



---

# Chelsio iSCSI Target Exported Kernel API Guide

---

CHELSIO COMMUNICATIONS, INC. Company Confidential  
Do Not Reproduce.  
This copy belongs to \_\_\_\_\_



---

Copyright © 2005 - 2015 Chelsio Communications Inc. All rights reserved.

This document contains proprietary information of Chelsio Communications. No part of the document may be reproduced. Reverse engineering of the hardware or software is prohibited and is protected by patent law.

This material or any portion of it may not be copied in any form or by any means, stored in a retrieval system, adopted or transmitted in any form or by any means (electronic, mechanical, photographic, graphic, optic or otherwise), or translated in any language or computer language without the prior written permission of Chelsio Communications.

Chelsio Communications makes no representation or warranties with respect to the contents herein and shall not be responsible for any loss or damage caused to the user by the direct or indirect use of this information.

All products or company names or trademarks mentioned herein are used for identification purposes only, and may be trademarks or registered trademarks of their respective owners.

---

## Table of Contents

<b>INTRODUCTION.....</b>	<b>5</b>
1.1    AUDIENCE.....	5
1.2    ABOUT THIS DOCUMENT.....	5
1.3    OTHER RELATED DOCUMENTS.....	5
1.4    DOCUMENT REVISION HISTORY.....	5
<b>ISCSI INTRODUCTION.....</b>	<b>7</b>
2.1    SCSI.....	7
2.2    iSCSI.....	7
<b>ISCSI ACCELERATION WITH CHELSIO'S TERMINATOR ASIC.....</b>	<b>12</b>
3.1    TCP OFFLOAD ENGINE.....	12
3.2    CHELSIO ULP iSCSI ACCELERATION.....	12
<b>CHELSIO iSCSI TARGET MODULE.....</b>	<b>14</b>
4.1    iSCSI SOFTWARE DESIGN.....	15
4.2    iSCSI PROCESSING CONTEXT.....	17
4.2.1 <i>Interface functions</i> .....	17
<b>REGISTER WITH iSCSI TARGET MODULE.....</b>	<b>18</b>
5.1    iSCSI TARGET CLASS.....	18
5.1.1 <i>Target Class Identification</i> .....	19
5.1.2 <i>Initiator Login Initialization</i> .....	19
5.1.3 <i>Login Phase Check</i> .....	21
5.1.4 <i>Retrieving CHAP Information</i> .....	22
5.1.5 <i>iSCSI Target Session Management</i> .....	22
5.1.6 <i>Target Discovery</i> .....	23
5.1.7 <i>Target Redirection</i> .....	24
5.2    iSCSI TARGET LUN CLASS.....	24
5.2.1 <i>Target LUN Class Identification</i> .....	26
5.2.2 <i>Target LUN Class Property</i> .....	26
5.2.2.1    LUN_CLASS_SCSI_PASS_THRU_BIT.....	26
5.2.2.2    LUN_CLASS_MULTI_PHASE_DATA_BIT.....	26
5.2.2.3    LUN_CLASS_HAS_CMD_QUEUE_BIT.....	26
5.2.3 <i>Target LUN Configuration Functions</i> .....	27
5.2.4 <i>Handling of iSCSI SCSI Command</i> .....	27
<b>ISCSI TARGET CONFIGURATION.....</b>	<b>30</b>

6.1	TARGET CONFIGURATION VIA KERNEL APIs.....	30
6.1.1	<i>Adding an iSCSI Target Node</i> .....	30
6.1.2	<i>Removing an iSCSI Target Node</i> .....	31
6.1.3	<i>Modify an iSCSI Target Node's Settings</i> .....	31
6.1.3.1	CHAP Configuration.....	32
	<b>ISCSI TARGET PROCESSING.....</b>	<b>34</b>
7.1	iSCSI SESSION.....	34
7.2	iSCSI DISCOVERY SESSION.....	34
7.3	iSCSI NORMAL SESSION.....	34
7.3.1	<i>The Login Phase</i> .....	35
7.3.1.1	First Login Request.....	35
7.3.1.2	Login Stage Transition.....	37
7.3.1.3	CHAP Processing.....	37
7.3.2	Session Addition.....	38
7.3.3	Session Removal.....	38
7.3.4	Session Abort.....	39
7.3.5	<i>iSCSI SCSI Command Processing</i> .....	39
7.3.5.1	iSCSI SCSI Command.....	39
7.3.5.2	iSCSI SCSI Command Read Operation.....	46
7.3.5.3	iSCSI SCSI Command Write Operation.....	50
7.3.5.4	Abort a SCSI Command.....	54
7.3.5.5	Debugging SCSI Command problem.....	55
7.3.6	<i>iSCSI Task Management Functions</i> .....	55
	<b>TARGET INFORMATION AND STATISTICS.....</b>	<b>58</b>
8.1	TARGET INFORMATION.....	60
8.2	SESSION INFORMATION.....	60
8.3	CONNECTION INFORMATION.....	60
8.4	PORTAL PERFORMANCE DATA.....	61
	<b>CHELSIO ISCSI TARGET API FILES.....</b>	<b>62</b>
1	<b>INTRODUCTION.....</b>	<b>51.1</b>
	AUDIENCE.....	51.2
	ABOUT THIS DOCUMENT.....	51.3
	OTHER RELATED DOCUMENTS.....	51.4
	DOCUMENT REVISION HISTORY.....	52
	<b>INTRODUCTION.....</b>	<b>72.1</b>
	<b>72.2.....</b>	<b>ISCSI73</b>
	<b>ISCSI ACCELERATION WITH CHELSIO'S TERMINATOR ASIC.....</b>	<b>103.1</b>
	<b>OFFLOAD ENGINE.....</b>	<b>103.2</b>

<b>CHELSIO ULP ISCSI ACCELERATION.....</b>	<b>104</b>
<b>CHELSIO ISCSI TARGET MODULE.....</b>	<b>124.1ISCSI</b>
<b>SOFTWARE DESIGN.....</b>	<b>134.2ISCSI</b>
<b>PROCESSING CONTEXT.....</b>	<b>144.2.1</b>
<b>INTERFACE FUNCTIONS.....</b>	<b>145</b>
<b>REGISTER WITH ISCSI TARGET MODULE.....</b>	<b>155.1ISCSI</b>
<b>TARGET CLASS.....</b>	<b>155.1.1</b>
<b>TARGET CLASS IDENTIFICATION.....</b>	<b>165.1.2</b>
<b>INITIATOR LOGIN INITIALIZATION.....</b>	<b>165.1.3</b>
<b>LOGIN PHASE CHECK.....</b>	<b>185.1.4</b>
<b>RETRIEVING CHAP INFORMATION.....</b>	<b>195.1.5ISCSI</b>
<b>TARGET SESSION MANAGEMENT.....</b>	<b>195.1.6</b>
<b>TARGET DISCOVERY.....</b>	<b>205.1.7</b>
<b>TARGET REDIRECTION.....</b>	<b>215.2ISCSI</b>
<b>TARGET LUN CLASS.....</b>	<b>215.2.1</b>
<b>TARGET LUN CLASS IDENTIFICATION.....</b>	<b>235.2.2</b>
<b>TARGET LUN CLASS PROPERTY.....</b>	<b>235.2.2.1</b>
<b>LUN_CLASS_SCSI_PASS_THRU_BIT.....</b>	<b>235.2.2.2</b>
<b>LUN_CLASS_MULTI_PHASE_DATA_BIT.....</b>	<b>235.2.2.3</b>
<b>LUN_CLASS_HAS_CMD_QUEUE_BIT.....</b>	<b>235.2.3</b>
<b>TARGET LUN CONFIGURATION FUNCTIONS.....</b>	<b>245.2.4</b>
<b>HANDLING OF ISCSI SCSI COMMAND.....</b>	<b>246ISCSI</b>
<b>TARGET CONFIGURATION.....</b>	<b>276.1</b>
<b>TARGET CONFIGURATION VIA KERNEL APIs.....</b>	<b>276.1.1</b>
<b>ADDING AN ISCSI TARGET NODE.....</b>	<b>276.1.2</b>
<b>REMOVING AN ISCSI TARGET NODE.....</b>	<b>286.1.3</b>
<b>MODIFY AN ISCSI TARGET NODE'S SETTINGS.....</b>	<b>286.1.3.1CHAP</b>
<b>CONFIGURATION.....</b>	<b>297ISCSI</b>
<b>TARGET PROCESSING.....</b>	<b>317.1ISCSI</b>
<b>SESSION.....</b>	<b>317.2ISCSI</b>
<b>DISCOVERY SESSION.....</b>	<b>317.3ISCSI</b>
<b>NORMAL SESSION.....</b>	<b>327.3.1THE</b>
<b>LOGIN PHASE.....</b>	<b>327.3.1.1FIRST</b>
<b>LOGIN REQUEST.....</b>	<b>327.3.1.2</b>
<b>LOGIN STAGE TRANSITION.....</b>	<b>347.3.1.3CHAP</b>
<b>PROCESSING.....</b>	<b>347.3.2</b>
<b>SESSION ADDITION.....</b>	<b>357.3.3</b>
<b>SESSION REMOVAL.....</b>	<b>357.3.4</b>
<b>SESSION ABORT.....</b>	<b>367.3.5ISCSI</b>
<b>SCSI COMMAND PROCESSING.....</b>	<b>367.3.5.1ISCSI</b>
<b>SCSI COMMAND.....</b>	<b>367.3.5.2ISCSI</b>
<b>SCSI COMMAND READ OPERATION.....</b>	<b>437.3.5.3ISCSI</b>
<b>SCSI COMMAND WRITE OPERATION.....</b>	<b>467.3.5.4</b>
<b>ABORT A SCSI COMMAND.....</b>	<b>487.3.5.5</b>
<b>DEBUGGING SCSI COMMAND PROBLEM.....</b>	<b>497.3.6ISCSI</b>
<b>TASK MANAGEMENT FUNCTIONS.....</b>	<b>498</b>

<b>TARGET INFORMATION AND STATISTICS.....</b>	<b>528.1</b>
<b>TARGET INFORMATION.....</b>	<b>548.2</b>
<b>SESSION INFORMATION.....</b>	<b>548.3</b>
<b>CONNECTION INFORMATION.....</b>	<b>548.4</b>
<b>PORTAL PERFORMANCE DATA.....</b>	<b>559</b>
<b>CHELSIO iSCSI TARGET API FILES.....</b>	<b>56</b>
<b>CHELSIO iSCSI SOFTWARE FEATURES.....</b>	<b>57</b>
<b>TEXT KEY SETTINGS.....</b>	<b>APPENDIX C.</b>
<b>CHELSIO TARGET CONFIGURATION KEYS.....</b>	<b>59APPENDIX D.LOG</b>
<b>AND DEBUG MESSAGES.....</b>	<b>60APPENDIX E.iSCSI</b>
<b>ACRONYMS.....</b>	<b>627</b>
<b>7.....</b>	<b>60</b>

# 1 Introduction

This document describes the Chelsio iSCSI Target APIs to be used with 3<sup>rd</sup>-party Storage Driver. The Chelsio iSCSI Target APIs enables the Storage Driver to communicate to the iSCSI Target Module.

- Section 2 introduces the iSCSI system.
- Section 3 explains the Terminator ASIC's ULP Acceleration and its APIs.
- Section 4 overviews Chelsio iSCSI Target Module Software.
- Section 5 defines how the Storage Driver could register with the Target by creating new iSCSI Target Class and iSCSI Target LUN class.
- Section 6 specifies how to add, remove and modify an iSCSI Target Node.
- Section 7 describes the exported APIs in the iSCSI Target processing flow.
- Section 8 describes the exported API for iSCSI Target information and Statistics
- Section 9 shows the source file layout for the data structures and the APIs described in this document.

## 1.1 Audience

The audience for the Chelsio iSCSI Target API Guide is primarily engineers who want to utilize the Chelsio iSCSI Target Module to support any specialized storage technology as the backend storage for iSCSI Targets.

## 1.2 About This Document

This manual contains the details of the data structures and kernel APIs exported by the Chelsio iSCSI Target Module.

## 1.3 Other Related Documents

For additional information about the Chelsio iSCSI Software and the Chelsio Terminator ASIC, refer to the following related documents listed below:

- ***Internet Small Computer Systems Interface (iSCSI)***, IETF RFC-3720, April 2004.
- ***PPP Challenge Handshake Authentication Protocol (CHAP)***, IETF RFC-1994, August 1996.
- ***iSCSI Software Installation and User Guide***, Chelsio communications.

## 1.4 Document Revision History

Revision	Comment	Date
0.1	Initial version.	07/16/2008
0.5	Updated read/write walkthrough and state diagram from multi-phase	06/02/2009

	data buffer handling	
0.6	Changed fp_chap_info_get function from blocking to non-blocking.	09/17/2009
0.7	Add support of multiple outstanding buffers for multi-phase data buffer handling. Removed requirement that only one outstanding buffer per phase.	09/29/2009
0.8	Add support for buffer overflow and underflow	06/23/2010
0.9	Add more clarification and examples in the buffer setup section as per customer request	07/22/2010
0.9.5	Updated tcp_endpoint structure with new iscsi_tcp_endpoints	02/01/2011
1.0	Update fp_first_login_check_done API	09/09/2011
1.1	Added a return value for fp_session_added(), and pass the value as part of the iscsi_scsi_command	08/22/2013
1.2	Added the private session handle to fp_tmf_execute(). Build #706	10/17/2013
1.3	Remove reference to t3	03/10/2015
1.4	Added details on target process context and API functions.	04/21/2015
1.5	prefix data structures and functions with ch(elsio)_.	06/12/2015

## 2 iSCSI Introduction

iSCSI (Internet Small Computer System Interface, also commonly referred to as Internet SCSI or “SCSI over IP”) is a storage networking standard that enables the transport of block-level I/O data over an IP network.

### 2.1 SCSI

The Small Computer Systems Interface (SCSI) is a popular family of protocols for communicating with I/O devices, especially storage devices.

There are two types of devices in the SCSI protocol; the SCSI *Initiators* (*clients*) start the communications and the *Targets* (*servers*) responds. The initiators are devices that request commands be executed. Targets are devices that carry out the commands.

A Command Descriptor Block (CDB) is used to carry a command from an application client to a device server. SCSI command execution results in an optional data phase and a status phase. In the data phase, data travels either from the initiator to the target as in a WRITE command or from the target to the initiator as in a READ command. In the status phase, the target returns the final status of the operation. The status response terminates an SCSI command or task.

### 2.2 iSCSI

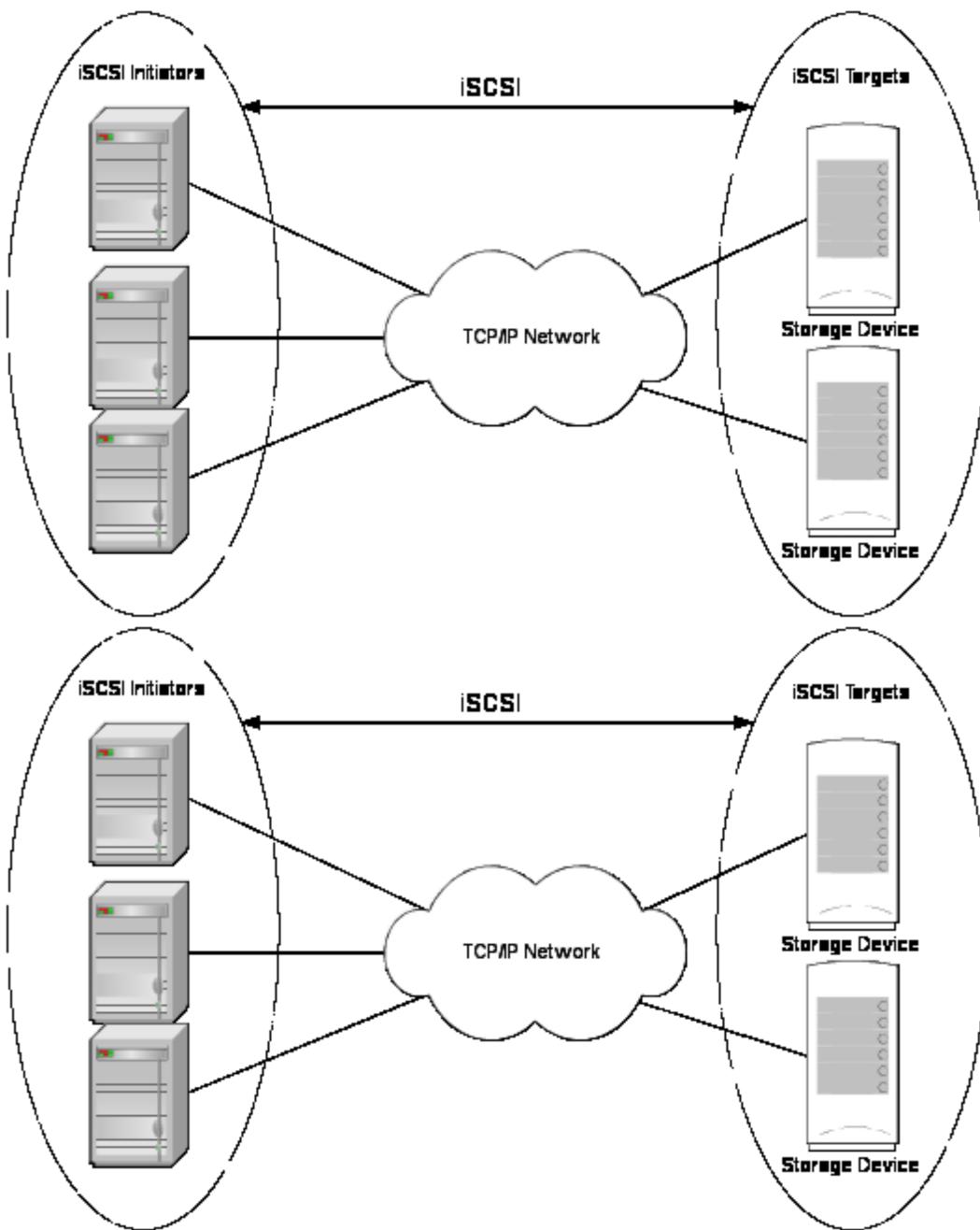
iSCSI replaces SCSI’s direct-attached cabling architecture with a network fabric. It works by encapsulating SCSI commands into packets and transporting them via TCP connections. This allows for the deployment of storage networks over a commodity IP network

The iSCSI protocol follows the SCSI’s client-server model. On opposing ends of the network are the iSCSI initiator client and the iSCSI target server.

The main responsibilities of the initiator are to establish a connection to an iSCSI target and to forward the SCSI Commands and transferring data to the target.

The iSCSI target’s primary functions are to broker the requests issued by the initiator to the physical storage and to forward the storage’s response back to the initiator. Essentially, the iSCSI target acts as the bridge between the network and the storage devices such as disks, or RAID arrays.

The figure below depicts an iSCSI network between multiple initiators and targets interconnected through a TCP/IP network.



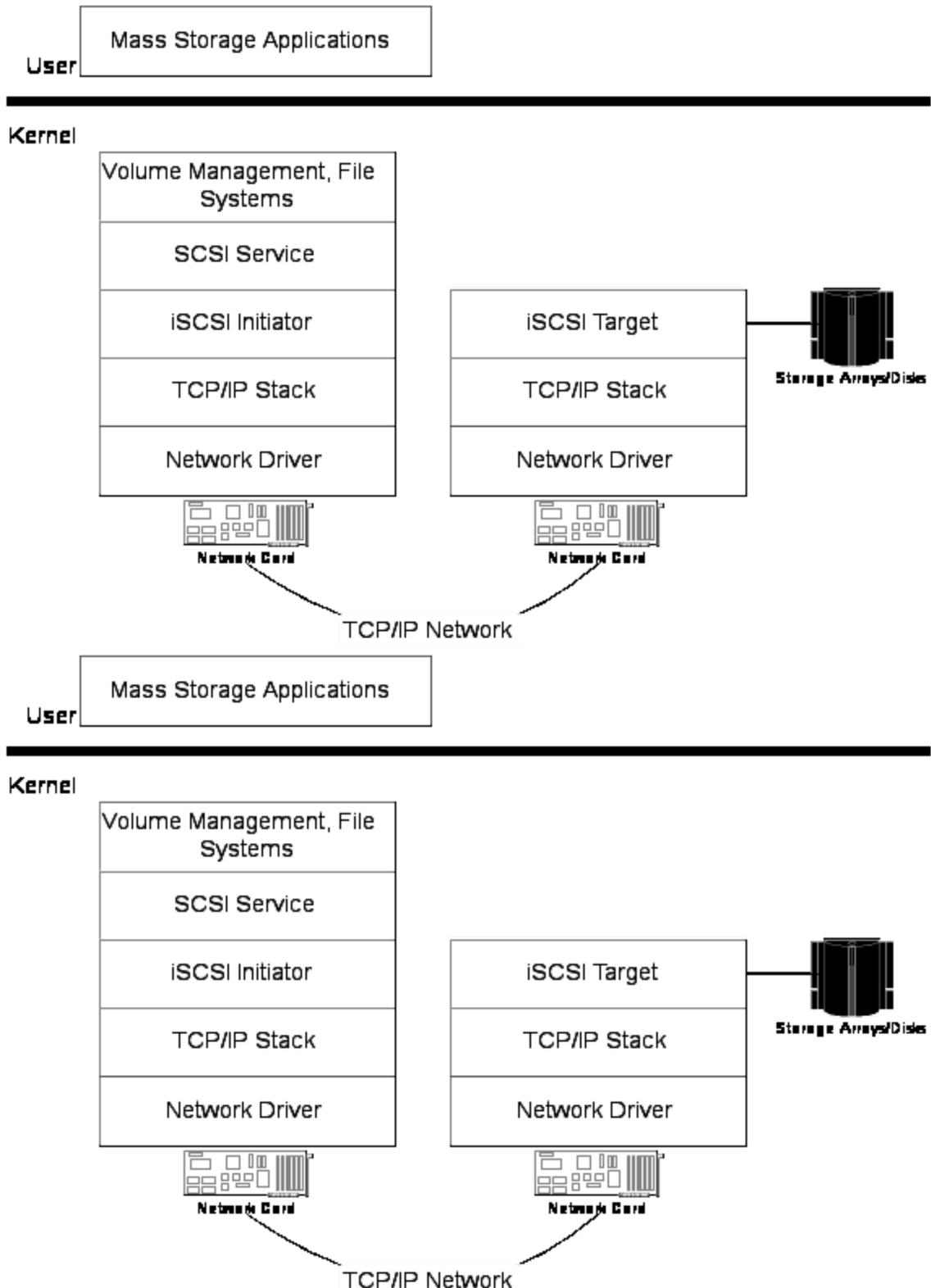
**Figure 2-1. iSCSI Network**

The messages exchanged between the Initiator and the target are called iSCSI Protocol Data Units (PDUs). They are used to carry SCSI and iSCSI commands from the Initiator to the Target, to transfer data in either direction, and to return responses from the Target to the Initiator.

Because iSCSI storage is presented to the OS as though it is attached locally, rather than as a network share, iSCSI's compatibility with existing software applications is practically guaranteed. It offers great flexibility as users can use the OS's native file

---

system on the remote storage devices over any IP network infrastructure just like any other local block-level device.



**Figure 2-1. iSCSI System Overview**

---

The simplified illustration above shows the software and hardware elements involved in an iSCSI system:

- The iSCSI Target exports the locally attached storage devices to iSCSI initiators. It receives and processes the SCSI requests from iSCSI initiators and delivers the responses back to the initiators.
- The iSCSI Initiator presents the remote storage devices exported by an iSCSI target to the local SCSI subsystem as if they are locally-attached SCSI devices.
- The iSCSI Initiator and Target communicates on top of TCP/IP network.

## 3 iSCSI Acceleration With Chelsio's Terminator ASIC

Chelsio's Terminator ASIC delivers full TCP offload and provides acceleration to iSCSI's CPU-intensive operations (such as PDU recovery, CRC digest computation and verification, Direct Data Placement into host memory buffers) across the PCI bus.

### 3.1 TCP Offload Engine

The Terminator's full TCP/IP offload engine handles the connection establishment and teardown, as well as all the exceptions including retransmission and re-ordering in the hardware.

Comparing with the TCP stacks resides in the host the Terminator's offload engine provides much higher performance due to the large reduction in packet and interrupt rate, and zero copy capability, all without requiring any application changes.

The Terminator TCP offload engine also offers a unique zero-copy "direct data placement" feature. It allows dramatic reductions not only in CPU load, but also in memory subsystem usage, which is oftentimes the actual bottleneck in today's systems.

In addition, hardware based retransmission and re-ordering virtually eliminate the impact of packet loss on performance, transforming the normally lossy Ethernet network into a high performance reliable fabric.

### 3.2 Chelsio ULP iSCSI Acceleration

In addition to offloading TCP processing from the host processor, the Terminator ASIC also supports iSCSI Upper Layer Protocol (ULP) acceleration and iSCSI Direct Data Placement (DDP) where the hardware handles the expensive byte touching operations, such as CRC computation and verification.

- **iSCSI PDU Digest Generation and Verification**

On transmitting, Terminator computes and inserts the Header Digest and the Data Digest into the iSCSI PDUs.

On receiving, the Terminator computes and verifies the Header Digest and the Data Digest of the iSCSI PDUs.

- **Direct Data Placement (DDP)**

---

Terminator can directly place the iSCSI Data-In or Data-Out PDU's payload into pre-posted final-destination host-memory buffers based on the Initiator Task Tag (ITT) or Target Task Tag (TTT) information contained in the PDU header. The "Page Pod" concept is used to manage the host memory buffers and the tags.

- **PDU Transmit and Recovery**

On transmitting, Terminator accepts the complete PDU (header + data) from the host driver, computes and inserts the digests, decomposes the PDU into multiple TCP segments if necessary, and transmits all the TCP segments onto the wire.

On receiving, Terminator recovers the iSCSI PDU by reassembling TCP segments, separating the header and data, calculating and verifying the digests, then forwards the header to the host. The payload data, if possible, will be directly placed into a pre-posted host DDP buffer. Otherwise, the payload data will be sent to the host too.

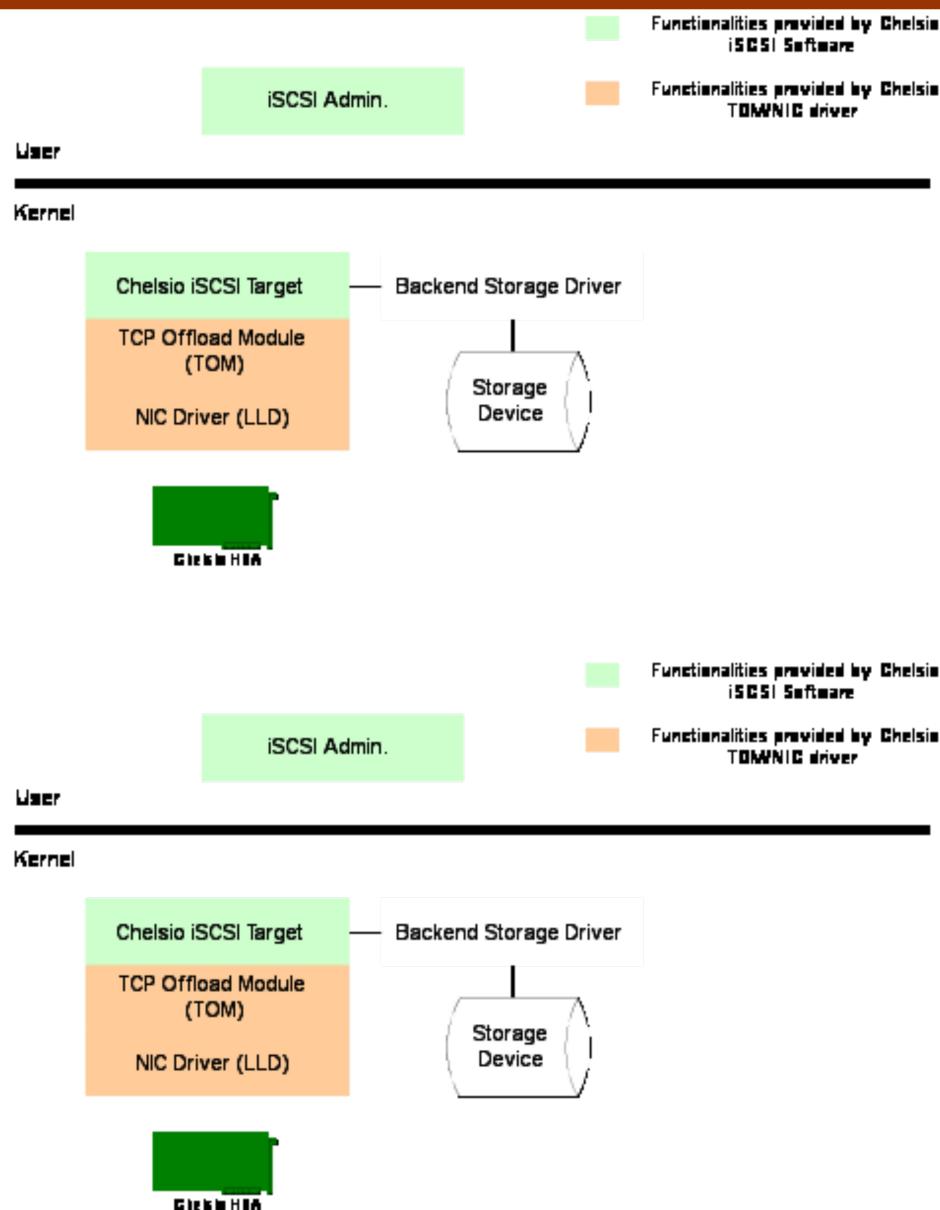
**NOTE:**

- The Terminator ASIC is unaware of iSCSI login, iSCSI sessions and other finer details of the iSCSI protocol. However, the Terminator is aware of TCP connections that carry iSCSI traffic and it performs the above functions for any TCP connection labeled with the iSCSI attribute.

## 4 Chelsio iSCSI Target Module

Chelsio's iSCSI Target Module implements the iSCSI target-mode protocol stack. It comes with the full support of Chelsio's Terminator ASIC, taking advantage of both the full TCP offload support and the ULP iSCSI acceleration.

The Chelsio's iSCSI Software can also operate on top of host's native TCP/IP stack over any third-party NIC.



**Figure 4-1. Chelsio iSCSI Target Module Overview**

The figure above shows a simplified view of the hardware and software components involved:

- The Chelsio protocol-offload Host Bus Adapters (HBAs) are installed in the target systems. These boards use the Terminator ASICs for offloading TCP/IP and accelerating iSCSI.

- The Chelsio TCP offload module (TOM) is a peer to the hosts' native TCP/IP stack. Used by the Chelsio iSCSI Target Module to offload TCP/IP processing, it enables the Host to bypass the native TCP/IP stack, and send and receive TCP data as a byte stream.
- The Chelsio Low-Level Driver (LLD) is the device driver for Chelsio's protocol engine.

**NOTE:**

1. The Chelsio's TCP offload module (TOM) and Low Level Driver (LLD) are outside of the scope of this document. For more details on these modules please refers to the related TOE driver documents.

## 4.1 iSCSI Software Design

The Chelsio iSCSI Target Module is designed with the following goals in mind:

- **Performance**

Highly optimized data path to utilize the full potential of Chelsio Terminator ASIC's offloading capabilities (TCP and iSCSI); specialized functions may be developed for better performance.

- **Inter-operability**

To the local host system, Chelsio iSCSI Target Module behaves in the exact same way as traditional direct-attached storage systems. No change is required for the operating system or high-level applications.

- **Storage Management**

The iSCSI Target Module provides target device virtualization. It comes with support for block devices (SCSI, IDE, Serial-ATA disk drives, etc) to be used easily. And it also provides abilities to utilize disk management features like Logical Volume Manager (LVM) and software RAID for flexible storage management and redundancy.

In addition, the iSCSI Target Module exports a set of kernel API to enable communications with any Storage Driver thus expanding the supports of additional target device types.

- **Maintainability**

As various operating systems are continue to evolve, the Chelsio iSCSI software tries to keep the design simple by using as much high-level interfaces as possible. The usage of well-tested, high-level standard OS/kernel interfaces has clear advantages in terms of stability, implementation and maintenance costs, and portability.

## 4.2 iSCSI Processing Context

The Target Module utilizes the kernel thread for the iSCSI protocol processing:

- 1 main thread

The main thread handles the iSCSI connections in the Login Phase.

- N worker threads (where N = # of logical CPUs in the host system)

The worker thread handles the iSCSI connections in the Full Feature Phase (FFP).

Once an iSCSI connection completes the login phase, the connection is handed from the main thread to one of the worker threads for continuing processing.

### 4.2.1 Interface functions

For the functions that the Storage Driver provides they will be invoked by the Target Module in one of the kernel thread context. Because these functions are executed in the Target Module kernel thread context, they are expected to be either as

- Blocking

The function is expected to complete its functionality entirely during the call and no callback mechanism into the Target Module is provided.

Most of the control plane functions fall into this category.

- Non-blocking

The function should do minimum processing and return the control to the Target Module as soon as possible. It is expected the bulk of work should be done by the Storage Driver outside of the Target Module kernel thread context. A callback mechanism into the Target Module is provided.

All data plane functions and some of the control plane functions fall into this category.

For the functions that are exported by the Target Module (i.e., called by the Storage Driver), similar to the non-blocking functions, they will perform minimum house-keeping and then wake up the related Target kernel thread to continue the iSCSI processing. Because the target kernel thread could become active before the function returns, there is a small window that the Target thread would call into one of the Storage Driver provided functions at the same time.

# 5 Register with iSCSI Target Module

To be able to communicate with the Target Module, the Storage Driver should first create and register a new class of iSCSI target and target LUN.

## 5.1 iSCSI Target Class

To add or remove a new class of iSCSI target, the Storage Driver should use the following exported APIs:

- `int chiscsi_target_class_register (chiscsi_target_class *target_class);`
- `int chiscsi_target_class_deregister (char *class_name);`

The structure `chiscsi_target_class` provides a series of hooks in various points in the iSCSI target protocol processing stack.

```
typedef struct chiscsi_target_class chiscsi_target_class;

struct chiscsi_target_class {
    ...
    char             *class_name;

    void (*fp_first_login_check) (unsigned long hndl,
                                char          *initiator_name,
                                char          *target_name,
                                chiscsi_tcp_endpoint *endpoint);

    void (*fp_login_stage_check) (unsigned long hndl,
                                unsigned char login_stage,
                                char          *initiator_name,
                                char          *target_name,
                                chiscsi_tcp_endpoint *endpoint);

    void (*fp_chap_info_get) (char          *initiator_name,
                            char          *target_name,
                            chap_info     *chap);

    unsigned long (*fp_session_added) (unsigned long sess_hdl,
                                      unsigned char isid[6],
                                      char          *initiator_name,
                                      char          *target_name);

    void (*fp_session_removed) (unsigned long sess_hdl,
                               char          *initiator_name,
                               char          *target_name);

    int (*fp_discovery_target_accessible) (unsigned long hndl
                                         char          *initiator_name,
                                         char          *target_name,
```

---

```

        chiscsi_tcp_endpoint *endpoint);
int (*fp_select_redirection_portal)(char      *target_name,
                                    char      *initiator_name,
                                    chiscsi_tcp_endpoint *endpoint);
};

}

```

### 5.1.1 Target Class Identification

Each iSCSI target class is identified with a unique character string “*class\_name*”. It must not be NULL.

### 5.1.2 Initiator Login Initialization

```

void (*fp_first_login_check) (      unsigned long  hndl,
                                    char          *initiator_name,
                                    char          *target_name,
                                    chiscsi_tcp_endpoint *endpoint);

```

where

- *hndl* is a value uniquely identifies the iSCSI connection
- *initiator\_name* points to the initiator name string
- *target\_name* points to the target name string
- *endpoint* contains the tcp 4-tuple.

The structure *chiscsi\_tcp\_endpoints* represents the tcp 4-tuples, and is defined as:

```

typedef struct chiscsi_tcp_endpoints      chiscsi_tcp_endpoints;
struct chiscsi_tcp_endpoints {
    unsigned int          f_ipv6:1;
    unsigned int          f.filler:31;
    struct tcp_endpoint   iaddr;           /* initiator tcp address */
    struct tcp_endpoint   taddr;           /* target tcp address */
    unsigned int          port_id;         /* for Chelsio HBA only */
};

```

The structure *tcp\_endpoints* represents the tcp endpoint, and is defined as:

```

#define ISCSI_IPADDR_LEN      16
struct tcp_endpoint {
    unsigned char  ip[ISCSI_IPADDR_LEN];
    unsigned int    port;
};

```

If the Storage Driver wants to be informed of an Initiator login attempt for a normal session, it could set the *fp\_first\_login\_check()*. This hook is the ideal place for actions such as

- preparing the target for a possible new iSCSI session and/or
- checking if the target resource is available and/or

- performing iSCSI Initiator access control

When the iSCSI protocol stack receives the first Login Request PDU for a normal session from a newly accepted TCP connection, it invokes *fp\_first\_login\_check()*.

The function *fp\_first\_login\_check()* is non-blocking. The Target Module will halt the processing of the particular connection until the Storage Driver notifies the Target Module via the exported API *chiscsi\_target\_first\_login\_check\_done()*.

```
void chiscsi_target_first_login_check_done(
    unsigned long hndl,
    unsigned char login_status_class,
    unsigned char login_status_detail,
    unsigned int max_cmd);
```

where

- *hndl* is the value uniquely identified the iSCSI connection, which is passed in via *fp\_first\_login\_check()*.
- *login\_status\_class* and *login\_status\_detail* are used to reject a login request.
- *max\_cmd* is the maximum number of outstanding SCSI commands this target can handle for the new session. If it is zero then the iSCSI target protocol stack will default to maximum of 128 outstanding SCSI commands per session.

*login\_status\_class* and *login\_status\_detail* should both be set to zero if the Storage Driver completes its processing successfully and wants the login phase to proceed.

In order to use Redirection, storage driver should set the *login\_status\_class* and *login\_status\_detail*, both to zero. Chelsio Target driver will treat Redirection status class as Success.

In the case of error, *login\_status\_class* should be set to one of the following values:

- #define ISCSI\_LOGIN\_STATUS\_CLASS\_INITIATOR\_ERROR 2
- #define ISCSI\_LOGIN\_STATUS\_CLASS\_TARGET\_ERROR 3

If *login\_status\_class* = *ISCSI\_LOGIN\_STATUS\_CLASS\_INITIATOR\_ERROR* then the possible values for *login\_status\_detail* are:

- #define ISCSI\_LOGIN\_STATUS\_DETAIL\_INIT\_ERROR 0
- #define ISCSI\_LOGIN\_STATUS\_DETAIL\_NO\_PERMS 2
- #define ISCSI\_LOGIN\_STATUS\_DETAIL\_TARGET\_NOT\_FOUND 3
- #define ISCSI\_LOGIN\_STATUS\_DETAIL\_TARGET\_REMOVED 4
- #define ISCSI\_LOGIN\_STATUS\_DETAIL\_INVALID\_REQUEST 11

If *login\_status\_class* = *ISCSI\_LOGIN\_STATUS\_CLASS\_TARGET\_ERROR* then *login\_status\_detail* should be set to one of the following values:

- #define ISCSI\_LOGIN\_STATUS\_DETAIL\_TARG\_ERROR 0

- 
- |                                                     |   |
|-----------------------------------------------------|---|
| • #define ISCSI_LOGIN_STATUS_DETAIL_SERVICE_UNAVAIL | 1 |
| • #define ISCSI_LOGIN_STATUS_DETAIL_NO_RESOURCES    | 2 |

### 5.1.3 Login Phase Check

The iSCSI login phase could proceed in two stages:

- |                                          |   |
|------------------------------------------|---|
| 1. #define ISCSI_LOGIN_STAGE_SECURITY    | 0 |
| 2. #define ISCSI_LOGIN_STAGE_OPERATIONAL | 1 |

In the login security stage the Initiator and the Target negotiate the authentication method (defined by iSCSI key “*AuthMethod*”) and perform the authentication if needed.

The login operational stage consists of text string negotiation of operating parameters using key=value pairs of login parameter exchanges.

*fp\_login\_stage\_check()* is invoked when the login security stage and the login operational stage are first entered for the leading connection of a normal session. This hook is the ideal place to perform initiator access control.

```
void (*fp_login_stage_check)(  
                           unsigned long hndl,  
                           unsigned char login_stage,  
                           char *initiator_name,  
                           char *target_name,  
                           chiscsi_tcp_endpoint *endpoint);
```

where

- *hndl* is a value uniquely identifies the iSCSI connection
- *login\_stage* is either *ISCSI\_LOGIN\_STAGE\_SECURITY* or *ISCSI\_LOGIN\_STAGE\_OPERATIONAL*.
- *initiator\_name* points to the initiator name string
- *target\_name* points to the target name string
- *endpoint* contains the tcp 4-tuple.

*fp\_login\_stage\_check()* is non-blocking. The Target Module will halt the processing of the connection until the Storage Driver informs the target the checking has been completed via the exported API *chiscsi\_target\_login\_stage\_check\_done()*.

```
void chiscsi_target_login_stage_check_done(  
                                         unsigned long hndl,  
                                         unsigned char login_status_class,  
                                         unsigned char login_status_detail);
```

where

- *hndl* is the value uniquely identified the iSCSI connection, which is passed in via *fp\_login\_stage\_check()*.
- *login\_status\_class* and *login\_status\_detail* are used to reject a login request.

## 5.1.4 Retrieving CHAP Information

If the Storage Driver manages its own CHAP database, it should set up *fp\_chap\_info\_get()*. This function is used by the Target Module to retrieve CHAP names and secrets from the Storage Driver during the login phase.

```
void (*fp_chap_info_get) (char *initiator_name,
                         char *target_name,
                         chap_info *chap);
```

where

- *initiator\_name* points to the initiator name string
- *target\_name* points to the target name string
- *chap* contains the CHAP name and secret.

The *chap\_info* structure is used to pass CHAP-related information from the Storage Driver to the Target Module.

```
typedef struct chap_info chap_info;

#define CHAP_FLAG_LOCAL_NAME_VALID 0x1
#define CHAP_FLAG_LOCAL_SECRET_VALID 0x2
#define CHAP_FLAG_REMOTE_NAME_VALID 0x4
#define CHAP_FLAG_REMOTE_SECRET_VALID 0x8
#define CHAP_FLAG_MUTUAL_REQUIRED 0x10

struct chap_info {
    unsigned char flag;
    unsigned char local_secret_length;
    unsigned char remote_secret_length;
    unsigned char filler;
    unsigned char local_secret[16];
    unsigned char local_name[256];
    unsigned char remote_secret[16];
    unsigned char remote_name[256];
};
```

## 5.1.5 iSCSI Target Session Management

If the Storage Driver wants to be informed of the session creation and termination it should set up *fp\_session\_added()* and *fp\_session\_removed()*.

The target module calls *fp\_session\_added()* whenever a session goes into service (i.e., its leading connection transits to the Full Feature Phase).

The function *fp\_session\_removed()* is called by the Target Module whenever a normal session has been terminated.

```

unsigned long (*fp_session_added)(unsigned long sess_hdl,
                                  unsigned char isid[6],
                                  char *initiator_name,
                                  char *target_name);
void (*fp_session_removed)(unsigned long sess_hdl,
                           char *initiator_name,
                           char *target_name);

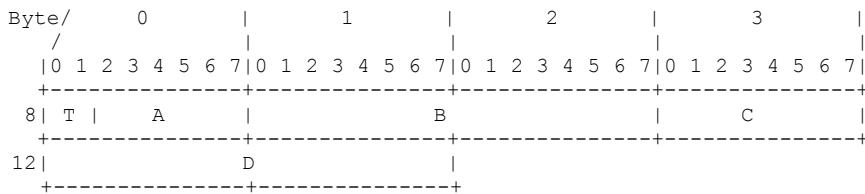
```

where

- *hdl* is the target session handle.
- *isid* is the initiator session id.
- *initiator\_name* points to the initiator name string
- *target\_name* points to the target name string

The ISID is a unique identifier that an Initiator assigns to its end point of the session. When combined with the iSCSI Initiator Name, it uniquely identifies its SCSI Initiator Port.

According to RFC3720 the ISID is a six-byte value:



For *fp\_session\_added()*, the Storage Driver can choose to pass back a unsigned value (or a casted pointer) to be associated with the target session handle *hdl*.

### 5.1.6 Target Discovery

*fp\_discovery\_target\_accessible()* should be set if the Storage Driver wants to control what targets an initiator should be able to see.

```

int (*fp_discovery_target_accessible)(
    unsigned long hndl,
    char *initiator_name,
    char *target_name,
    isccsi_tcp_endpoint *endpoint);

```

where

- *hdl* is a value uniquely identifies the iSCSI connection
- *initiator\_name* points to the initiator name string
- *target\_name* points to the target name string
- *endpoint* contains the tcp 4-tuple.

*fp\_discovery\_target\_accessible()* is a blocking call. The Storage Driver should return zero (0) if it does not want the target named “*target\_name*” to be seen by the initiator

named “*initiator\_name*” or returns one (1) otherwise. It should return a value less than zero (<0) in the case of error.

### 5.1.7 Target Redirection

An iSCSI Target can redirect an initiator to use a different IP address and port (often called a portal) instead of the current one. The redirected-to target portal can either be on the same machine, or on another system.

Chelsio configuration key *ShadowMode* is used to distinguish whether the redirected target portals are on a different system or not. If *ShadowMode*=Yes, then the redirected-to target portals are on another system.

To configure target redirection, specify the redirected-to target portals, and add their portal group tag to local listening portal. In the following example, any login requests received on *10.193.184.81:3260* will be redirected to *10.193.184.85:3261*:

```
PortalGroup=2@10.193.184.85:3261
PortalGroup=1@10.193.184.81:3260,[2]
```

*fp\_select\_redirection\_portal()* should be set if the Storage Driver wants to control which redirected-to portal the initiator should be connected to.

```
int (*fp_select_redirection_portal)( char *target_name,
                                     char *initiator_name,
                                     chiscsi_tcp_endpoint *endpoint);
```

where

- *target\_name* points to the target name string
- *initiator\_name* points to the initiator name string
- *endpoint* contains the tcp 4-tuple.

*fp\_select\_redirection\_portal()* is a blocking call.

The Storage Driver should fill the field *endpoint->taddr* as the redirected-to portal and return zero (0). It should return a value less than zero (<0) in the case of error.

## 5.2 iSCSI Target LUN Class

An *chiscsi\_target\_lun\_class* defines a LUN type and its behavior for a particular *chiscsi\_target\_class*.

To relate a LUN class with an existing iSCSI target class uses the following exported API:

- int chiscsi\_target\_lun\_class\_register(
 chiscsi\_target\_lun\_class \*lun\_class,

- 
- int chiscsi\_target\_lun\_class\_deregister(  
          char                               \*target\_class\_name);  
          char          \*lun\_class\_name,  
          char          \*target\_class\_name);

A *chiscsi\_target\_class* must contain at least one *chiscsi\_target\_lun\_class*.

An iSCSI target node can accommodate multiple LUNs. Similarly, a *chiscsi\_target\_class* can have multiple *chiscsi\_target\_lun\_class* registered to it.

```
enum lun_class_property_bits {
    LUN_CLASS_SCSI_PASS_THRU_BIT,
    LUN_CLASS_MULTI_PHASE_DATA_BIT,
    LUN_CLASS_HAS_CMD_QUEUE_BIT, /* LU maintains its own queue, this
                                  should be the default mode, so it
                                  won't exe. in the iscsi target stack
                                  context */
};

typedef struct chiscsi_target_lun_class chiscsi_target_lun_class;

struct chiscsi_target_lun_class {
    ...
    char                               *class_name;
    unsigned int                       property;

    int (*fp_config_parse_options) (unsigned long hndl,
                                    char              *buf,
                                    int               buflen,
                                    char              *ebuf);

    int (*fp_attach)                  (unsigned long hndl,
                                    char              *ebuf,
                                    int               ebuflen);

    void (*fp_detach)                (unsigned long hndl);

    int (*fp_reattach)               (unsigned long hndl,
                                    unsigned long hndl_new,
                                    char              *ebuf,
                                    int               ebuflen);

    int (*fp_flush)                  (unsigned long hndl);

    int (*fp_scsi_cmd_cdb_rcved) (chiscsi_scsi_command *scmd);

    int (*fp_scsi_cmd_data_xfer_status)(chiscsi_scsi_command *scmd);

    int (*fp_scsi_cmd_abort)        (chiscsi_scsi_command *scmd);

    void (*fp_scsi_cmd_abort_status) (unsigned int sc_lun,
                                    unsigned int sc_cmds,
                                    unsigned int sc_xfer_sgcnt,
```

```

        unsigned char *sc_xfer_sreq_buf,
        void         *sc_sdev_hdl);

int (*fp_tmf_execute)
        (unsigned long sess_tclass,
         unsigned long hndl,
         unsigned char immediate_cmd,
         unsigned char tmf_func,
         unsigned int lun,
         chisci_scsi_command *scmd);
};


```

## 5.2.1 Target LUN Class Identification

Each iSCSI target LUN class is identified with a unique character string “*class\_name*”. It must not be NULL.

## 5.2.2 Target LUN Class Property

### 5.2.2.1 LUN\_CLASS\_SCSI\_PASS\_THRU\_BIT

If the Storage Driver wants to handle all the SCSI command (encapsulated in *SCSI Command PDU*) received from the Initiator itself, it should set the *LUN\_CLASS\_SCSI\_PASS\_THRU\_BIT* in the property (i.e., *property = 1 << LUN\_CLASS\_SCSI\_PASS\_THRU\_BIT*).

### 5.2.2.2 LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT

It is preferred that the Storage Driver could supply all the data buffers for a SCSI command in one shot -- i.e., for a read command all the read data is returned together and for a write command all the write data buffers can be allocated and returned together.

If the Storage Driver wants to allocate and return the data buffers for the SCSI command in chunks (i.e., the data buffers are allocated and returned in multiple steps), it should set the *LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT* in the property (i.e., *property = 1 << LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT*).

### 5.2.2.3 LUN\_CLASS\_HAS\_CMD\_QUEUE\_BIT

It is recommended that the Storage Driver set *LUN\_CLASS\_HAS\_CMD\_QUEUE\_BIT* in the property (i.e., *property = 1 << LUN\_CLASS\_HAS\_CMD\_QUEUE\_BIT*), which means

---

the Storage Driver maintains its own queue of scsi tasks/commands and its own thread to execute these scsi commands (i.e., not in the Target Module execution context).

### 5.2.3 Target LUN Configuration Functions

There are ways to configure an iSCSI target node, its LUNs and other iSCSI specific parameters:

1. via a set of exported kernel APIs. These APIs are covered in detailed in the next chapter.
2. via the Chelsio-defined configuration file (default to */etc/chiscsi/chiscsi.conf*).

To be able to use the second method, the Chelsio-defined configuration file, to manage the iSCSI target nodes, the Storage Driver must set the following function hooks properly:

- *fp\_config\_parse\_option()* is invoked when parsing the “*TargetDevice*” key-value line in the configuration file.
- *fp\_attach()* attaches to or opens a LUN.
- *fp\_reattach()* will be called when a target node is reloaded.
- *fp\_detach()* detaches or closes a LUN.
- *fp\_flush()* flushes the data in the cache. It is called before a LUN is taken offline (i.e. being detached).

The above hooks do not need to be set if using the first method, via exported kernel APIs.

### 5.2.4 Handling of iSCSI SCSI Command

After an iSCSI session completes the Login Phase successfully, it moves into the Full Feature Phase (FFP). Once in the Full Feature Phase, an iSCSI initiator can then start to perform I/Os.

The Storage Driver should always set the following function pointers:

```
int (*fp_scsi_cmd_cdb_rcved)      (chiscsi_scsi_command *scmd);  
  
int (*fp_scsi_cmd_data_xfer_status) (chiscsi_scsi_command *scmd,  
                                    unsigned char *sc_xfer_sreq_buf,  
                                    unsigned int sc_xfer_sgcnt,  
                                    unsigned int sc_xfer_offset,  
                                    unsigned int sc_xfer buflen);  
  
int (*fp_scsi_cmd_abort)          (chiscsi_scsi_command *scmd);  
  
void (*fp_scsi_cmd_abort_status)  (unsigned int    sc_lun,
```

```

        unsigned int    sc_cmdsn,
        unsigned int    sc_itt,
        unsigned int    sc_xfer_sgcnt,
        unsigned char   *sc_xfer_sreq_buf,
        void           *sc_sdev_hdl);

int (*fp_tmf_execute)          (unsigned long sess_tclass,
                               unsigned long hndl,
                               unsigned char immediate_cmd,
                               unsigned char tmf_func,
                               unsigned int lun,
                               chisci_scsi_command *scmd);

```

*fp\_scsi\_cmd\_cdb\_rcved()* is called by the Target Module when the SCSI command CDB is received from the initiator.

*fp\_scsi\_cmd\_data\_xfer\_status()* is used by the Target Module to pass data transfer status to the Storage Driver.

*fp\_scsi\_cmd\_abort()* is used by the Target Module to abort a scsi command that has been forwarded to the Storage Driver.

When the Storage Driver initiates the request to abort a scsi command, the Target Module uses *fp\_scsi\_cmd\_abort\_status()* to inform the Storage Driver of the abort is done.

*fp\_tmf\_execute()* is called by the Target Module for the *Task Management Function Request PDUs* received.

If any private data is kept by the Storage driver for an *chisci\_scsi\_command*, the following function pointer can be set optionally:

```
void (*fp_scsi_cmd_cleanup)      (chisci_scsi_command *scmd);
```

*fp\_scsi\_cmd\_cleanup()* would be called by the Target Module right before the *chisci\_scsi\_command* structure is released. This function can be used to clean up any private data kept by the Storage driver for the particular *chisci\_scsi\_command*.

All of the above functions are non-blocking.

In order to inform the Target Module the requested operation has been completed, the Target Module exported the following callback function APIs:

```
int chisci_scsi_cmd_buffer_ready (chisci_scsi_command *scmd,
                                  unsigned char *sc_xfer_sreq_buf,
                                  unsigned int sc_xfer_sgcnt,
                                  unsigned int sc_xfer_offset,
```

```
        unsigned int sc_xfer_buflen);  
  
int chisci_scsi_cmd_execution_status(chisci_scsi_command *scmd,  
                                     unsigned char *sc_xfer_sreq_buf,  
                                     unsigned int sc_xfer_sgcnt,  
                                     unsigned int sc_xfer_offset,  
                                     unsigned int sc_xfer_buflen);  
  
int chisci_tmf_execution_done      (unsigned long hndl,  
                                     unsigned char tmf_response,  
                                     chisci_scsi_command *scmd);
```

The Storage Driver should use the following exported API to request the Target Module to abort a SCSI command:

```
int chisci_scsi_cmd_abort          (chisci_scsi_command *scmd);  
int chisci_scsi_cmd_abort_status  (chisci_scsi_command *scmd);
```

All of the SCSI command related functions mentioned above will be explained in detail in Chapter 7.

## 6 iSCSI Target Configuration

Chelsio Target Module supports two ways of configuring an iSCSI target node:

- via the configuration file. This method uses the ioctl interface to pass the configuration information from user space to kernel space.
- via exported kernel APIs. This method is preferred if the configuration management is done from the kernel space.

The following sections described the exported kernel functions by the Target Module.

### 6.1 Target Configuration via Kernel APIs

The following sections explain the exported target configuration APIs in detail.

#### 6.1.1 Adding an iSCSI Target Node

The Storage Driver should use *chiscsi\_target\_add()* to create a new iSCSI target node.

```
int chiscsi_target_add(          void *sdev_priv,  
                           char *target_name,  
                           char *target_class_name,  
                           char *config_buffer,  
                           int buflen);
```

where

- *sdev\_priv* points to any target node private data maintained by the Storage Driver.
- *target\_name* points to the iSCSI target node name string. The string should be a globally unique name.
- *target\_class\_name* is the name of the iSCSI target class this target node belongs to (i.e. *iscsi\_target\_class.class\_name*).
- *config\_buffer* should contain all of the configuration parameters in the form of *<key>=<value>*, each key-value pair should be separated by a *NULL* character, similar to the *iscsi* key-value pairs contained in the login PDUs.
- *buflen* is the length of the *config\_buffer*, including the ending *NULL* character.

Any key that is not presented in the *config\_buffer* will take the default value if applicable.

The function *chiscsi\_target\_add()* is blocking. It returns zero if the target node has been added and started successfully or less than zero (< 0) in the case of error.

The “*PortalGroup*” includes the portal group tag and a list of associated target IP address(es) and port number(s) that service the login request. *PortalGroup* is one of the Chelsio configuration keys. Its format is described in the Appendix.

Additionally, the Storage driver can also specify the following parameters in the *config\_buffer*:

- The iSCSI parameter settings.
- The Initiator Access Control List

### 6.1.2 Removing an iSCSI Target Node

The function *chiscsi\_target\_remove()* is provided to the Storage Driver to stop and remove an existing iSCSI target node.

```
int chiscsi_target_remove(          void *sdev_priv,  
                           char *target_name);
```

where

- *sdev\_priv* points to the target node private data maintained by the Storage Driver.
- *target\_name* points to the iSCSI target node name string.

The function *chiscsi\_target\_remove()* is blocking. It returns zero if the target node is stopped and removed successfully or less than zero (< 0) in the case of error.

When a target node is removed, all the sessions logged into this target will be terminated.

### 6.1.3 Modify an iSCSI Target Node's Settings

The function *chiscsi\_target\_reconfig()* could be used by Storage Driver to change the settings of an existing target node.

```
int chiscsi_target_reconfig(          void *sdev_priv,  
                               char *target_name,  
                               char *target_class_name,  
                               char *config_buffer,  
                               int buflen);
```

where

- *sdev\_priv* points to any target node private data maintained by the Storage Driver.
- *target\_name* points to the iSCSI target node name string. *target\_class\_name* is the name of the iSCSI target class this target node belongs to (i.e. *chiscsi\_target\_class.class\_name*).

- *config\_buffer* should contain all of the configuration parameters in the form of *<key>=<value>*. Each key-value pair should be separated by a *NULL* character, similar to the iSCSI key-value pairs contained in the login PDUs.
- *buflen* is the length of the *config\_buffer*, including the ending *NULL* character.

Any key that is not presented in the *config\_buffer* will take the default value if applicable.

The function is blocking, the Target Module will parse the parameter buffer (pointed by *config\_buffer*). If any text key format error is detected the Target Module will return a value less than zero (< 0). The functions will return 0 if no error is detected and the target is reconfigured successfully.

To make an iSCSI target node usable, at a minimum, it should have at least one target portal group with a minimum of one listening portal.

#### 6.1.3.1 CHAP Configuration

CHAP is a protocol that is used to authenticate the peer of a connection and uses the notion of a challenge and response, (i.e., the peer is challenged to prove its identity).

The Target Module supports Challenge-Handshake Authentication Protocol (CHAP) for normal sessions.

The iSCSI key *AuthMethod* controls if an initiator needs to be authenticated or not. It defaults to “*None,CHAP*”.

To force CHAP authentication on all the connection to a particular target node, the target node should have its “*AuthMethod*” set to “*CHAP*” only (i.e., *AuthMethod=CHAP*).

Two authentication options are available if CHAP is chosen as the authentication method:

1. One-way CHAP method (also called uni-directional CHAP), and
2. Mutual CHAP method (also called bi-directional CHAP).

With one-way CHAP the target uses CHAP to authenticate the initiator. The initiator does not authenticate the target.

With mutual CHAP, the target uses CHAP to authenticate the initiator. And the initiator also uses CHAP to authenticate the target.

The Chelsio configuration key *Auth\_CHAP\_Policy* controls which CHAP authentication (*Oneway* or *Mutual*) needs to be performed if CHAP authentication is

---

needed on a connection. It defaults to oneway CHAP (i.e.,  
*Auth\_CHAP\_Policy=Oneway*).

To force Mutual CHAP to be performed on all the connections to a target node,  
*“Auth\_CHAP\_Policy”* should be set to “*Mutual*” (i.e., *Auth\_CHAP\_Policy=Mutual*).

## 7 iSCSI Target Processing

This chapter describes the target interface and the exported APIs usage in the iSCSI target processing flows.

### 7.1 iSCSI Session

The iSCSI Protocol defines two types of sessions:

- **Normal Session**: a session in which SCSI commands, data and responses can be transferred between an iSCSI initiator and an iSCSI target.
- **Discovery Session**: a session only opened for target discovery.

To exchange SCSI commands, data, and responses, an iSCSI session must exist between the Initiator and the Target.

An iSCSI session identifies all of the connections between an initiator port and a target port. It is defined as a group of one or more TCP connections over which commands and data for a particular SCSI session are transferred.

A session may consist of one or more TCP/IP connections. Across all connections within a session, an initiator sees one and the same target. When multiple commands are striped across multiple TCP connections, the iSCSI session delivers the commands in the order in which they are received.

### 7.2 iSCSI Discovery Session

The Discovery session is used by the Initiators to find out available iSCSI targets at a given IP address.

After the initiator completes the login successfully, it will issue a *Text Request PDU* with the iSCSI key “*SendTarget*”.

If the function *fp\_discovery\_target\_accessible()* (of *chisci\_target\_class*) is defined, the Target Module will first invoke the function to query the Storage Driver if the target should be included in the discovery response.

After receiving the available target list in the *Text Response PDU* from the Target, the Initiator would terminate the session.

## 7.3 iSCSI Normal Session

The iSCSI Normal Session is required if an Initiator wants to perform SCSI I/Os from/to an iSCSI target.

### 7.3.1 The Login Phase

When the Initiator is ready to connect to a target, it first establishes a TCP connection to the Target's listening portal (defined by the Chelsio Configuration key "PortalGroup") and then it begins the iSCSI login phase.

iSCSI login is a mechanism used to authenticate the involved parties, negotiates the session's parameters, and marks the connection as belonging to an iSCSI session. The login phase must be completed on each TCP connection before it can be used to transport SCSI commands. The login process may establish a new session, or add the connection to an existing session.

If the Target wants to reject the login, the iSCSI Target Module issues a negative Login response to the Initiator, with the Status Class/Detail (standard iSCSI status code defined in the iSCSI RFC 3720) as determined by the Target System. In this immediate reject case, the iSCSI Target will terminate the TCP connection and clean up any resources associated with this connection.

#### 7.3.1.1 First Login Request

Byte/	0		1		2		3	
/								
0 1 2 3 4 5 6 7   0 1 2 3 4 5 6 7   0 1 2 3 4 5 6 7   0 1 2 3 4 5 6 7								
+-----+-----+-----+-----+-----+-----+-----+-----+								
0  . 1  0x03   T C . . CSG NSG  Version-max   Version-min								
+-----+-----+-----+-----+-----+-----+-----+-----+								
4  TotalAHSLength   DataSegmentLength								
+-----+-----+-----+-----+-----+-----+-----+								
8  ISID								
+-----+-----+-----+-----+-----+-----+-----+								
12    TSIH								
+-----+-----+-----+-----+-----+-----+-----+								
16  Initiator Task Tag								
+-----+-----+-----+-----+-----+-----+-----+								
20  CID   Reserved								
+-----+-----+-----+-----+-----+-----+-----+								
24  CmdSN								
+-----+-----+-----+-----+-----+-----+-----+								
28  ExpStatSN or Reserved								
+-----+-----+-----+-----+-----+-----+-----+								
32  Reserved								
+-----+-----+-----+-----+-----+-----+-----+								
36  Reserved								
+-----+-----+-----+-----+-----+-----+-----+								
40  Reserved							/	
+-----+-----+-----+-----+-----+-----+-----+							/	
48  DataSegment - Login Parameters in Text request Format							/	
+-----+-----+-----+-----+-----+-----+-----+							/	

### Figure 7-1. Login Request PDU Format

When the iSCSI Target Module receives the first Login Request PDU without C (continue) bit set from a newly accepted TCP connection, it first verifies

- the request is to establish a new normal session
- the requested target node exists
- the *chiscsi\_target\_class* of the requested target node has *fp\_first\_login\_check()* set

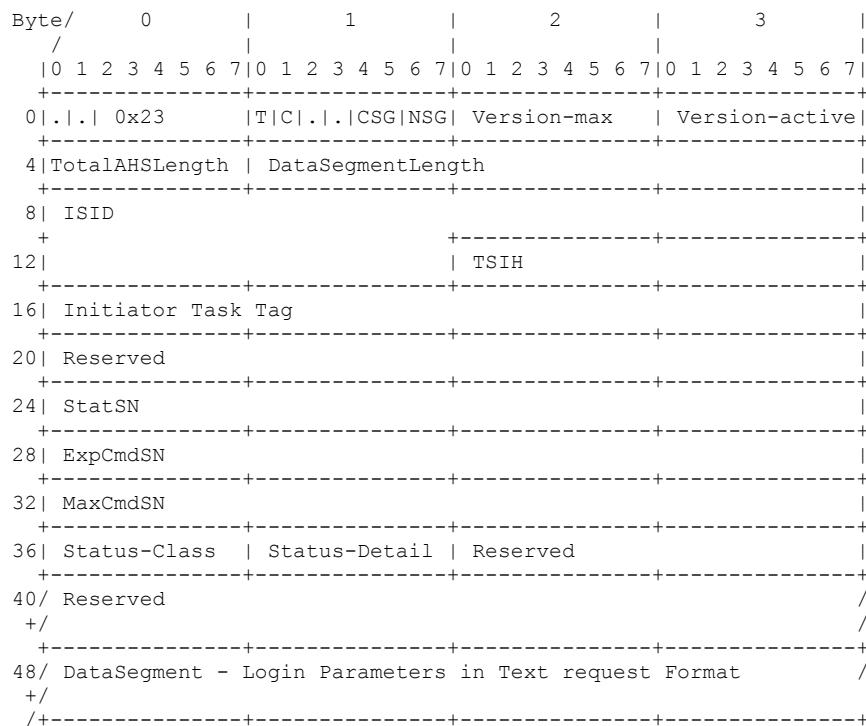
Then the Target Module invokes *fp\_first\_login\_check()*.

Because *fp\_first\_login\_check()* is assumed to be non-blocking, at this point, the Target Module will halt the login processing on this connection until the Storage Driver calls the exported API *chiscsi\_target\_first\_login\_check\_done()*.

If the Storage Driver calls *chiscsi\_target\_first\_login\_check\_done()* with a non-zero *login\_status\_class*, the Target Module will reject the Initiator's login request and terminates the underlying TCP connection.

In this case the *Login Response PDU* that is constructed and sent to the Initiator would have

- “*Status-Class*” = *login\_status\_class* and
- “*Status-Detail*” = *login\_status\_detail*



### Figure 7-1. Login Responses PDU Format

---

If both *login\_status\_class* and *login\_status\_detail* are zeroes, indicating the Storage Driver is ready to proceed, the Target Module will continue the login processing of the connection.

The *max\_cmd* passed in from the Storage Driver via *chiscsi\_target\_first\_login\_check\_done()* will be used as the maximum iSCSI command window for the session, if it is zero, the Target Module will default to maximum of 128 command per session.

### 7.3.1.2 Login Stage Transition

The iSCSI Login Phase could proceed in two stages indicated by the “CSG” field of the *Login Request PDU*:

- *ISCSI\_LOGIN\_STAGE\_SECURITY*
- *ISCSI\_LOGIN\_STAGE\_OPERATIONAL*

If the *chiscsi\_target\_class* of the requested target node has *fp\_login\_stage\_check()* set, the Target Module would also invoke *fp\_login\_stage\_check()* upon receiving the first *Login Request PDU* without C-bit (continue) set, similar to the case for *fp\_first\_login\_check()*.

If the login phase starts from *ISCSI\_LOGIN\_STAGE\_SECURITY* *fp\_login\_stage\_check()* could be called again when the target agrees to transit to the next stage *ISCSI\_LOGIN\_STAGE\_OPERATIONAL*.

Similar to *fp\_first\_login\_check()*, *fp\_login\_stage\_check()* is non-blocking. The Target Module will halt the login processing on this connection until the Storage Driver calls the exported API *chiscsi\_target\_login\_stage\_check\_done()*.

The target will reject the initiator’s login request if the *login\_status\_class* from *chiscsi\_target\_login\_stage\_check\_done()* is non-zero.

### 7.3.1.3 CHAP Processing

The Initiator and the Target may decide to authenticate each other in the security stage of the login phase.

In this stage, the target would invokes *fp\_chap\_info\_get()* when it receives CHAP\_N and CHAP\_R from the initiator.

The target Module will set the *chap\_info* structure as follows:

- the flag will have *CHAP\_FLAG\_REMOTE\_NAME\_VALID* set,

- the value of CHAP\_N received will be copied into *remote\_name*, the string will be terminated by a NULL character.

The Storage Driver should

- Set *CHAP\_FLAG\_REMOTE\_SECRET\_VALID* bit in the flag
- fill in the *remote\_secret* and *remote\_secret\_length*

If the mutual CHAP information is available, it should also

- Set *CHAP\_FLAG\_LOCAL\_NAME\_VALID* and *CHAP\_FLAG\_LOCAL\_SECRET\_VALID* bits in the flag
- fill in the *remote\_name* and *remote\_name\_length*
- fill in the *remote\_secret* and *remote\_secret\_length*
- If the Mutual CHAP is required, set *CHAP\_FLAG\_MUTUAL\_CHAP\_REQUIRED*

*fp\_chap\_info\_get()* is non-blocking. In case of error, the Target Module will reject the login and terminate the connection.

### 7.3.2 Session Addition

A session is created or associated with a connection when the connection moves into the login phase.

Once the leading connection of a newly created session completes the login phase successfully and moves into the full feature phase (FFP), the Target Module will call *fp\_session\_added()* to inform the Storage Driver of the new active session.

*fp\_session\_added()* is a blocking call.

### 7.3.3 Session Removal

In the normal session termination scenario, when the iSCSI Initiator decides to end the iSCSI session, it sends a *Logout Request PDU* to the target. After the target responds with a *Logout Response PDU*, the associated session will be closed and all the resources related to this session will be released.

A session could also be terminated abnormally. The Target may decide that it is unable to continue providing service to a specific Initiator or all Initiators.

An abnormal session termination can be caused by reasons such as

- Administrative action to take the target offline.
- Initiator access control violation.
- Local TCP state machine encountered an unrecoverable error such as a timeout, and the connection is the only connection of that session.

- 
- Received PDU has header digest error or iSCSI protocol error on a connection, and the connection is the only connection of that session.

In both cases the Target Module will call *fp\_session\_removed()* to inform the Storage Driver of the session is being terminated.

*fp\_session\_removed()* is a blocking call.

### 7.3.4 Session Abort

The Storage Driver could also force the target to terminate an established session via the exported API *chiscsi\_target\_session\_abort()*.

The handle of the session abort request is similar to abnormal session termination in the previous section.

The Storage Driver will still be called via *fp\_session\_removed()* when the session termination is completed

### 7.3.5 iSCSI SCSI Command Processing

The Target Module supports SCSI uni-directional read and write commands.

#### 7.3.5.1 iSCSI SCSI Command

There is an *chiscsi\_scsi\_command* structure for every *SCSI command PDU* received from the initiator.

The fields listed below are public and can be accessed in the Storage Driver provided functions. They will contain meaningful values during the lifetime of the scsi command. Most of them are read-only except when marked as writeable in the following sections.

The Storage Driver could use “*sc\_sdev\_hdl*” to keep any per SCSI command information/data structures for its own private use. This pointer will not be accessed in the Target Module.

```
typedef struct chiscsi_scsi_command          chiscsi_scsi_command;  
struct chiscsi_scsi_command {  
    ...  
    /* value returned by fp_session_added */
```

```

unsigned long          sc_tclass_sess_priv;

chiscsi_tcp_endpoint *sc_ieps;
void                  *sc_sessl;      /* the iscsi session handle */
void                  *sc_conn;       /* the iscsi connection handle */

unsigned int           sc_flag;

#define SC_FLAG_READ           0x1
#define SC_FLAG_WRITE          0x2
#define SC_FLAG_SENSE          0x4
#define SC_FLAG_TMF_ABORT      0x8
#define SC_FLAG_SESS_ABORT     0x10
#define SC_FLAG_CMD_ABORT      0x20
#define SC_FLAG_XFER_OVERFLOW   0x40
#define SC_FLAG_XFER_UNDERFLOW 0x80
#define SC_FLAG_XFER_BI_OVERFLOW 0x100
#define SC_FLAG_XFER_BI_UNDERFLOW 0x200
#define SC_FLAG_IMMEDIATE_CMD   0x400

unsigned char           sc_attribute; /* ATTR field of SCSI Command PDU */
unsigned char           sc_cmdlen;    /* length of the cdb, >= 16 */
unsigned char           *sc_cmd;       /* CDB from SCSI Command PDU */

unsigned int            sc_CmdSN;    /* iscsi cmdSN */
unsigned int            sc_itt;       /* iscsi initiator task tag */
unsigned int            sc_xfer_len;  /* transfer length specified in
                                         SCSI Command PDU */

unsigned int            sc_lun;       /* lun from SCSI Command PDU*/

/* for constructing SCSI_RESPONSE response, filled by the storage */
unsigned char           sc_status;
unsigned char           sc_response;
unsigned char           sc_sense_key;
unsigned char           sc_sense_asc;
unsigned char           sc_sense_ascq;

unsigned char           sc_sense_buf[SCSI_SENSE_BUFFERSIZE];
unsigned int            sc_xfer_residualcount;

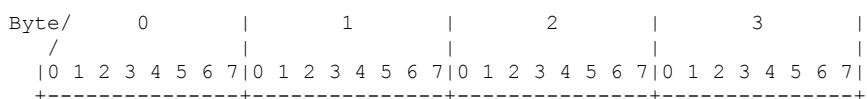
void                   *sc_sdev_hdl; /* by set the Storage Driver for its own
                                         private data if there is any */

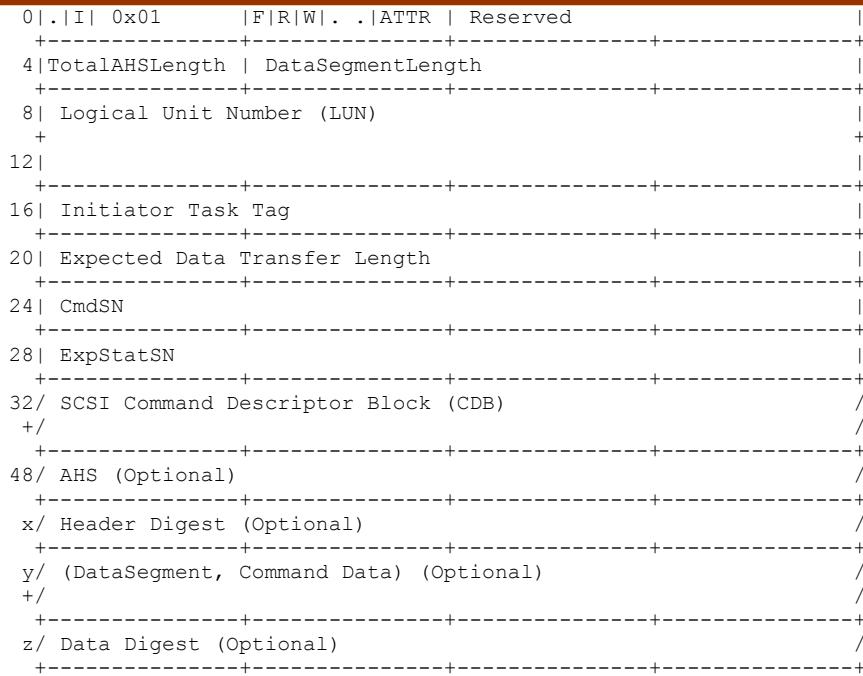
...
};


```

When the iSCSI target module receives an SCSI Command PDU, it sets up an *chiscsi\_scsi\_command* structure according to the received *SCSI Command PDU*.

The format of the *SCSI Command PDU* is as follows:





**Figure 7-1. SCSI Command PDU Format**

Bi-directional SCSI command (i.e. both R-bit and W-bit are set) will be rejected.

The *sc\_ieps* field of *chisci\_scsi\_command* contains the tcp 4-tuple.

The *chisci\_scsi\_command* will be marked as a write command (*SC\_FLAG\_WRITE*) if the *SCSI Command PDU* has its w-bit set. Otherwise, it will be marked as a read command (*SC\_FLAG\_READ*).

The *chisci\_scsi\_command* is then set up according to the *SCSI Command PDU*:

- *sc\_attribute* = PDU’s “*ATTR*” field
- *sc\_xfer\_len* = PDU’s “*expected data transfer length*” field
- *sc\_lun* = PDU’s “*Logical Unit Number (LUN)*” field
- *sc\_cmd* = PDU’s “*SCSI Command Descriptor Block (CDB)*” field
- *sc\_cmdlen* = 16
- *sc\_itt* = PDU’s “*Initiator Task Tag*” field
- *sc\_Cmdsn* = PDU’s “*CmdSN*” field

The possible values of *sc\_attribute* are:

- #define ISCSI\_TASK\_ATTRIBUTES\_UNTAGGED 0
- #define ISCSI\_TASK\_ATTRIBUTES\_SIMPLE 1
- #define ISCSI\_TASK\_ATTRIBUTES\_ORDERED 2
- #define ISCSI\_TASK\_ATTRIBUTES\_HEAD\_OF\_QUEUE 3
- #define ISCSI\_TASK\_ATTRIBUTES\_ACA 4

Once a SCSI command CDB (i.e., *SCSI Command PDU*) is received, the Target Module hands the *chiscsi\_scsi\_command* to the Storage Driver by calling *fp\_scsi\_cmd\_cdb\_rcved()*.

If the command is a read command the Storage Driver should queue the command up for execution. Once the command is executed, it notifies the Target Module via *chiscsi\_scsi\_cmd\_execution\_status()*. The Target Module will then send the data out and return the data transfer status to the Storage Driver via *fp\_scsi\_cmd\_data\_xfer\_status()*. See section 7.3.5.2 below.

If the Lun has *LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT* is set, the Storage Driver should pass the next chunk of read data to the Target Module via *chiscsi\_scsi\_cmd\_execution\_status()* once it is available. And the Target Module calls *fp\_scsi\_cmd\_data\_xfer\_status()* to indicating the chunk of read data has been transmitted.

If the command is a write command the Storage Driver should allocate the data buffers needed. After the Target Module get notified that the write data buffers are available by *chiscsi\_scsi\_cmd\_buffer\_ready()* it will solicit remaining write data from the initiator by sending out Ready-To-Transfer PDUs. Once all the write data are received, the command and data are forwarded to the Storage Driver via *fp\_scsi\_cmd\_data\_xfer\_status()*. See section below.

In the case of *LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT* is set for the LUN, the Storage Driver should pass the next chunk of write data to the Target Module via *chiscsi\_scsi\_cmd\_buffer\_ready()* once it is available. And the Target Module calls *fp\_scsi\_cmd\_data\_xfer\_status()* to indicate the chunk of write data has been received. Please note the buffers returned from the Target Module via *fp\_scsi\_cmd\_data\_xfer\_status()* could be merged from one or multiple of the buffers given by the Storage Driver.

## Data Buffer Setup

As mentioned above, the read or write data buffer are allocated by the Storage Driver and pointed by *sc\_xfer\_sreq\_buf*. The *sc\_xfer\_sgcnt* is the total number of *chiscsi\_sgvec* structures.

An *chiscsi\_sgvec* structure describes the location and size of a buffer. An array of *chiscsi\_sgvec* structures can be used to represent blocks of data buffers which are logically related but may not be contiguous in the memory.

```
typedef struct chiscsi_sgvec      chiscsi_sgvec;
{
    unsigned int   sg_flag;
#define ISCSI_SG_SBUF_DMABLE 0x1
```

---

```

void          *sg_page;      /* host page if any */
unsigned char *sg_addr;     /* start of the buffer */
unsigned long sg_dma_addr;  /* physical address */
unsigned int   sg_offset;    /* used w/ sg_page: offset into the page */
unsigned int   sg_length;   /* length of the buffer */

...
};


```

All of the fields listed above should be set up properly by the Storage Driver provided functions.

The Storage Driver should allocate *sc\_xfer\_sgcnt* of *chiscsi\_sgvec* structures and initialize each *chiscsi\_sgvec* properly. These structures together with the data buffers allocated should be persistent until the data transfer from or to the buffers are complete.

Following are the options for setting up the buffers:

- If the data buffers are host page based
  - Set *sg\_page*, *sg\_addr*, *sg\_offset*, *sg\_length*.
  - The Target Module will use Linux Kernel API *pci\_map\_page()* to obtain the physical address of the page if needed.
  - *sg\_offset* field is relevant only when *sg\_page* is set.
- If the Storage Driver manages its own memory (i.e, the memory is outside of kernel space or cannot be read or write by the Target Module), it should set *sg\_dma\_addr*.
  - Set *sg\_dma\_addr*, *sg\_length* and the dma flag (i.e., *sg\_flag* = *ISCSI\_SG\_SBUF\_DMABLE*).
  - It is recommended that *sg\_addr* is set if Target Module can read and write to the buffer address..
  - If *sg\_addr* will not be set (i.e., *sg\_addr* = NULL) then the Storage driver must turn off the unsolicited data and immediate data of the corresponding target node:
    - ImmediateData=No
    - InitialR2T=Yes
- Following are the possible combinations for setting buffer addresses:
  - *sg\_length* must be set
  - *sg\_page*=NULL, *sg\_addr* not set, no dmable flag: Not allowed,
  - *sg\_page* set, *sg\_offset* set, no dmable flag: Allowed..
  - *sg\_page*=NULL, *sg\_dma\_addr* set, dmable flag set: Allowed
  - *sg\_page* set, *sg\_dma\_addr* set, dmable flag set: Allowed,
   
*sg\_dma\_addr* takes precedence over *sg\_page*.

Following is an example for write command requesting 16K data, with dma-able memory. In this case *sc\_xfer\_sgcnt*=4, *sc\_xfer\_offset*=0 and *sc\_xfer\_buflen* 16K. One of the sglist entry would look like this.

- *sg\_dma\_addr=<physical\_addr>*
- *sg\_length=4096*
- *sg\_addr=<virtual\_addr>*
- *sg\_page=NULL*
- *sg\_offset=0*
- *sg\_flag |= ISCSI\_SBUG\_SG\_DMABLE*

## Multiphase data buffers

If *LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT* is set, the Storage Driver should set *sc\_xfer\_offset* and *sc\_xfer\_buflen*, in *chiscsi\_scsi\_cmd\_buffer\_ready()* and *chiscsi\_scsi\_cmd\_execution\_status()* properly:

- *sc\_xfer\_offset* is the offset into the total data length (i.e., *sc\_xfer\_len*) indicated by the command.
- *sc\_xfer\_buflen* is the total length of the data buffers pointed by *sc\_xfer\_sreq\_buf*.

For example, with *LUN\_CLASS\_MULTI\_PHASE\_DATA\_BIT* set, for a SCSI read command requesting 256KB of data (*sc\_xfer\_len*=256K), the Storage Driver returns the read data in two phases: 64KB then followed by 192KB. For the first phase of 64KB of data, *sc\_xfer\_offset* = 0 and *sc\_xfer\_buflen* = 64KB. And *sc\_xfer\_offset* = 64KB and *sc\_xfer\_buflen* = 192KB for the second phase of 192KB of data.

## Requirement for DDP

For write command, the Terminator can perform Direct Data Placement (DDP) for direct placement of iSCSI PDU payload data into pre-posted final-destination host-memory buffers.

In order for this mechanism to work, the following guideline must be observed.

For total *n* number of data buffers spanning over multiple phases

- The memory must be DMA-able and/or page based
- All data buffers from 2 ~ (n – 1) should be of the same length and length should be either 4K, 8K, 16K, or 64K (i.e. *sg\_offset* = 0 and *sg\_length* are of the same value).
- The first (1) buffer can be partially used (i.e., *sg\_offset* >= 0 and *sg\_length* <= the length of buffers from 2 ~ (n – 1)).
- The last (*sc\_xfer\_sgcnt*) buffer can be partially used from the beginning (i.e., *sg\_offset* = 0 and *sg\_length* <= the length of buffers from 2 ~ (n – 1)).

It is preferable to not setup DDP for small IOs (< 2K), hence *sg\_addr* should be provided in these cases.

---

Following is an example of multiphase, DDP buffer setup for a 64KB write command with 8K data buffer length(*sg\_length*), with partially used first and last page.

Phase 1: *chiscsi\_scsi\_cmd\_buffer\_ready()*, offset=0 buflen 20K, sgcnt 3: Three entries in sglist, with *sg\_offset*=4K,0,0 and *sg\_length*=4K,8K,8K.

Phase 2: *chiscsi\_scsi\_cmd\_buffer\_ready()*, offset 20K buflen 32K, sgcnt 4: Four entries in sglist with *sg\_offset*=0,0,0,0 and *sg\_length*=8K,8K,8K,8K

Phase 3: *chiscsi\_scsi\_cmd\_buffer\_ready()*, offset=52K buflen 12K sgcnt 2: Two entries in sglist with *sg\_offset*=0,0 and *sg\_length*=8K, 4K

Storage Driver may pass phase 2 buffers to iSCSI target stack before phase 1 buffer has been returned back by the Target Module.

Both *fp\_scsi\_cmd\_cdb\_rcved()* and *fp\_scsi\_cmd\_data\_xfer\_status()* are non-blocking calls, the target will halt the processing of the command until the Storage Driver notifies the Target Module.

If the Storage Driver could not execute the SCSI command or the execution results in error, it should set the following fields properly in the *chiscsi\_scsi\_command*:

- *sc\_status*
- *sc\_response*
- *sc\_sense\_key*
- *sc\_sense\_asc*
- *sc\_sense\_ascq*
- *sc\_sense\_buf*
- *sc\_xfer\_residualcount*

The Target Module will return the execution status to the Initiator based on the fields mentioned above.

## Buffer Overflow and Underflow

The iSCSI Target module provides two flags to handle buffer overflow and buffer underflow, namely `SC_FLAG_XFER_OVERFLOW` and `SC_FLAG_XFER_UNDERFLOW`. The phase in which the flags are set will be assumed to be the last phase which means that Storage Driver should not send any additional buffers after this.

For buffer overflow, the Storage Driver needs to:

1. Set the `SC_FLAG_XFER_OVERFLOW` flag and
2. Set the residual count in `sc->sc_xfer_residualcount`

The Residual Count indicates the number of bytes that were not transferred because the initiator's Expected Data Transfer Length was not sufficient. iSCSI Target Module will transfer only Expected Data Transfer Length of bytes but send *fp\_scsi\_cmd\_data\_xfer\_status()* for all of the buffers that were provided.

For buffer underflow, The Storage Driver needs to:

1. Set the flag `SC_FLAG_XFER_UNDERFLOW`
2. Set residual count in `sc->sc_xfer_residualcount`
3. Set new `sc_xfer_len = bufferlen`

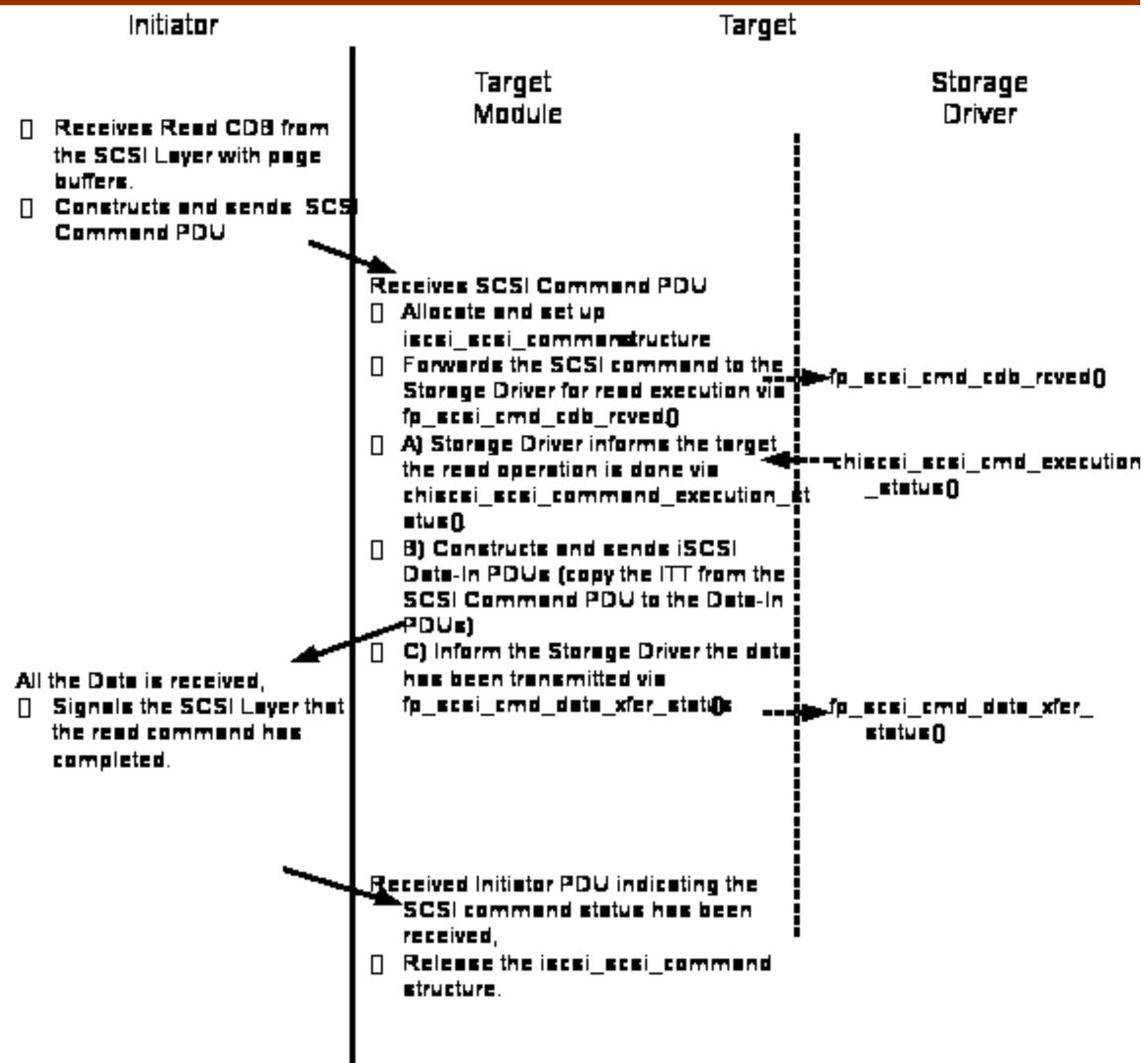
The Residual Count indicates the number of bytes that were not transferred out of the number of bytes that were expected to be transferred.

Additionally the Storage driver should set the appropriate `sc->sc_response` and `sc->sc_status` in `chisci_scsi_command` structure.

### **7.3.5.2 iSCSI SCSI Command Read Operation**

For SCSI read request, the Target Module will transfer the data from the memory allocated by the Storage Driver to the Initiator.

The diagram below describes the sequence of actions of a SCSI read operation:

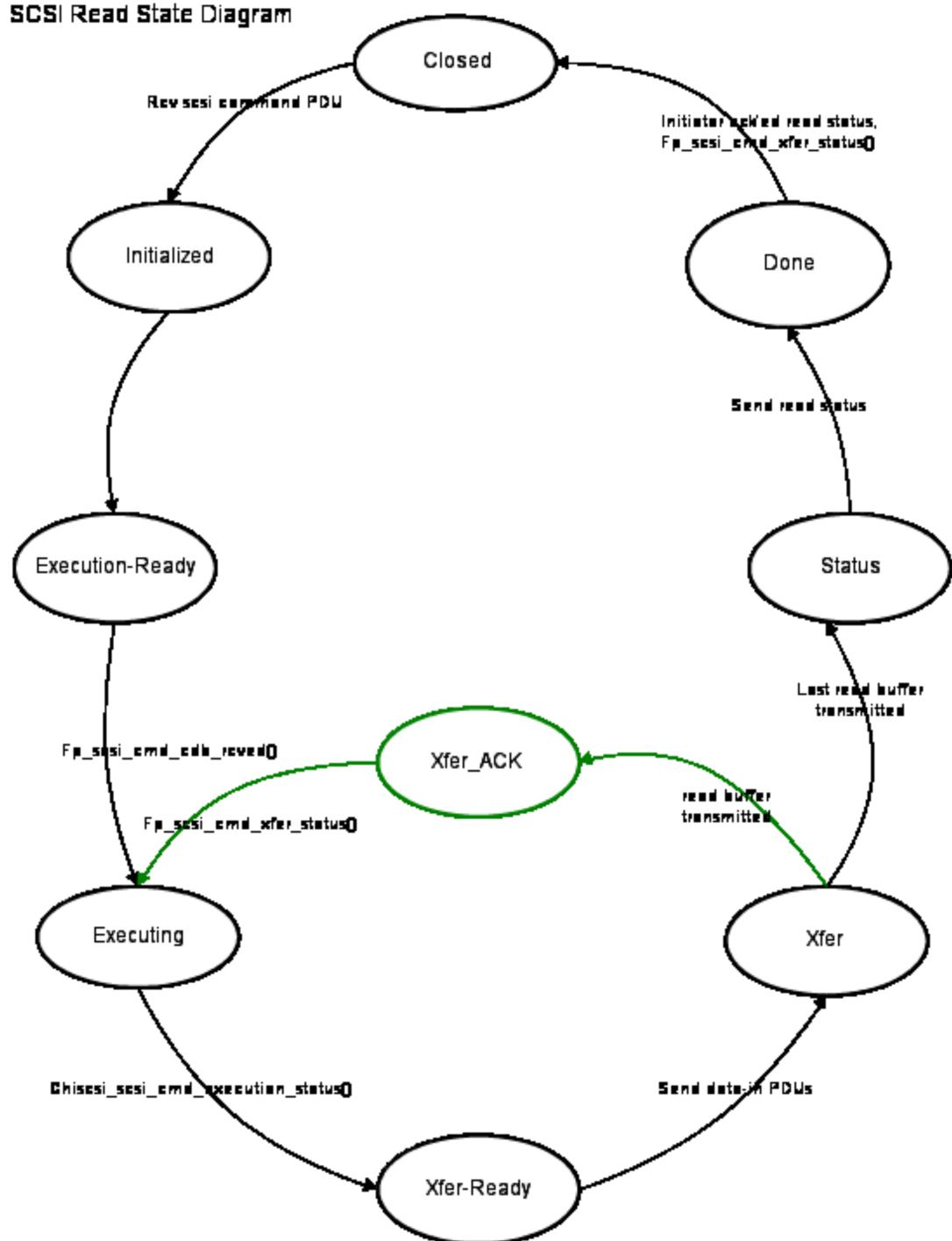


**NOTE:**

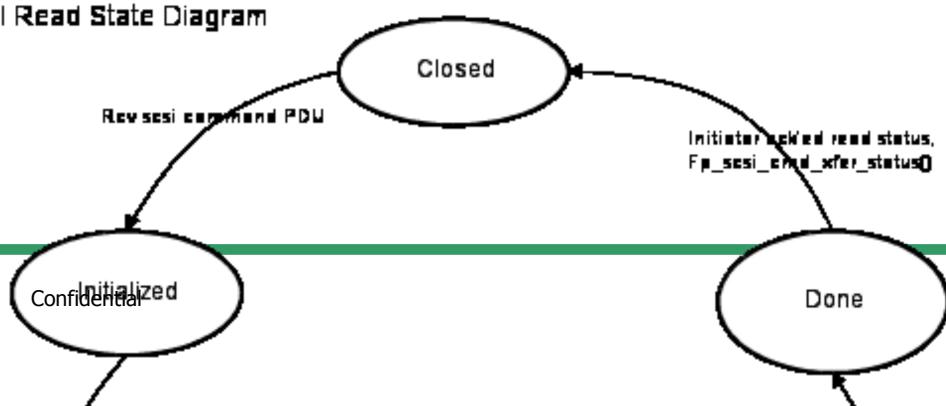
Step A), B), and C) could be repeated multiple times in the case of the Lun has `LUN_CLASS_MULTI_PHASE_DATA_BIT` set.

Figure 7-1. iSCSI SCSI Read Walkthrough

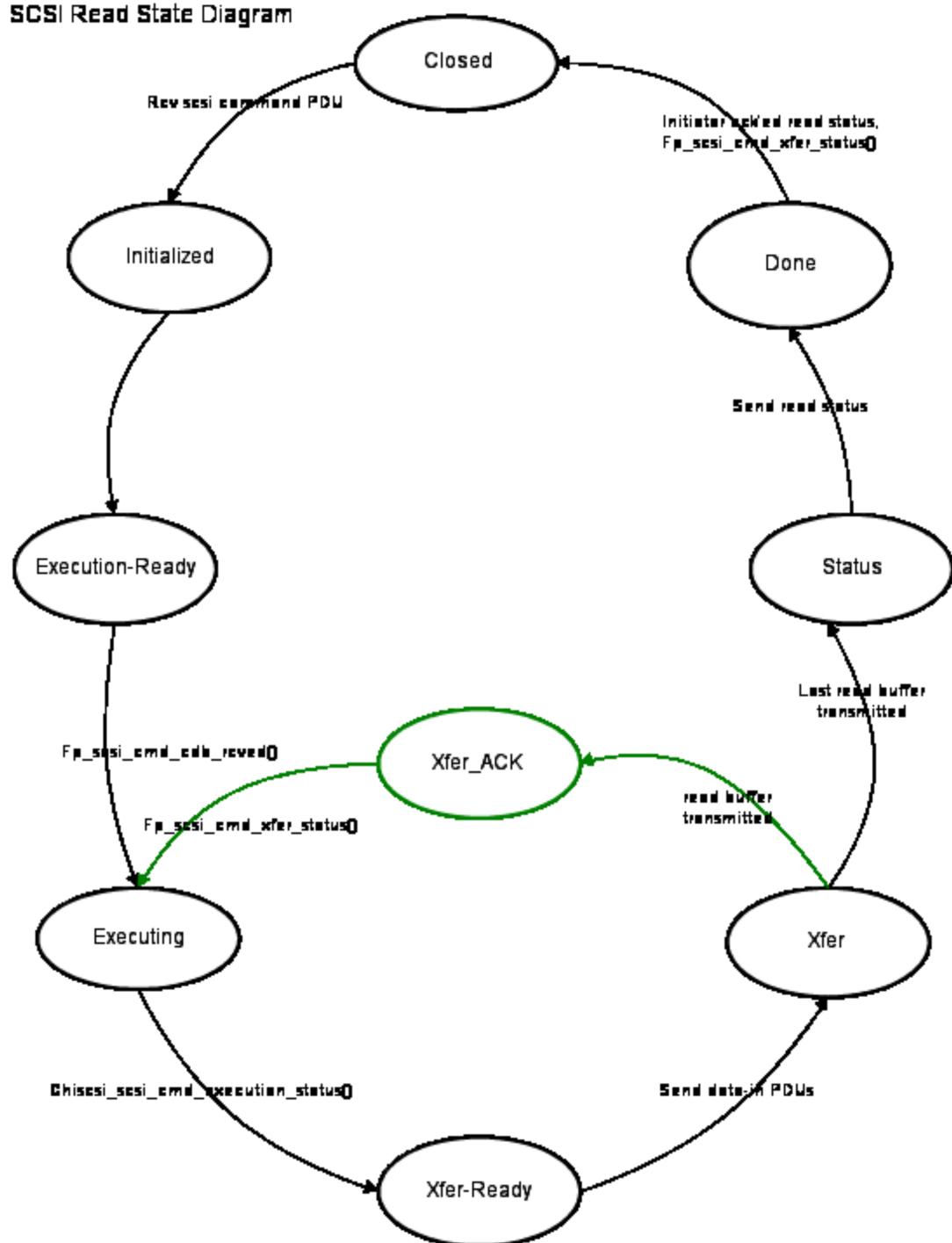
**SCSI Read State Diagram**



**SCSI Read State Diagram**



**SCSI Read State Diagram**



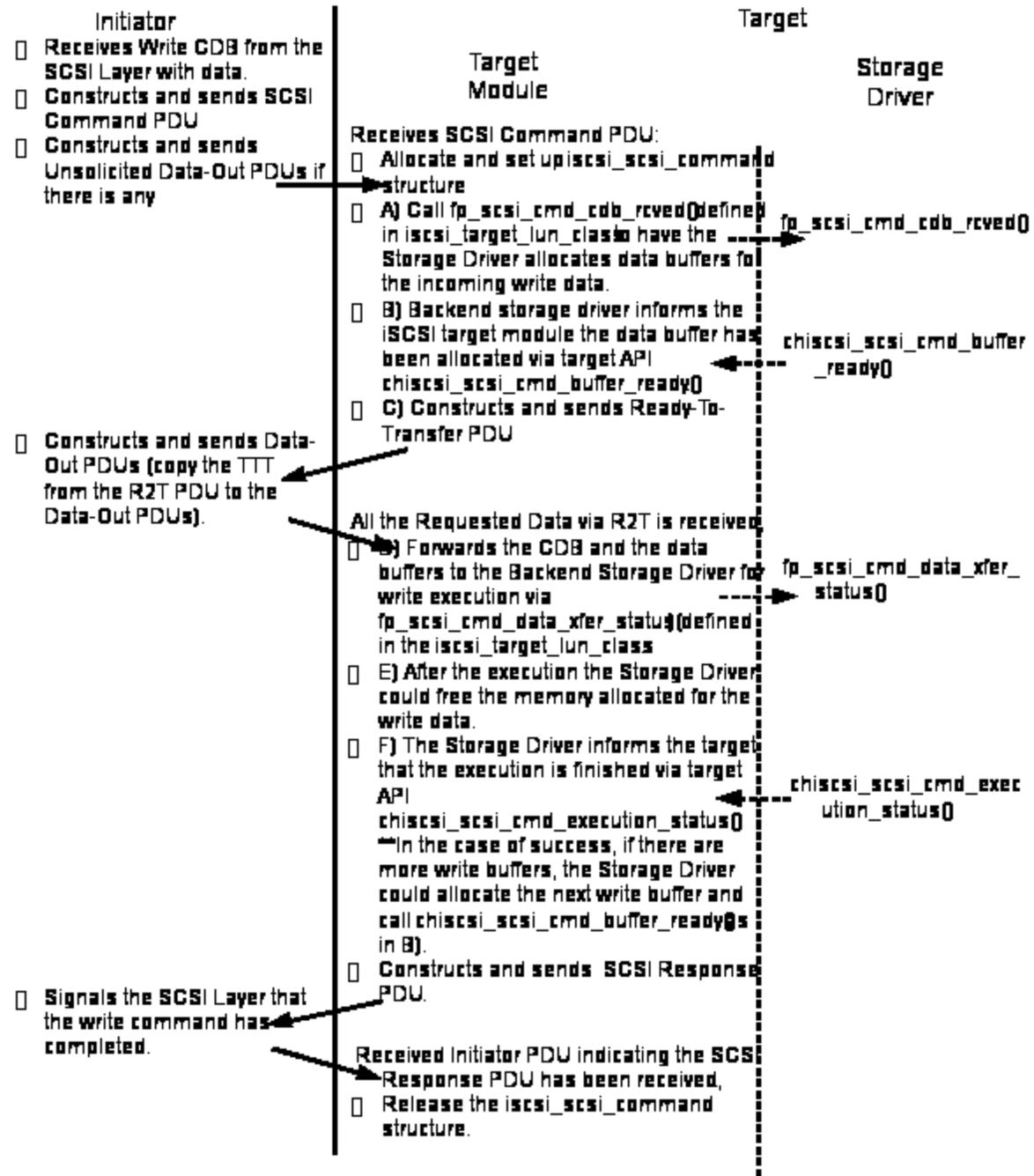
**Figure 7-1. iSCSI SCSI Read State Diagram**

### 7.3.5.3 iSCSI SCSI Command Write Operation

For SCSI write request, the Target Module will first call *fp\_scsi\_cmd\_cdb\_rcved()* to have the Storage Driver allocate the write data buffers. It will copy or DMA the data received from the initiator to the memory allocated by the Storage Driver. The SCSI command and data then is handed back to the Storage Driver for write execution via *fp\_scsi\_cmd\_data\_xfer\_status()*.

The diagram below describes the action sequence of a SCSI write operation:

## SCSI Write Walk-Through

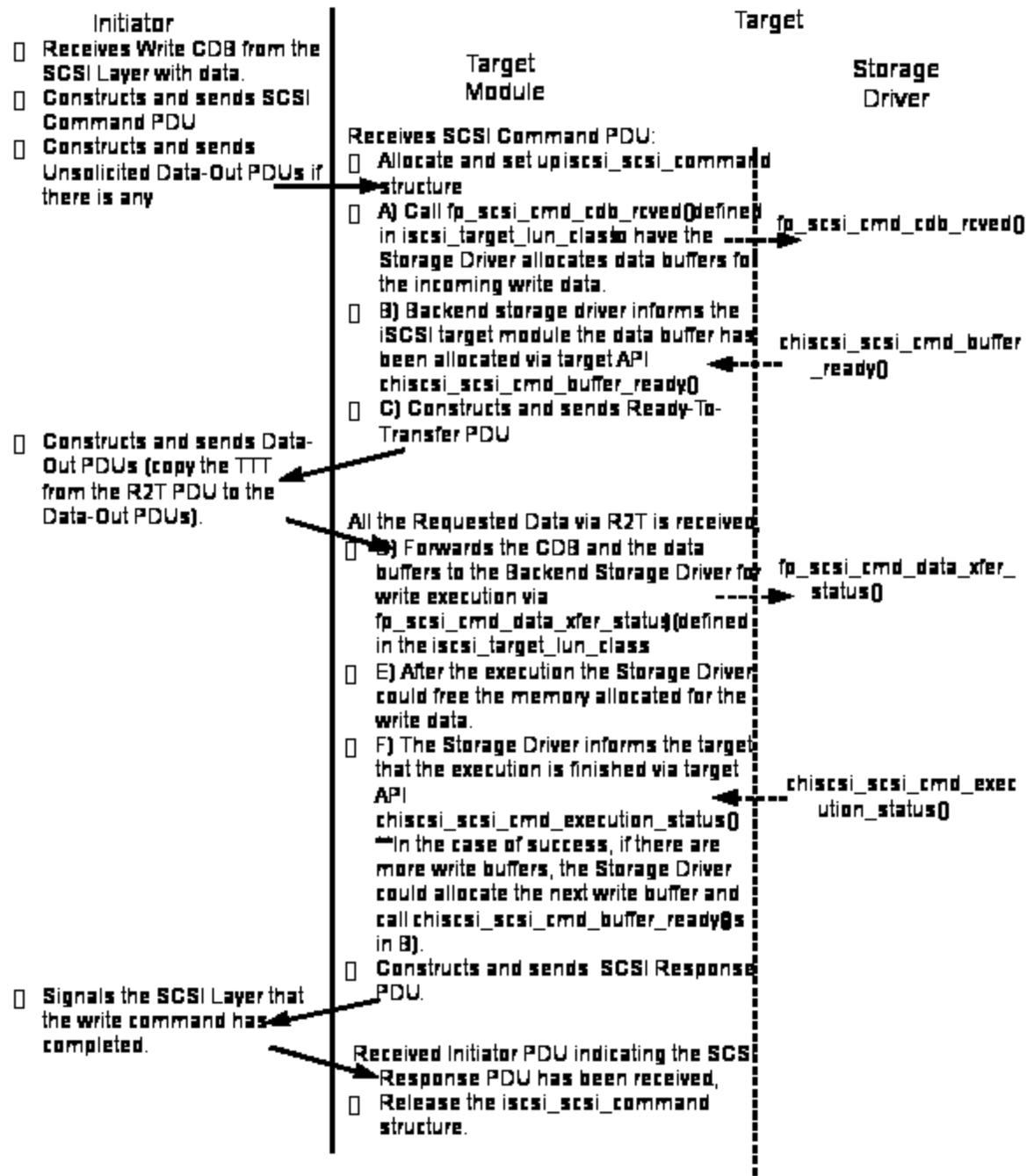


**NOTE:** In the case of the Lun has `LUN_CLASS_MULTI_PHASE_DATA_BIT` set, Step A), B), C), D), E) and F) could be repeated multiple times (the long path), or Step B), C), D), and E) could be repeated multiple times (the short path).

## SCSI Write Walk-Through



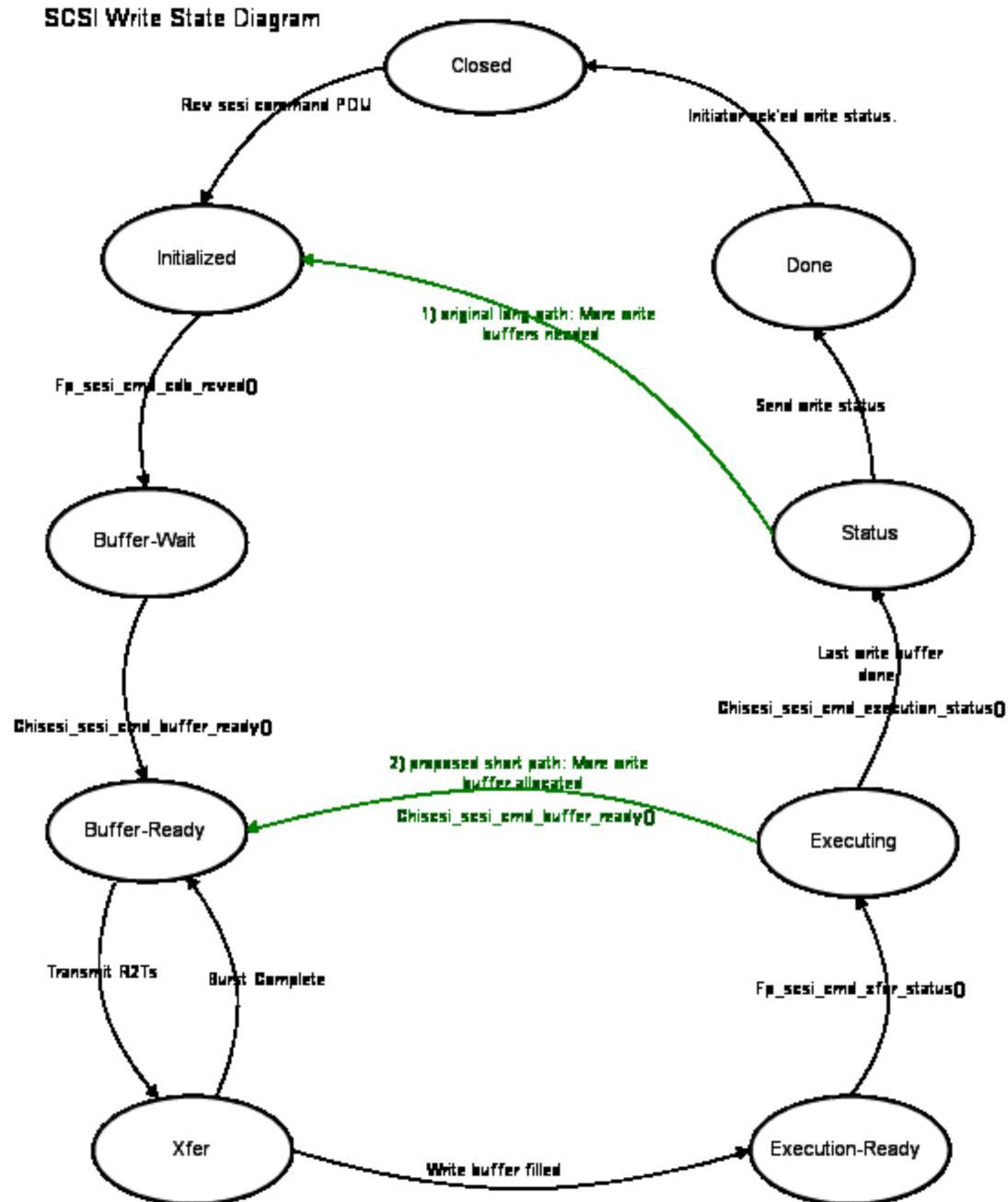
### SCSI Write Walk-Through



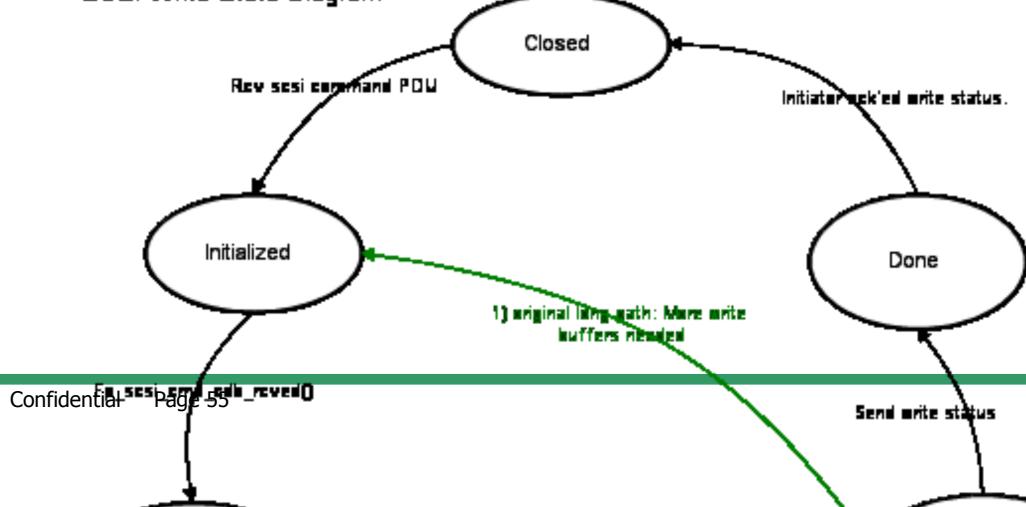
**NOTE:** In the case of the Lun has `LUN_CLASS_MULTI_PHASE_DATA_BIT` set, Step A), B), C), D), E) and F) could be repeated multiple times (the long path), or Step B), C), D), and E) could be repeated multiple times (the short path).

Figure 7-1. iSCSI SCSI Write Walkthrough

### SCSI Write State Diagram



### SCSI Write State Diagram



SCSI Write State Diagram

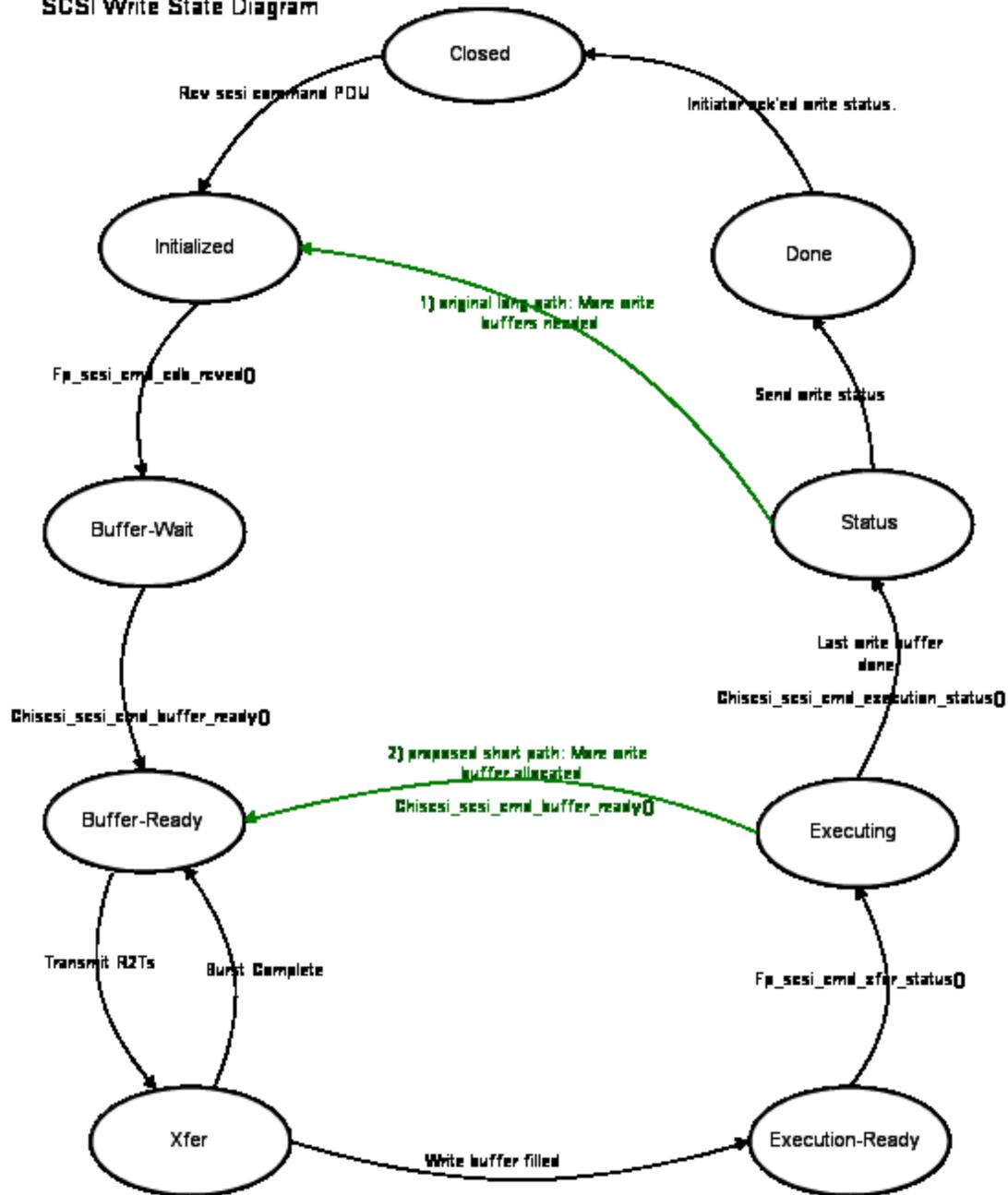


Figure 7-1. iSCSI SCSI Write State Diagram

#### 7.3.5.4 Abort a SCSI Command

When a target session has to be terminated due to either Initiator Requests or iSCSI protocol errors, the Target Module will issue `fp_scsi_cmd_abort()` for every outstanding SCSI command that has been forwarded to the Storage Driver (either via `fp_scsi_cmd_cdb_rcved()` or `fp_scsi_cmd_data_xfer_status()`).

Each of the *chiscsi\_scsi\_command* will have *SC\_FLAG\_SESS\_ABORT* set.

After cleaning up the command resources the Storage Driver should notify the Target Module by invoking *chiscsi\_scsi\_cmd\_abort\_status()*. The Storage Driver should set up the *chiscsi\_scsi\_command* similar to the case where the execution results in error.

*fp\_scsi\_cmd\_abort()* is non-blocking.

If the Storage Driver wants to abort a SCSI command that it has already executed or allocated memory for, it should call the Target Module's exported API function *chiscsi\_scsi\_cmd\_abort()*. The storage driver should set up the *chiscsi\_scsi\_command* similar to the case where the execution results in error.

*chiscsi\_scsi\_cmd\_abort()* is non-blocking.

After the Target Module aborts the command it will inform the Storage Driver via *fp\_scsi\_cmd\_abort\_status()*.

### 7.3.5.5 Debugging SCSI Command problem

The Target Module provides a function to dump out an *chiscsi\_scsi\_command* information:

```
void chiscsi_iscsi_command_dump (chiscsi_scsi_command *scmd)
```

*chiscsi\_iscsi\_command\_dump()* will print out the *chiscsi\_scsi\_command* structure information into the syslog.

### 7.3.6 iSCSI Task Management Functions

When the Target Module receives a *Task Management Function Request PDU*, it may invoke *fp\_tmf\_execute()* defined in *iscsi\_target\_lun\_class*.

The possible Task Management Functions (TMF) include

- #define ISCSI\_TMF\_FUNCTION\_ABORT\_TASK 1
- #define ISCSI\_TMF\_FUNCTION\_ABORT\_TASK\_SET 2
- #define ISCSI\_TMF\_FUNCTION\_CLEAR ACA 3
- #define ISCSI\_TMF\_FUNCTION\_CLEAR\_TASK\_SET 4
- #define ISCSI\_TMF\_FUNCTION\_LOGICAL\_UNIT\_RESET 5
- #define ISCSI\_TMF\_FUNCTION\_TARGET\_WARM\_RESET 6
- #define ISCSI\_TMF\_FUNCTION\_TARGET\_COLD\_RESET 7
- #define ISCSI\_TMF\_FUNCTION\_TASK\_REASSIGN 8

The target will handle the TMF locally under the following conditions (i.e., the Storage Driver will not be involved):

- Function is *ISCSI\_TMF\_FUNCTION\_TASK REASSIGN*, or
- Function is *ISCSI\_TMF\_FUNCTION\_CLEAR ACA*, or
- Function is *ISCSI\_TMF\_FUNCTION\_ABORT\_TASK* and
  - a. the target could not find the requested SCSI command, or
  - b. the SCSI command has already completed execution (i.e. *chisci\_scsi\_cmd\_execution\_status()* has already been called for this command), or
  - c. the SCSI command has not been sent to the Storage Driver for execution.

In the rest of the scenarios the target will forward the TMF request to the Storage Driver via *fp\_tmf\_execute()*.

```
int (*fp_tmf_execute) (unsigned long sess_tclass,
                      unsigned long hndl,
                      unsigned char immediate_cmd,
                      unsigned char tmf_func,
                      unsigned int lun,
                      chisci_scsi_command *scmd);
```

where

- *sess\_tclass* is the private session hndl returned by the Storage Driver from *fp\_session\_added()*.
- *lun* is valid when the function is *ISCSI\_TMF\_FUNCTION\_ABORT\_TASK*, or *ISCSI\_TMF\_FUNCTION\_CLEAR\_TASK\_SET*, or *ISCSI\_TMF\_FUNCTION\_ABORT\_TASK\_SET*, or *ISCSI\_TMF\_FUNCTION\_LOGIN\_UNIT\_RESET*.
- *Scmd* will point to a valid *chisci\_scsi\_command* structure when the function is *ISCSI\_TMF\_FUNCTION\_ABORT\_TASK*.

In the case where the scsi command is in the data transfer stage (i.e., *chisci\_scsi\_cmd\_buffer\_ready()* has been called for the command by the Storage Driver), the Target Module will

- set *chisci\_scsi\_command.sc\_flag = SC\_FLAG\_TMF\_ABORT*, and
- invoke *fp\_scsi\_cmd\_abort()*.

For each scsi command being aborted, after cleaning up the command resources the Storage Driver should notify the Target Module by invoke *chisci\_scsi\_cmd\_execution\_status()*. The Storage Driver should setup the *chisci\_scsi\_command* similar to the cases where the execution results in error.

*fp\_tmf\_execute()* is non-blocking, the Target Module will not respond to the initiator with *Task Management Function Response PDU* until the Storage Driver notify the Target Module of the TMF response code via *chisci\_tmf\_execution\_done()*.

The possible Task Management Function (TMF) response code include

- #define ISCSI\_RESPONSE\_TMF\_COMPLETE 0
- #define ISCSI\_RESPONSE\_TMF\_INVALID\_TASK 1
- #define ISCSI\_RESPONSE\_TMF\_INVALID\_LUN 2
- #define ISCSI\_RESPONSE\_TMF\_TASK\_STILL\_ALLEGANT 3
- #define ISCSI\_RESPONSE\_TMF\_NO\_TASK\_FAILOVER 4
- #define ISCSI\_RESPONSE\_TMF\_NOT\_SUPPORTED 5
- #define ISCSI\_RESPONSE\_TMF\_AUTH\_FAILED 6
- #define ISCSI\_RESPONSE\_TMF\_FUNCTION\_REJECTED 0xff

## 8 Target Information and Statistics

To allow Storage Driver to retrieve target settings and statistical data, the following structures are defined:

```
struct chisci_perf_info {  
    unsigned long read_bytes;  
    unsigned long write_bytes;  
    unsigned long read_cmd_cnt;  
    unsigned long write_cmd_cnt;  
};  
  
struct iscsi_session_settings {  
    unsigned char initiatial_r2t:1;  
    unsigned char immediate_data:1;  
    unsigned char erl:2;  
    unsigned char data_pdu_in_order:1;  
    unsigned char data_sequence_in_order:1;  
    unsigned char filler:2;  
  
    unsigned int max_conns;  
    unsigned int max_r2t;  
    unsigned int first_burst;  
    unsigned int max_burst;  
    unsigned int time2wait;  
    unsigned int time2retain;  
};  
  
struct iscsi_conn_settings {  
    unsigned char header_digest[2];  
    unsigned char data_digest[2];  
  
    unsigned int portal_group_tag;  
    unsigned int max_recv_data_segment;  
    unsigned int max_xmit_data_segment;  
};  
  
struct iscsi_chap_settings {  
    unsigned char chap_en:1;  
    unsigned char chap_required:1;  
    unsigned char mutual_chap_forced:1;  
    unsigned char filler:5;  
  
    unsigned int challenge_length;  
};  
  
struct iscsi_target_config_settings {  
    unsigned char acl_en:1;  
    unsigned char isns_register:1;  
    unsigned char shadow_mode:1;  
    unsigned char filler:5;  
  
    unsigned int sess_max_cmds;
```

```
};

struct chisci_target_info {
    char name[256];
    char alias[256];

    struct iscsi_session_settings sess_keys;
    struct iscsi_conn_settings conn_keys;
    struct iscsi_chap_settings chap;
    struct iscsi_target_config_settings config_keys;

    unsigned char auth_order;

    /* private to Chelsio Stack */
    unsigned long hndl;
};

struct chisci_session_info {
    char peer_name[256];
    char peer_alias[256];

    struct iscsi_session_settings sess_keys;

    unsigned char type;
    unsigned char isid[6];
    unsigned char conn_cnt;
    unsigned short tsih;

    unsigned int cmdsn;
    unsigned int maxcmdsn;
    unsigned int expcmdsn;

    struct chisci_perf_info perf;
    /* private to Chelsio Stack */
    unsigned long hndl;
};

struct chisci_connection_info {
    struct chisci_tcp_endpoint tcp_endpoints;

    struct iscsi_conn_settings conn_keys;

    unsigned char offloaded:1;
    unsigned char filler:7;

    unsigned int cid;
    unsigned int statsn;
    unsigned int expstatsn;

    /* private to Chelsio Stack */
    unsigned long hndl;
};
```

## 8.1 Target Information

The Storage driver could get detail information about the target using the exported API *chiscsi\_target\_get\_information()*.

```
int chiscsi_get_target_info (char *tname,
                           struct chiscsi_target_info *target_info);
```

Where,

- *tname* – is the target\_name provided by the storage driver.
- *target\_info* – is the output structure which will return the details information on the target specified by target\_info structure.

## 8.2 Session Information

The Storage driver could get information about established session parameters using the exported API *chiscsi\_target\_get\_session()*. The Storage Driver can issue this API call anytime it wants to collect the detail information about the established sessions.

```
int chiscsi_get_session_info (char *tname,
                            char *iname,
                            int sess_info_max,
                            struct chiscsi_session_info *sess_info);
```

Where,

- *tname* – is the target\_name provided by the storage driver.
- *iname* – is the initiator name provided by the storage driver.
- *sess\_info\_max* – the number of *chiscsi\_session\_info* structures pointed by *sess\_info*.
- *sess\_info* – is the pointer to *chiscsi\_session\_info* structure. This structure will be returned as output once the data has been filled in by chiscsi target stack.

When there are less than *sess\_info\_max* active sessions between a given initiator and a target, *chiscsi\_target\_get\_session()* will return the number of sessions filled in. Otherwise, it will return the total numbers of active sessions. The *sess\_info* would contains the information of the first *sess\_info\_max* sessions.

## 8.3 Connection Information

The Storage Driver can retrieve connection information by calling exported API *chiscsi\_get\_connection\_info()*.

```
int chiscsi_get_connection_info (unsigned long sess_hdl,
                               int conn_idx,
                               struct chiscsi_connection_info *conninfo)
```

---

The connection information API should always be called after session API so that sess\_hdl and conn\_cnt parameter can be used to retrieve connection details.

## 8.4 Portal Performance Data

The Storage Driver can retrieve portal performance data by calling exported API *chiscsi\_get\_perf\_info()*.

```
int chiscsi_get_perf_info          (int ipv6,  
                                   unsigned int *ip_addr,  
                                   unsigned int port,  
                                   struct chiscsi_perf_info *pdata)
```

Where,

- *ipv6* – whether *\*ip\_addr* is ipv6 address or not.
- *address* – ip address for the target portal for which the performance statistics are to be collected.
- *port* – port number of the target portal. If port number is not provided Chelsio target stack will use default (3260) port number.
- *pdata* – pointer to the structure *chiscsi\_perf\_data* which will be filled in by Chelsio iSCSI stack

A new structure to store performance statistics is created for storage driver to pass pointer to when calling the exported API.

```
struct chiscsi_perf_info {  
    unsigned long          read_bytes;  
    unsigned long          write_bytes;  
    unsigned long          read_cmd_cnt;  
    unsigned long          write_cmd_cnt;  
};
```

Where,

- *read\_bytes* – number of bytes read by the target portal.
- *write\_bytes* – number of bytes written by the target portal.
- *read\_cmd\_cnt* – number of read command executed by the target portal.
- *write\_cmd\_cnt* – number of write commands executed by the target portal.

## 9 Chelsio iSCSI Target API Files

All the data structure definitions and iSCSI Target API function declarations described in the previous chapters are under includes/common/.

Name	Description
iscsi_target_class.h	<i>iscsi_target_class</i> and <i>iscsi_target_lun_class</i> data structures and related APIs.
chiscsi_sgvec.h	scatter gather vector representation <i>chiscsi_sgvec</i> .
chiscsi_scsi_command.h	Describes an iSCSI SCSI Command request and associated data <i>chiscsi_scsi_command</i> .
iscsi_chap.h	<i>chap_info</i> representation
iscsi_info.h	Target, session, connection settings and performance counters

**Table 9-1. Target API Header Files**

## Appendix A. Chelsio iSCSI Software Features

This table below summarizes the major capabilities supported by Chelsio iSCSI Software:

Feature	Y/N	Comment
Multi-Path I/O	Y	
Discovery	Y	
“iqn.” Naming	Y	
“eui.” Naming	Y	
Login	Y	
Text Key List Negotiations	Y	
Text Key Simple Value Negotiations	Y	
Security Negotiation	Y	
Authentication Method	Y	
Connection Reinstatement	N	
Session Reinstatement	Y	
Session Continuation and Failure	N	
Logout	Y	
CHAP	Y	One way and Mutual
Access Control List	Y	IP address and/or Initiator Name based
Target LUN Masking	Y	
Error Recovery level	Y	Level 0 only
Reject PDU usage	Y	
Connection Timeout	Y	
Immediate Data	Y	
Unsolicited SCSI Data Out	Y	
Outstanding R2T	Y	Support multiple outstanding R2Ts
Task Mgmt Function	Y	
Asynchronous Message	Y	
Text Request/Response	Y	
Login Request/Response	Y	
Logout Request/Response	Y	
SNACK request	N	
NOP-Out / NOP-In	Y	
iSNS Client	Y	

**Table I. iSCSI Software Features**

## Appendix B. iSCSI Text Key Settings

iSCSI Text Entity Parameters pass iSCSI protocol control information to the Chelsio iSCSI module. This information is unique for each target node.

The parameters follow the IETF iSCSI standard RFC 3720 in both definition and syntax. The descriptions below are mostly from this RFC.

Please refer to the Chelsio iSCSI User Guide for detailed description of supported values.

## Appendix C. Chelsio Target Configuration Keys

Chelsio Target Configuration Entity Parameters pass control information to the Chelsio iSCSI module. The parameters are specific to Chelsio's implementation of the iSCSI Target and are unique to each target.

The parameters consist of information that can be put into the following categories:

1. Challenge Handshake Authentication Protocol (CHAP).
2. Target specific settings.

All of the following parameters can have multiple instances in one target entity block (i.e., they can be declared multiple times for one particular target).

- Portal Group
- Storage Device
- Access Control List (ACL)

Please refer to the User Guide for detailed description of supported values.

## Appendix D. Log and Debug messages.

Chelsio iSCSI Target driver will perform sanity check on the buffer settings and will print following error messages to dmesg.

### Buffer Setup Errors

Error Message	Reason
<code>"chiscsi_scsi_cmd_buffer_ready: itt 0x5a, data buf NULL, 0+4096/8192"</code>	Buffer pointer passed in chiscsi_scsi_cmd_buffer_ready is NULL
<code>"chiscsi_scsi_cmd_buffer_ready: itt 0x5a, sg 0, exp. sgl, 0+4096/8192"</code>	sgcnt passed in chiscsi_scsi_cmd_buffer_ready is zero
<code>"chiscsi_scsi_cmd_execution_status: itt 0x5a, buf overflow, 4096! =0+8192."</code>	Received more buffer (0+8192) than ExpectedDataTransferLen (4096) while SC_FLAG_XFER_OVERFLOW flag was not set
<code>"chiscsi_scsi_cmd_execution_status: itt 0x5a, buf underflow, 16384!=0+8192"</code>	Received more buffer(0+8192) than ExpectedDataTransferLen (16384) while SC_FLAG_XFER_UNDERFLOW flag ‘and’ MULTIPHASE bit not set
<code>"chiscsi_scsi_cmd_buffer_ready: itt 0x5a, buf off 4096+4096, exp 0+4096"</code>	Buffers are out of order. offset + buflen (4096+4096) sent by storage driver is different than what Chelsio iSCSI stack expects (0+4096)

### Other Error Messages

Error Message	Reason
<code>chiscsi: sc 0xffff880231c88000, itt 0x69000000, held by storage driver</code>	iSCSI stack is waiting for callback and cannot proceed until reply comes from controller. TMF is issued from initiator in this case
<code>chiscsi:No DDP: sgcnt 2 sgl 0xffff88022d79b0e0, sg_flag 0x0, sg_addr NULL, Cannot copy data!</code>	If sg_addr is not set then command will fail.
<code>chiscsi:2/3, len 3072 xferoff 8192 NOT full pages</code>	One of the middle phase has an incomplete page, which means buffer not suitable for DDP. In this example, in one of the middle phases (indicated with the xferoff of 8192) 2 <sup>nd</sup> sglist entry out of 3, has sg_length less than page size

## Sending debug messages to Chelsio

To trace the flow of the complete chiscsi\_scsi\_command in Chelsio iSCSI stack, turn on the debug messages by executing following command:

```
iscsicl -G iscsi_verbose_level=0xf,0xc82000
```

Execute a read/write command from initiator after this and send the log to Chelsio for debugging purpose.

## Appendix E. iSCSI Acronyms

ACL	Access Control List
CHAP	Challenge-Handshake Authentication Protocol
CPL	Chelsio Protocol Language, the Terminator's programming interface
CRC	Cyclic Redundancy Check
DDP	Direct Data Placement
ERL	Error Recovery Level (0, 1, 2)
EUI	Extended Unique Identifier
FFP	Full Feature Phase
HBA	Host Bus Adaptor
H/W	Hardware
iSCSI	Internet Small Computer System Interface
iSNS	Internet Storage Name Service
ISID	Initiator Session ID
LU	Logical Unit
LVM	Logical Volume Management
MPIO	Multi-Path I/O
MCS	Multiple Connections per Session
NIC	Network Interface Card
N/A	Not applicable
NOP	No Operation
PDU	Protocol Data Unit
R2T	Request To Transmit
SLP	Service Location Protocol
SNACK	Selected Negative Acknowledgement
SNMP	Simple Network Management Protocol
SSID	Session ID
S/W	Software
TBD	To Be Determined
TCP	Transmission Control Protocol
TCB	TCP Control Block
Terminator	Chelsio's ASIC for protocol offload
TOE	TCP/IP Offload Engine
ULP	Upper Layer Protocol