

*Supplement to InfiniBand™  
Architecture Specification  
Volume 1 Release 1.2*



*Annex A12:  
Support for iSCSI  
Extensions for RDMA*

September 8, 2006

Copyright © 2006 by InfiniBand™ Trade Association.  
All rights reserved.

All trademarks and brands are the property of their respective owners.

This document contains information proprietary to the InfiniBand™ Trade Association. Use or disclosure without written permission by an officer of the InfiniBand™ Trade Association is prohibited.

**Table 0 Revision History**

Revision	Date	
0.9	09/08/2006	Draft document for general IBTA member review

**LEGAL DISCLAIMER**

**This specification provided “AS IS” and without any warranty of any kind, including, without limitation, any express or implied warranty of non-infringement, merchantability or fitness for a particular purpose.**

**In no event shall IBTA or any member of IBTA be liable for any direct, indirect, special, exemplary, punitive, or consequential damages, including, without limitation, lost profits, even if advised of the possibility of such damages.**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42

## TABLE OF CONTENTS

---

<b>Annex A12:</b>	<b>Support for iSCSI Extensions for RDMA (iSER) .....</b>	<b>6</b>	1
A12.1	Introduction .....	6	2
A12.2	Glossary .....	6	3
A12.3	Convention .....	7	4
A12.4	iSER Connection Establishment .....	7	5
A12.5	iSER Extension for Handling Send with Invalidate .....	10	6
A12.6	iSER Extension for Handling Virtual Address .....	10	7
A12.7	Compliance Summary .....	11	8
			9
			10
			11
			12
			13
			14
			15
			16
			17
			18
			19
			20
			21
			22
			23
			24
			25
			26
			27
			28
			29
			30
			31
			32
			33
			34
			35
			36
			37
			38
			39
			40
			41
			42

BETA

---

## LIST OF FIGURES

---

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42

IBTA

## LIST OF TABLES

---

		1
		2
		3
		4
Table 0	Revision History .....	2
Table 1	iSER Service ID Format .....	8
Table 2	iSER CM REQ Message Private Data Format .....	8
Table 3	iSER CM REP Message Private Data Format.....	9
Table 4	Expanded iSER Header for Supporting Virtual Address .....	11
		10
		11
		12
		13
		14
		15
		16
		17
		18
		19
		20
		21
		22
		23
		24
		25
		26
		27
		28
		29
		30
		31
		32
		33
		34
		35
		36
		37
		38
		39
		40
		41
		42

## ANNEX A12: SUPPORT FOR iSCSI EXTENSIONS FOR RDMA (iSER)

### A12.1 INTRODUCTION

The iSCSI protocol is an IETF standard (RFC3720) which maps the SCSI Architecture Model over the TCP protocol. SCSI commands are carried by iSCSI requests, and SCSI responses and status are carried by iSCSI responses. Other iSCSI protocol exchanges and SCSI Data are also transported in iSCSI PDUs. iSCSI Extensions for RDMA (iSER) is an IETF standard which adds the RDMA data transfer capability to iSCSI by layering iSCSI on top of an RDMA-Capable Protocol such as InfiniBand. (As of this writing, the iSER specification has not yet been released as an IETF RFC standard. The current version is draft-ietf-ips-iser-05. Please refer to the IETF website at [www.ietf.org](http://www.ietf.org) for the final version.) An RDMA-Capable Protocol provides RDMA Read and Write services, which enable data to be transferred directly into SCSI I/O Buffers without intermediate data copies. Using iSER as a storage protocol in an InfiniBand environment also means that existing iSCSI infrastructure (sometimes referred to as "iSCSI ecosystem") including but not limited to MIB, bootstrapping, negotiation, naming & discovery, and security can be utilized as well.

The IETF iSER standard requires certain features which are optional to implement in InfiniBand, including the following:

- Zero-Based Virtual Address
- Send with Invalidate

**CA12-1:** The iSER standard as released by IETF, shall be followed when iSER is used as a storage protocol in an InfiniBand environment except for the handling of Zero-Based Virtual Address and Send with Invalidate.

This annex describes the connection establishment issues for iSER including iSER Service ID and CM Private Data format in an InfiniBand environment and how Zero-Based Virtual Address and Send with Invalidate should be handled if they are not supported.

### A12.2 GLOSSARY

#### Inbound RDMA Read Queue Depth (IRD)

The maximum number of incoming outstanding RDMA Read Requests that the RDMA-Capable Controller can handle on a particular RDMA-Capable Protocol Stream at the Data Source. For InfiniBand, the equivalent for IRD is the Responder Resources.

#### Initiator

This refers to the iSCSI initiator node which originates device service and task management requests to be processed by an iSCSI target node.

### Outbound RDMA Read Queue Depth (ORD)

The maximum number of outstanding RDMA Read Requests that the RDMA-Capable Controller can initiate on a particular RDMA-Capable Protocol Stream at the Data Sink. For InfiniBand, the equivalent for ORD is the Initiator Depth.

### Steering Tag (STag)

An identifier of a Tagged Buffer on a node (local or remote) as defined in "An RDMA Protocol Specification" (RDMAP) and "Direct Data Placement over Reliable Transports" (DDP). (As of this writing, the RDMAP and DDP specifications have not yet been released as IETF RFC standards. The current versions are draft-ietf-rddp-rdmap-05 and draft-ietf-rddp-ddp-05. Please refer to the IETF website at [www.ietf.org](http://www.ietf.org) for the final version.) For Infiniband, an STag for remote access is known as an R-Key, and an STag for local access is known as an L-Key, and both will be considered STags in iSER.

### Target

This refers to the iSCSI target node which receives device service and task management requests for processing.

## A12.3 CONVENTION

Though the iSER specification uses the same big endian byte ordering convention as defined in section [1.5.1 Byte Ordering on page 66](#), the rule for bit ordering is different. In the iSER specification, for a byte, bit 0 is the most significant bit and bit 7 is the least significant bit. Similarly, for a word, bit 0 is the most significant bit and bit 31 is the least significant bit. In this annex, the convention for bit ordering and byte ordering as defined in section [1.5.1 Byte Ordering on page 66](#) is used.

## A12.4 ISER CONNECTION ESTABLISHMENT

iSER connection setup uses InfiniBand CM MADs, with additional iSER information exchanged in the private data. Please refer to [Chapter 12: Communication Management on page 650](#) for more details.

**CA12-2:** All iSER messages shall be transmitted via the InfiniBand RC channel.

**CA12-3:** iSER shall use the RDMA-Aware Service ID as defined in [Annex A11: RDMA IP CM Service on page 6](#). The iSER Service ID shall conform to the format as defined in [Table 1 iSER Service ID Format on page 8](#).

**Table 1 iSER Service ID Format**

Byte Location	Description	Value
0 to 4	Service ID for an RDMA-Aware ULP	As defined in the RDMA IP CM Service Annex
5	Protocol	0x06 (TCP)
6 to 7	Destination port number	Set by iSCSI/iSER (usually the iSCSI well-known TCP port number)

Discovery of target portals (IP address and port number of iSCSI targets) is done using the iSCSI discovery mechanisms as described in the IETF document “iSCSI Naming and Discovery” (RFC3721).

For connection establishment, the initiator sends a CM REQ using the iSER Service ID. The CM Private Data field contains additional parameters relating to the iSER connection and IB options support status.

**CA12-4:** The iSER CM REQ Message Private Data shall conform to the format as defined in [Table 2 iSER CM REQ Message Private Data Format on page 8](#).

**Table 2 iSER CM REQ Message Private Data Format**

Byte	Bit	Description	Value
0-35		CM REQ Message Extension	As defined in the RDMA IP CM Service Annex
36-39	31	Zero Based Virtual Address Exception	0 - Initiator supports Zero Based Virtual Address (ZBVA) 1 - Initiator does not support Zero Based Virtual Address (ZBVA)
	30	Send with Invalidate Exception	0 - Initiator supports Send with Invalidate 1 - Initiator does not support Send with Invalidate
	29-0	reserved	Set to zero and ignored on receive
40-91		reserved	Set to zero and ignored on receive

**CA12-5:** Before accepting a CM REQ, a target shall check the content of the non-Private Data portion of the CM REQ first, followed by the CM REQ Message Extension in the Private Data, and then the rest of the Private Data, in that order.

**CA12-6:** To prevent address spoofing attempts by user applications, the source IP address and the port number shall be filled in by privileged kernel mode. The passive side shall verify that the CM REQ Message contains a privileged Q-key and its value is 0x80010000.

If a target does not support iSER, or if the Protocol or Destination Port Number specified in the Service ID are unacceptable, it returns a CM REJ indicating “Invalid Service ID”. The initiator may attempt connection using iSCSI over IPoIB.

Other errors in the CM REQ resulting in a CM REJ are handled as described in [Section 12.6.7, “REJ - Reject,” on page 662](#).

**CA12-7:** A target shall use the semantics as specified in the RDMA IP CM Service Annex to handle the CM REQ Message Extension in the Private Data.

If a target supports iSER and the content of the Private Data are acceptable, it returns a CM REP Message.

**CA12-8:** The iSER CM REP Message Private Data shall conform to the format as defined in [Table 3 iSER CM REP Message Private Data Format on page 9](#).

**Table 3 iSER CM REP Message Private Data Format**

Byte	Bit	Description	Value
0-3	31	Zero Based Virtual Address Exception	0 - ZBVA shall be used for this connection 1 - VA shall be used for this connection
	30	Send with Invalidate Exception	0 - The target shall issue Send with Invalidate as needed 1 - The target shall issue Send instead of Send with Invalidate
	29 - 0	reserved	Set to zero and ignored on receive

Upon receiving the CM REP Message from the target, the initiator returns the CM RTU Message to complete the connection establishment.

**CA12-9:** An initiator shall proceed with iSCSI login exchanges after completing the CM MAD exchanges to establish a connection.

**CA12-10:** During the iSCSI login exchanges, the RDMAExtensions key shall be negotiated to “Yes”.

During the CM MAD exchanges, the Responder Resources (equivalent to IRD in iSER) and Initiator Depth (equivalent to ORD in iSER) are negotiated. The iSER initiator and target should take note of the negotiated

values and ensure that in the subsequent iSCSI login negotiations and iSER Hello/HelloReply exchanges that the IRD/ORD values reflect the negotiated values of Responder Resources and Initiator Depth.

## A12.5 ISER EXTENSION FOR HANDLING SEND WITH INVALIDATE

**CA12-11:** If both the initiator and the target support Send with Invalidate, then the target shall issue Send with Invalidate for that connection whenever Send with Invalidate is intended to be used as specified in the iSER specification. If either the initiator or the target does not support Send with Invalidate, then the target shall issue Send instead of Send with Invalidate for that connection.

**CA12-12:** An initiator shall set the Send with Invalidate Exception flag in CM REQ as described in [Table 2 iSER CM REQ Message Private Data Format on page 8](#) if it does not support Send with Invalidate. This shall be done per connection.

**CA12-13:** A target shall set the Send with Invalidate Exception flag in CM REP as described in [Table 3 iSER CM REP Message Private Data Format on page 9](#) if either the initiator or the target does not support Send with Invalidate. This shall be done per connection.

## A12.6 ISER EXTENSION FOR HANDLING VIRTUAL ADDRESS

**CA12-14:** If both the initiator and the target support Zero-Based Virtual Address, then Zero-Based Virtual Address shall be used for that connection. If either the initiator or the target does not support Zero-Based Virtual Address, then Virtual Address shall be used for that connection.

**CA12-15:** An initiator shall set the Zero Based Virtual Address Exception flag in CM REQ as described in [Table 2 iSER CM REQ Message Private Data Format on page 8](#) if it does not support Zero Based Virtual Address. This shall be done per connection.

**CA12-16:** A target shall set Zero-Based Virtual Address Exception flag in CM REP as described in [Table 3 iSER CM REP Message Private Data Format on page 9](#) if either the initiator or the target does not support Zero Based Virtual Address. This shall be done per connection.

**CA12-17:** When Zero Based Virtual Address is not supported, both the initiator and the target shall use the expanded iSER header as defined in

[Table 4 Expanded iSER Header for Supporting Virtual Address on page 11](#) for iSCSI control-type PDUs in the connection.

**Table 4 Expanded iSER Header for Supporting Virtual Address**

Byte	Bit	Name	Value
0-3	31-28	Operation code	0x1 - iSCSI Control Type PDU
	27	Write STag Valid (WSV)	When set, indicates that the Write STag and the Write Virtual Address fields are used
	26	Read STag Valid (RSV)	When set, indicates that the Read STag and the Read Virtual Address fields are used
	25-0	reserved	Set to zero and ignored on receive
4-7	31-0	Write STag	- Contains the IB R-key for the SCSI Write command when WSV is set to b'1 - Set to zero and ignored on receive when WSV is set to b'0
8-11	31-0	Write Virtual Address High	- Contains bits 63 to 32 of the IB Virtual Address for the SCSI Write command when WSV is set to b'1 - Set to zero and ignored on receive when WSV is set to b'0
12-15	31-0	Write Virtual Address Low	- Contains bits 31 to 0 of the IB Virtual Address for the SCSI Write command when WSV is set to b'1 - Set to zero and ignored on receive when WSV is set to b'0
16-19	31-0	Read STag	- Contains the IB R-key for the SCSI Read command when RSV is set to b'1 - Set to zero and ignored on receive when RSV is set to b'0
20-23	31-0	Read Virtual Address High	- Contains bits 63 to 32 of the IB Virtual Address for the SCSI Read command when RSV is set to b'1 - Set to zero and ignored on receive when RSV is set to b'0
24-27	31-0	Read Virtual Address Low	- Contains bits 31 to 0 of the IB Virtual Address for the SCSI Read command when RSV is set to b'1 - Set to zero and ignored on receive when RSV is set to b'0

**A12.7 COMPLIANCE SUMMARY**

In order to claim compliance for the Support for iSCSI Extensions for RDMA, a product shall meet all requirements specified in this section.

- CA12-1: IETF iSER standard compliance . . . . . Page 6
- CA12-2: iSER uses only RC . . . . . Page 7
- CA12-3: iSER uses RDMA-Aware Service ID . . . . . Page 8
- CA12-4: iSER CM REQ Message Private Data format . . . . . Page 8

● CA12-5: CM REQ checking order . . . . .	Page 8	1
● CA12-6: iSER CM REQ Message contains a privileged Q-key . . .	Page 9	2
● CA12-7: Handling of iSER CM REQ Message Extension . . . . .	Page 9	3
● CA12-8: iSER CM REP Message Private Data format . . . . .	Page 9	4
● CA12-9: iSCSI login exchanges after connection establishment . .	Page 9	5
● CA12-10: Negotiation of iSCSI RDMA Extensions key . . . . .	Page 9	6
● CA12-11: Use of Send vs Send with Invalidate in iSER . . . . .	Page 10	7
● CA12-12: Use of Send Invalidate Exception flag by initiator . . . . .	Page 10	8
● CA12-13: Use of Send with Invalidate Exception flag by target . . .	Page 10	9
● CA12-14: Use of Zero-Based Virtual Address vs. Virtual Address .	Page 10	10
● CA12-15: Use of ZBVA Exception flag by initiator . . . . .	Page 10	11
● CA12-16: Use of ZBVA Exception flag by target . . . . .	Page 10	12
● CA12-17: Use expanded iSER header with ZBVA exception . . . . .	Page 10	13
		14
		15
		16
		17
		18
		19
		20
		21
		22
		23
		24
		25
		26
		27
		28
		29
		30
		31
		32
		33
		34
		35
		36
		37
		38
		39
		40
		41
		42