.1.6.4 Adjustments to Section 11 Security Considerations

The following paragraphs are 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 the same as that of the underlying RDMA-Capable Protocol Layer. For iWARP, this is described in [RDMAP] and [RDDPSEC].
>
> Since iSER-assisted iSCSI protocol is still functionally iSCSI from a security considerations perspective, all of the iSCSI security requirements as described in [RFC3720] and [RFC3723] apply.
>
> 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, all the security protocol mechanisms applicable to that RDMA-Capable Protocol Layer is also applicable to an iSCSI/iSER connection.

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). And the non-IP protocol SHOULD implement (optional to use) a packet by packet security protocol equal in strength to the IPsec protocol specified by [iSCSI].

3.1.7 Adjustments to 13.2 Informational References

Add the following references:

[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

4 IANA Considerations

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

None are known at this time.

5 References

5.1 Informative References

[DA] M. Chadalapaka et al., "Datamover Architecture for iSCSI", IETF Internet-draft, draft-ietf-ips-iwarp-da-03.txt (work in progress), June 2005

[DDP] H. Shah et al., "Direct Data Placement over Reliable Transports", IETF Internet-draft draft-ietf-rddp-ddp-04.txt (work in progress), February 2005

[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-ietf-ips-iser-04.txt (work in progress), July 2005

[iSNS] Josh Tseng et. al., Internet Storage Name Service (iSNS), IETF Internet-draft, draft-ietf-ips-isns-22.txt (work in progress), February 2004

[MPA] P. Culley et al., "Marker PDU Aligned Framing for TCP Specification", IETF Internet-draft draft-ietf-rddp-mpa-02.txt (work in progress), February 2005

[RDMAP] R. Recio et al., "An RDMA Protocol Specification", IETF Internet-draft draft-ietf-rddp-rdmap-04.txt (work in progress), April 2005

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

[SLP] M. Bakke et. al., "Finding iSCSI Targets and Name Servers by Using SLPv2", RFC 4018, April 2005

[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-v1.0a, May 2003

6 Appendix

6.1 Architectural discussion of iSER over InfiniBand

This entire appendix is hereby included as Appendix B in the iSCSI Extensions for RDMA document [iSER].

The following is an explanation of how an InfiniBand network (with Gateways) would be structured. It is intended to provide insight on how iSER is used in an InfiniBand environment and be generally informational. It is informational only and it is intended to put the idea of an iSER operating on InfiniBand into perspective for the readers of this document.

6.2 The Host side of the InfiniBand iSCSI & iSER connections

Figure 2 (iSCSI, and iSER on IB) defines the topologies in which iSCSI and iSER will be able to operate on an InfiniBand Network.

```
   +---------+ +---------+ +---------+ +---------+ +---  -----+
   |  Host   | |  Host   | |   Host  | |   Host  | |   Host  |
   |         | |         | |         | |         | |         |
   +---+-+---+ +---+-+---+ +---+-+---+ +---+-+---+ +---+-+---+
   |HCA| |HCA| |HCA| |HCA| |HCA| |HCA| |HCA| |HCA| |HCA| |HCA|
   +-v+ +-v-+  +-v-+ +-v-+ +-v-+ +-v-+ +-v-+ +-v-+ +-v-+ +-v-+
     |----+------|-----+-----|-----+-----|-----+-----|-----+---> To IB
   IB|         IB |         IB |         IB |         IB |    SubNet2 SWTCH
   +-v-----------v-----------v-----------v-----------v---------+
   |                  InfiniBand Switch for Subnet1            |
   +---+-----+--------+-----+--------+-----+------------v------+
       | TCA |        | TCA |        | TCA |            |
       +-----+        +-----+        +-----+            | IB
      /  IB   \      /  IB   \      /       \      +--+--v--+--+
     |  iSER   |    |  iSER   |    |  IPoIB  |     |  | TCA |  |
     | Gateway |    | Gateway |    | Gateway |     |  +-----+  |
     |   to    |    |   to    |    |   to    |     | Storage   |
     |  iSCSI  |    |  iSER   |    |   IP    |     | Controller|
     |   TCP   |    |  iWARP  |    |Ethernet |     +-----+-----+
     +---v-----|    +---v-----|    +----v----+
         | EN            | EN              | EN
         +---------------+----------------+----> to IP based storage

           Ethernet links that carry iSCSI or iWARP
```

Figure 2 - 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 the InfiniBand [IPoIB] protocol to IP protocol, and in the iSCSI case, permits iSCSI to be operated on an IB Network between the Hosts and the [IPoIB] Gateway.

6.3 The Storage side of iSCSI & iSER mixed network environment

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

```
          |                |                |
          |                |                |
    +--+--v--+----------+--v--+----------+--v--+--+
    |  | IB  |          |iWARP|          | EN  |  |
    |  |     |          | TCP |          | NIC |  |
    |  |(TCA)|          | RNIC|          |     |  |
    |  +-----|          +-----+          +-----+  |
    |   TPG-1            TPG-2            TPG-4   |
    |  9.1.3.3          9.1.2.4          9.1.2.6  |
    |                                             |
    |                  Storage Controller         |
    |                                             |
    +---------------------------------------------+
```

Figure 3 - Storage Controller with TCP, iWARP, and IB Connections

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

6.4 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 exists.

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

6.5 IBTA Connection specifications

It is outside the scope of this document, but it is expected that the InfiniBand Trade Association (IBTA) has or will define:

- The iSER ServiceID

- A Means for permitting a Host to establish a connection with a peer InfiniBand end-node, and that peer indicating when that end-node supports iSER, so the Host would be able to fall back to iSCSI/TCP over [IPoIB].

- A 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].

- A Means for combining the IB ServiceID for iSER and the IP port number such that the IB Host can use normal IB connection processes, yet ensure that the iSER target peer can actually connect to the required IP port number.

7 Author’s Address

John Hufferd
Brocade
1745 Technology Drive
San Jose, CA 95110, USA
Phone: +1-408-333-5244
Email: jhufferd@brocade.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

8 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

9 Full Copyright Statement