

INTERNET DRAFT
draft-ietf-ips-iwarp-da-00.txt

Mallikarjun Chadalapaka
HP
John Hufferd
IBM
Julian Satran
IBM
Hemal Shah
Intel

Expires March 2005

Datamover Architecture for iSCSI (DA)

1 Status of this Memo

By submitting this Internet-Draft, we certify that any applicable patent or other IPR claims of which we are aware have been disclosed, or will be disclosed, and any of which we become aware will be disclosed, in accordance with RFC 3668.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than a "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

2 Abstract

iSCSI is a SCSI transport protocol that maps the SCSI family of application protocols onto TCP/IP. The Datamover Architecture for iSCSI (DA) defines an abstract model in which the movement of data between iSCSI end nodes is logically separated from the rest of the iSCSI protocol in order to allow iSCSI to adapt to innovations available in new IP transports. The new Datamover protocol provides a reliable transport for all iSCSI PDUs, but actually moves the data required for certain iSCSI PDUs without involving the remote iSCSI layer itself. This document begins with an introduction of a few new abstractions, defines a layered architecture for iSCSI and Datamover protocols, and then models the interactions within an iSCSI end node between the iSCSI layer and the Datamover layer that happen in order to transparently perform remote data movement within an IP fabric. It is intended that this definition would help map iSCSI to generic RDMA-capable IP fabrics in the future comprising TCP, SCTP, and possibly other underlying network transport layers.

Table of Contents

1	Status of this Memo	1
2	Abstract	2
3	Definitions and acronyms	6
3.1	Definitions	6
3.2	Acronyms	6
4	Motivation	8
5	Architectural layering of iSCSI and Datamover layers ..	10
6	Design Overview	12
7	Architectural Concepts	14
7.1	iSCSI PDU types	14
7.1.1	iSCSI data-type PDUs.....	14
7.1.2	iSCSI control-type PDUs.....	15
7.2	Data_Descriptor	15
7.3	Connection_Handle	15
7.4	Operational_Primitive	16
7.5	Transport Connection	16
8	Datamover layer and Datamover protocol	17
9	Operational Primitives provided by the Datamover layer	19
9.1	Send_Control	19
9.2	Put_Data	20
9.3	Get_Data	20
9.4	Allocate_Connection_Resources	21
9.5	Deallocate_Connection_Resources	22
9.6	Enable_Datamover	23
9.7	Connection_Terminate	23
9.8	Notice_Key_Values	24
9.9	Deallocate_Task_Resources	24
10	Operational Primitives provided by the iSCSI layer ...	25
10.1	Control_Notify.....	25
10.2	Connection_Terminate_Notify.....	26
10.3	Data_Completion_Notify.....	26
10.4	Data_ACK_Notify.....	27
11	Datamover Interface (DI)	29
11.1	Overview.....	29
11.2	Interactions for handling asynchronous notifications.	29
11.2.1	Connection termination.....	29
11.2.2	Data transfer completion.....	29
11.2.3	Data acknowledgement.....	30
11.3	Interactions for sending an iSCSI PDU.....	31
11.3.1	SCSI Command.....	31
11.3.2	SCSI Response.....	32
11.3.3	Task Management Function Request.....	32
11.3.4	Task Management Function Response.....	32
11.3.5	SCSI Data-out & SCSI Data-in.....	33
11.3.6	Ready To Transfer (R2T).....	33
11.3.7	Asynchronous Message.....	34
11.3.8	Text Request.....	34
11.3.9	Text Response.....	34

11.3.10	Login Request	34
11.3.11	Login Response	35
11.3.12	Logout Command	36
11.3.13	Logout Response	36
11.3.14	SNACK Request	36
11.3.15	Reject	36
11.3.16	NOP-Out	37
11.3.17	NOP-In	37
11.4	Interactions for receiving an iSCSI PDU.....	37
11.4.1	SCSI Command.....	37
11.4.2	SCSI Response.....	38
11.4.3	Task Management Function Request.....	38
11.4.4	Task Management Function Response.....	38
11.4.5	SCSI Data-out & SCSI Data-in.....	38
11.4.6	Ready To Transfer (R2T).....	39
11.4.7	Asynchronous Message.....	39
11.4.8	Text Request.....	39
11.4.9	Text Response.....	40
11.4.10	Login Request	40
11.4.11	Login Response	40
11.4.12	Logout Command	40
11.4.13	Logout Response	41
11.4.14	SNACK Request	41
11.4.15	Reject	41
11.4.16	NOP-Out	41
11.4.17	NOP-In	41
12	Security Considerations	42
13	IANA Considerations	43
14	References and Bibliography	44
14.1	Normative References.....	44
14.2	Informative References.....	44
15	Authors' Addresses	45
16	Acknowledgements	46
17	Appendix	50
17.1	Design considerations for a Datamover protocol.....	50
17.2	Examples of Datamover interactions.....	51
18	Full Copyright Statement	60
19	Intellectual Property Statement	61

Table of Figures

Figure 1	Datamover Architecture diagram, with the RDMAP example.....	10
Figure 2	A successful iSCSI login on initiator.....	52
Figure 3	A successful iSCSI login on target.....	52
Figure 4	A failed iSCSI login on initiator.....	53
Figure 5	A failed iSCSI login on target.....	53
Figure 6	iSCSI does not enable the Datamover.....	54
Figure 7	A normal iSCSI connection termination.....	55

Figure 8 An abnormal iSCSI connection termination.....55
Figure 9 A SCSI Write data transfer.....56
Figure 10 A SCSI Read data transfer.....57
Figure 11 A SCSI Read data acknowledgement.....58
Figure 12 Task resource cleanup on abort.....59

3 Definitions and acronyms

3.1 Definitions

I/O Buffer - A buffer that is used in a SCSI Read or Write operation so SCSI data may be sent from or received into that buffer.

Datamover protocol - A Datamover protocol is a data transfer wire protocol for iSCSI that meets the requirements stated in section 8.

Datamover layer - A Datamover layer is a protocol layer within an end node that implements the Datamover protocol.

Datamover-assisted - An iSCSI connection is said to be "Datamover-assisted" when a Datamover layer is enabled for moving control and data information on that connection.

3.2 Acronyms

Acronym	Definition

DA	Datamover Architecture
DDP	Direct Data Placement Protocol
DI	Datamover Interface
IANA	Internet Assigned Numbers Authority
IETF	Internet Engineering Task Force
I/O	Input - Output
IP	Internet Protocol
iSCSI	Internet SCSI
iSER	iSCSI Extensions for RDMA
ITT	Initiator Task Tag
LO	Leading Only

MPA	Marker PDU Aligned Framing for TCP
PDU	Protocol Data Unit
RDDP	Remote Direct Data Placement
RDMA	Remote Direct Memory Access
R2T	Ready To Transfer
R2TSN	Ready To Transfer Sequence Number
RDMA	Remote Direct Memory Access
RDMAP	Remote Direct Memory Access Protocol
RFC	Request For Comments
SAM	SCSI Architecture Model
SCSI	Small Computer Systems Interface
SN	Sequence Number
SNACK	Selective Negative Acknowledgment - also Sequence Number Acknowledgement for data
TCP	Transmission Control Protocol
TTT	Target Transfer Tag

4 Motivation

There are new industry and standards initiatives to develop Remote Direct Memory Access (RDMA) and Remote Direct Data Placement (RDDP) technologies to work over IP fabrics. The principal value proposition of these technologies is that they enable one end node to place data in the final intended buffer on the remote end node, thus eliminating the data copy that traditionally happens in the receive path to move the data to the final buffer. The data copy avoidance in turn eliminates unnecessary memory bandwidth consumption, substantially decreases the reassembly buffer size requirements, and preserves CPU cycles that would otherwise be spent in copying.

The iSCSI specification ([RFC3720]) defines a very detailed data transfer model that employs SCSI Data-In PDUs, SCSI Data-Out PDUs, and R2T PDUs, in addition to the SCSI Command and SCSI Response PDUs that respectively create and conclude the task context for the data transfer. In the traditional iSCSI model, the iSCSI protocol layer plays the central role in pacing the data transfer and carrying out the ensuing data transfer itself. An alternative architecture would be for iSCSI to delegate a large part of this data transfer role to a separate protocol layer exclusively designed to move data, which in turn is possibly aided by a data movement and placement technology such as RDMA.

If iSCSI were operating in such RDMA environments, iSCSI would be shielded from the low-level data transfer mechanics but would only be privy to the conclusion of the requested data transfer. Thus, there would be an effective "off-loading" of the work that an iSCSI protocol layer is expected to perform, compared to today's iSCSI end nodes. For such RDMA environments, it is highly desirable that there be a standard architecture to separate the data movement part of the iSCSI protocol definition from the rest of the iSCSI functionality. This architecture precisely defines what a Datamover layer is and also describes the model of interactions between the iSCSI layer and the Datamover layer (section 8). In order to satisfy this need, this document presents a Datamover Architecture (DA) and also summarizes a reasonable model for interactions between the iSCSI layer and the Datamover layer for each of the iSCSI PDUs that are defined in [RFC3720]. Note that while DA is motivated by the advent of RDMA over TCP/IP technology, the architecture is

not dependent on RDMA in its design. DA is intended to be a generic architectural framework for allowing different types of Datamovers based on different types of RDMA and transport protocols. Adoption of this model will help iSCSI proliferate into more environments.

5 Architectural layering of iSCSI and Datamover layers

Figure 1 illustrates an example of the architectural layering of iSCSI and Datamover layers, in conjunction with a TCP/IP implementation of RDMAP/DDP layers in an iSCSI end node. Note that RDMAP/DDP/MPA, and TCP protocol layers are shown here only as an example and in reality, DA is completely oblivious to protocol layers below the Datamover layer. The RDMAP/DDP/MPA protocol stack provides a generic transport service with direct data placement. There is no need to tailor the implementation of this protocol stack to the specific ULP to benefit from these services.

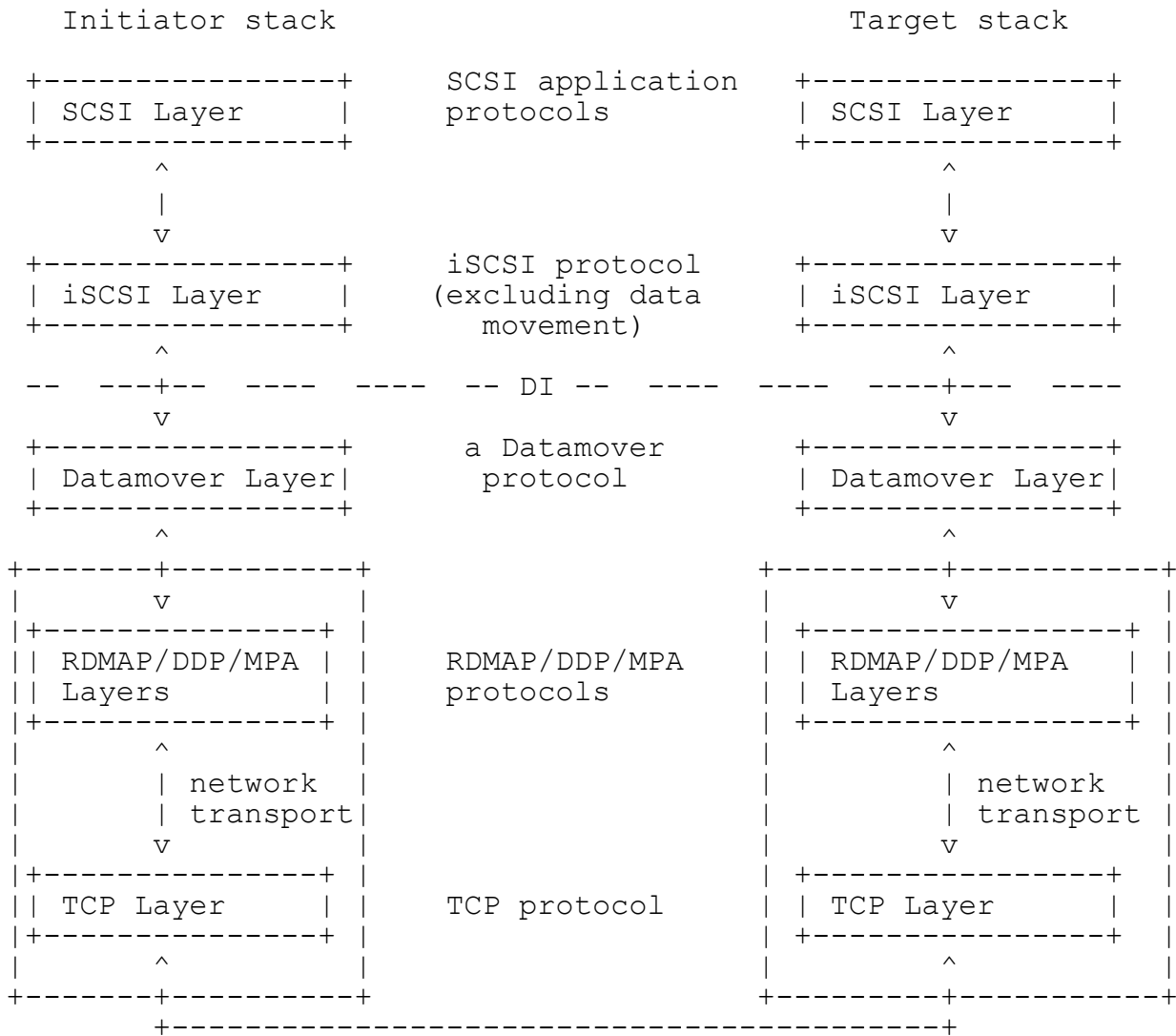


Figure 1 Datamover Architecture diagram, with the RDMAP example

The scope of this document is limited to:

1. Defining the notion of a Datamover layer and a Datamover protocol (section 8),
2. Defining the functionality distribution between the iSCSI layer and the Datamover layer along with the communication model between the two (Operational Primitives), and,
3. Modeling the interactions between the blocks labeled as "iSCSI Layer" and "Datamover Layer" in Figure 1 - i.e. defining the interface labeled as "DI" in the figure - for each defined iSCSI PDU, based on the Operational Primitives.

6 Design Overview

This document discusses and defines a model for interactions between the iSCSI layer and a "Datamover layer" (see section 8) operating within an iSCSI end node, presumably communicating with one or more iSCSI end nodes with similar layering. The model for interactions for handling different iSCSI operations is called the "Datamover Interface" (DI, section 11), while the architecture itself is called "Datamover Architecture" (DA). It is likely that the architecture will have implications on the Datamover wire protocols as DA places certain requirements and functionality expectations on the Datamover layer. However, this document itself neither defines any new wire protocol for the Datamover layer, nor any potential modifications to the iSCSI wire protocol to employ the Datamover layer. The scope of this document is strictly limited to specifying the architectural framework and the minimally required interactions that happen within an iSCSI end node to leverage the Datamover layer.

The design ideas behind DA can be summarized thus -

- 1) DA defines an abstract procedural interface definition of iSCSI layer's interactions with a Datamover layer below - i.e. DA models the interactions between the logical "bottom" interface of iSCSI and the logical "top" interface of a Datamover.
- 2) DA guides the wire protocol for a Datamover layer by defining the iSCSI knowledge that the Datamover layer may utilize in its protocol definition (as an example, this draft completely limits the notion of "iSCSI session" to the iSCSI layer).
- 3) DA is designed to allow implementing the Datamover layer either in hardware or in software.
- 4) DA is not a wire protocol spec, but an architecture that also models the interactions between iSCSI and Datamover layers operating within an iSCSI end node.
- 5) DA by design seeks to model the iSCSI-Datamover interactions in a way that the modeling is independent of the specifics of either a particular iSCSI revision, or a specific instantiation of a Datamover layer.

- 6) DA introduces and relies on the notion of a defined set of Operational Primitives (could be seen as entry point definitions in implementation terms) provided by each layer to the other to carry out the request-response interactions.
- 7) DA is intended to allow Datamover protocol definitions with minimal changes to existing iSCSI implementations.
- 8) DA is designed to allow the iSCSI layer to completely rely on the Datamover layer for all the data transport needs.
- 9) DA models the architecturally required minimal interactions between an operational iSCSI layer and a Datamover layer to realize the iSCSI-transparent data movement. There may be several other interactions in a typical implementation in order to bootstrap a Datamover layer (or an iSCSI layer) into operation, and they are outside the scope of this document.

Note that in summary, DA is architected to support many different Datamover protocols operating under the iSCSI layer. One such example of a Datamover protocol is iSER ([iSER]).

7 Architectural Concepts

7.1 iSCSI PDU types

This section defines the iSCSI PDU classification terminology, as defined and used in this document. Out of the set of legal iSCSI PDUs defined in [RFC3720], as we will see in section 7.1.1, the iSCSI layer does not request a SCSI Data-Out PDU carrying solicited data for transmission across the Datamover Interface per this architecture. For this reason, the SCSI Data-Out PDU carrying solicited data is excluded in the iSCSI PDU classification we introduce in this section. The rest of the legal iSCSI PDUs that may be exchanged across the Datamover Interface are defined to consist of two classes:

- 1) iSCSI data-type PDUs
- 2) iSCSI control-type PDUs

7.1.1 iSCSI data-type PDUs

An iSCSI data-type PDU is defined as an iSCSI PDU that causes data transfer, transparent to the remote iSCSI layer, to take place between the peer iSCSI nodes on a full feature phase iSCSI connection. A data-type PDU, when requested for transmission by the sender iSCSI layer, results in the associated data transfer without the participation of the remote iSCSI layer, i.e. the PDU itself is not delivered as-is to the remote iSCSI layer. The following iSCSI PDUs constitute the set of iSCSI data-type PDUs -

- 1) SCSI Data-In PDU
- 2) R2T PDU

In an iSCSI end node structured as an iSCSI layer and a Datamover layer as defined in this document, the solicitation for Data-out (i.e. R2T PDU) is not delivered to the initiator iSCSI layer, per the definition of an iSCSI data-type PDU. The data transfer is instead performed via the mechanisms known to the Datamover layer (e.g. RDMA Read). This in turn implies that a SCSI Data-Out PDU for solicited data is never requested for transmission across the Datamover Interface at the initiator.

7.1.2 iSCSI control-type PDUs

Any iSCSI PDU that is not an iSCSI data-type PDU and also not a solicited SCSI Data-out PDU is defined as an iSCSI control-type PDU. Specifically, it is to be noted that SCSI Data-Out PDUs for unsolicited Data are defined as iSCSI control-type PDUs.

7.2 Data_Descriptor

A Data Descriptor is an information element that describes an iSCSI/SCSI data buffer, provided by the iSCSI layer to its local Datamover layer or by the Datamover layer to its local iSCSI layer for identifying the data associated respectively with the requested or completed operation.

In implementation terms, a Data Descriptor may be a scatter-gather list describing a local buffer, the exact structure of which is subject to the constraints imposed by the operating environment on the local iSCSI node.

7.3 Connection_Handle

A Connection Handle is an information element that identifies the particular iSCSI connection for which an inbound or outbound iSCSI PDU is intended. A connection handle is unique for a given pair of an iSCSI layer instance and a Datamover layer instance. The Connection_Handle qualifier is used in all invocations of any Operational Primitive for connection identification.

Note that the Connection_Handle is conceptually different from the Connection Identifier (CID) defined by the iSCSI specification. While the CID is a unique identifier of an iSCSI connection within an iSCSI session, the uniqueness of the Connection_Handle extends to the entire iSCSI layer instance coupled with the Datamover layer instance, across possibly multiple iSCSI sessions.

In implementation terms, a Connection Handle could be an opaque identifier exchanged between the iSCSI layer and the Datamover layer at the connection login time. One may also consider it to be similar in scope of uniqueness to a socket identifier. The exact structure and modalities of exchange of a Connection Handle between the two layers is implementation-specific.

7.4 Operational Primitive

An Operational Primitive, in this document, is an abstract functional interface procedure that requests another layer to perform a specific action on the requestor's behalf or notifies the other layer of some event. The Datamover Interface between an iSCSI layer instance and a Datamover layer instance within an iSCSI end node uses a set of Operational Primitives to define the functional interface between the two layers. Note that not every invocation of an Operational Primitive may elicit a response from the requested layer. This document describes the types of Operational Primitives that are implicitly required and provided by the iSCSI protocol layer as defined in [RFC3720], and the semantics of these Primitives.

Note that ownership of buffers and data structures is likely to be exchanged between the iSCSI layer and its local Datamover layer in invoking the Operational Primitives defined in this architecture. The buffer management details, including how buffers are allocated and released, are implementation-specific and thus are outside the scope of this document.

7.5 Transport Connection

The term "Transport Connection" is used in this document as a generic term to represent the end-to-end logical connection as defined by the underlying reliable transport protocol. For this revision of this document, a Transport Connection means only a TCP connection.

8 Datamover layer and Datamover protocol

This section introduces the notion of a "Datamover layer" and "Datamover protocol" as meant in this document, and defines the requirements on a Datamover protocol.

A Datamover layer is the implementation component that realizes a Datamover protocol functionality in an iSCSI-capable end node, in communicating with other iSCSI end nodes with similar capabilities. More specifically, a "Datamover layer" MUST provide the following functionality and the "Datamover protocol" MUST consist of the wire protocol required to realize the following functionality -

- 1) guarantee that all the necessary data transfers take place when the local iSCSI layer requests transmitting a command (in order to complete a SCSI command, for an initiator), or sending/receiving an iSCSI data sequence (in order to complete part of a SCSI command, for a target).
- 2) transport an iSCSI control-type PDU as-is to the peer Datamover layer when requested to do so by the local iSCSI layer.
- 3) provide notification and delivery to the iSCSI layer upon arrival of an iSCSI control-type PDU.
- 4) provide an end-to-end data acknowledgement of SCSI read data to the target iSCSI layer, when requested.
- 5) provide an asynchronous notification upon completion of a requested data transfer operation that moved data without involving the iSCSI layer.
- 6) place the SCSI data into the I/O buffers or pick up the SCSI data for transmission out of the data buffers that the iSCSI layer had requested to be used for a SCSI I/O.
- 7) guarantee an error-free (i.e. must have at least the same level of assurance of data integrity as the CRC32C iSCSI data digest), reliable, in-order delivery transport mechanism over IP fabrics in performing the data transfer, and asynchronously notify the iSCSI layer upon iSCSI connection termination.

Note that this architecture expects that each compliant Datamover protocol will define the precise means of satisfying the requirements specified in this section.

9 Operational Primitives provided by the Datamover layer

While the iSCSI specification itself does not have a notion of Operational Primitives, any iSCSI layer implementing the iSCSI specification functionally requires the following Operational Primitives from its Datamover layer. These Operational Primitives are invoked by the iSCSI layer as appropriate. Unless otherwise stated, all the following Operational Primitives may be used both on the initiator side and the target side. In general programming terminology, this set of Operational Primitives may be construed as "down calls".

- 1) Send_Control
- 2) Put_Data
- 3) Get_Data
- 4) Allocate_Connection_Resources
- 5) Deallocate_Connection_Resources
- 6) Enable_Datamover
- 7) Connection_Terminate
- 8) Notice_Key_Values
- 9) Deallocate_Task_Resources

9.1 Send_Control

Input qualifiers: Connection_Handle, iSCSI PDU-specific qualifiers

Return Results: Not specified.

An iSCSI layer requests its local Datamover layer to transmit an iSCSI control-type PDU to the peer iSCSI layer operating in the remote iSCSI node by this Operational Primitive. The Datamover layer performs the requested operation, and may add its own protocol headers in doing so. The iSCSI layer MUST NOT invoke the Send_Control Operational Primitive on an iSCSI connection that is not yet Datamover-assisted.

An initiator iSCSI layer requesting the transfer of a SCSI command PDU or a target iSCSI layer requesting the transfer of a SCSI response PDU are examples of invoking the Send_Control Operational Primitive.

9.2 Put_Data

Input qualifiers: Connection_Handle, contents of a SCSI Data-In PDU header, Data_Descriptor, Notify_Enable

Return Results: Not specified.

An iSCSI layer requests its local Datamover layer to transmit the data identified by the Data_Descriptor for the SCSI Data-In PDU to the peer iSCSI layer on the remote iSCSI node by this Operational Primitive. The Datamover layer performs the operation by using its own protocol means, completely transparent to the remote iSCSI layer. The iSCSI layer MUST NOT invoke the Put_Data Operational Primitive on an iSCSI connection that is not yet Datamover-assisted.

The Notify_Enable qualifier is used to request the local Datamover layer to generate or to not generate the eventual local completion notification to the iSCSI layer for this Put_Data invocation. For detailed semantics of this qualifier, see section 10.3.

A Put_Data Primitive may only be invoked by an iSCSI layer on the target to its local Datamover layer.

A target iSCSI layer requesting the transfer of an iSCSI read data sequence (also known as a read burst) is an example of invoking the Put_Data Operational Primitive.

9.3 Get_Data

Input qualifiers: Connection_Handle, contents of an R2T PDU, Data_Descriptor, Notify_Enable

Return Results: Not specified.

An iSCSI layer requests its local Datamover layer to retrieve certain data identified by the R2T PDU from the peer iSCSI layer on the remote iSCSI node into the buffer identified by the Data_Descriptor by invoking this Operational Primitive. The Datamover layer performs the operation by using its own protocol means, completely transparent to the remote iSCSI layer. The iSCSI layer MUST NOT invoke the Get_Data Operational Primitive on an iSCSI connection that is not yet Datamover-assisted.

The Notify_Enable qualifier is used to request the local Datamover layer to generate or to not generate the eventual local completion notification to the iSCSI layer for this Get_Data invocation. For detailed semantics of this qualifier, see section 10.3.

A Get_Data Primitive may only be invoked by an iSCSI layer on the target to its local Datamover layer.

A target iSCSI layer requesting the transfer of an iSCSI write data sequence (also known as a write burst) is an example of invoking the Get_Data Operational Primitive.

9.4 Allocate_Connection_Resources

Input qualifiers: Connection_Handle[, Resource_Descriptor]

Return Results: Status.

By invoking this Operational Primitive, an iSCSI layer requests its local Datamover layer to perform all the Datamover-specific resource allocations required for the full feature phase of an iSCSI connection. The Connection_Handle identifies the connection the iSCSI layer is requesting the resource allocation for in order to eventually transition the connection to be a Datamover-assisted iSCSI connection. Note that the Datamover layer however does not allocate any Datamover-specific task-level resources upon invocation of this Primitive.

An iSCSI layer, in addition, optionally specifies the implementation-specific resource requirements for the iSCSI connection to the Datamover layer, by passing an input qualifier called Resource_Descriptor. The exact structure of a Resource_Descriptor is implementation-dependent, and hence structurally opaque to DA.

A return result of Status=success means that the Allocate_Connection_Resources invocation corresponding to that Connection_Handle succeeded. If an Allocate_Connection_Resources invocation is made for a Connection_Handle for which an earlier invocation succeeded, the return_Status must be success and the request will be ignored by the Datamover layer. A return result of Status=failure means that the Allocate_Connection_Resources invocation corresponding to that Connection_Handle failed. There MUST NOT be more than one Allocate_Connection_Resources Primitive invocation outstanding for a given Connection_Handle at any time.

The iSCSI layer must invoke the Allocate_Connection_Resources Primitive before the invocation of the Enable_Datamover Primitive.

9.5 Deallocate_Connection_Resources

Input qualifiers: Connection_Handle

Return Results: Not specified.

By invoking this Operational Primitive, an iSCSI layer requests its local Datamover layer to deallocate all the Datamover-specific resources that may have been allocated earlier for the Transport Connection identified by the Connection_Handle. The iSCSI layer may invoke this Operational Primitive when the Datamover-specific resources associated with the Connection_Handle are no longer necessary (such as the Login failure of the corresponding iSCSI connection).

9.6 Enable_Datamover

Input qualifiers: Connection_Handle,
Transport_Connection_Descriptor [, Final_Login_Response_PDU]

Return Results: Not specified.

By invoking this Operational Primitive, an iSCSI layer requests its local Datamover layer to assist all further iSCSI exchanges on the iSCSI connection (i.e. to make the connection Datamover-assisted) identified by the Connection_Handle, for which the Datamover-specific resource allocation was earlier made. The iSCSI layer MUST NOT invoke the Enable_Datamover Operational Primitive for an iSCSI connection unless there was a corresponding prior resource allocation.

The Final_Login_Response_PDU input qualifier is applicable only for a target, and contains the final Login Response that concludes the iSCSI Login phase and which must be sent as a byte stream as expected by the initiator iSCSI layer. When this qualifier is used, the target-Datamover layer MUST transmit this final Login Response before Datamover assistance is enabled for the Transport Connection.

The iSCSI layer identifies the specific Transport Connection associated with the Connection_Handle to the Datamover layer by specifying the Transport_Connection_Descriptor. The exact structure of this Descriptor is implementation-dependent.

9.7 Connection_Terminate

Input qualifiers: Connection_Handle

Return Results: Not specified.

By invoking this Operational Primitive, an iSCSI layer requests its local Datamover layer to terminate the Transport Connection and deallocate all the connection and task resources associated with the Connection_Handle. When this Operational Primitive invocation returns to the iSCSI layer, the iSCSI layer may assume the full ownership of all the iSCSI-level resources, e.g. I/O Buffers, associated with the connection. This Operational Primitive may be invoked only

with a valid `Connection_Handle` and the Transport Connection associated with the `Connection_Handle` must already be Datamover-assisted.

9.8 Notice_Key_Values

Input qualifiers: `Connection_Handle`, Number of keys, a list of Key-Value pairs

Return Results: Not specified.

By invoking this Operational Primitive, an iSCSI layer requests its local Datamover layer to take note of the negotiated values of the listed keys for the Transport Connection. This Operational Primitive may be invoked only with a valid `Connection_Handle` and the Key-Value pairs MUST be the current values that were successfully agreed upon by the iSCSI peers for the connection. The Datamover layer may use the values of the keys to aid the Datamover operation as it deems appropriate. The specific keys to be passed in as input qualifiers and the point(s) in time this Operational Primitive is invoked are implementation-dependent.

9.9 Deallocate_Task_Resources

Input qualifiers: `Connection_Handle`, ITT

Return Results: Not specified.

By invoking this Operational Primitive, an iSCSI layer requests its local Datamover Layer to deallocate all Datamover-specific resources that earlier may have been allocated for the task identified by the ITT qualifier. The iSCSI layer uses this Operational Primitive during exception processing when one or more active tasks are to be terminated without corresponding SCSI Response PDUs. This Primitive MUST be invoked for each active task terminated without a SCSI Response PDU. This Primitive MUST NOT be invoked by the iSCSI layer when a SCSI Response PDU normally concludes a task. When a SCSI Response PDU normally concludes a task (even if the SCSI Status was not a success), the Datamover layer is assumed to have automatically deallocated all Datamover-specific task resources for that task.

10 Operational Primitives provided by the iSCSI layer

While the iSCSI specification itself does not have a notion of Operational Primitives, any iSCSI layer implementing the iSCSI specification would have to provide the following Operational Primitives to its local Datamover layer. These Operational Primitives are invoked by the Datamover layer as appropriate and when the iSCSI connection is Datamover-assisted. Unless otherwise stated, all the following Operational Primitives may be used both on the initiator side and the target side. In general programming terminology, this set of Operational Primitives may be construed as "up calls".

- 1) Control_Notify
- 2) Connection_Terminate_Notify
- 3) Data_Completion_Notify
- 4) Data_ACK_Notify

10.1 Control_Notify

Input qualifiers: Connection_Handle, an iSCSI control-type PDU.

Return Results: Not specified.

A Datamover layer notifies its local iSCSI layer, via this Operational Primitive, of the arrival of an iSCSI control-type PDU from the peer Datamover layer on the remote iSCSI node. The iSCSI layer processes the control-type PDU as defined in [RFC3720].

A target iSCSI layer being notified of the arrival of a SCSI Command is an example of invoking the Control_Notify Operational Primitive.

Note that implementations may choose to describe the "iSCSI control-type PDU" qualifier in this notification using a

Data_Descriptor (section 7.2) and not necessarily one contiguous buffer.

10.2 Connection_Terminate_Notify

Input qualifiers: Connection_Handle

Return Results: Not specified.

A Datamover layer notifies its local iSCSI layer on an unsolicited termination or failure of an iSCSI connection providing the Connection_Handle associated with the iSCSI Connection. The iSCSI Layer MUST consider the Connection_Handle to be invalid upon being so notified. The iSCSI layer processes the connection termination as defined in [RFC3720]. The Datamover layer MUST deallocate the connection and task resources associated with the terminated connection before notifying the iSCSI layer of the termination via this Operational Primitive.

A target iSCSI layer being notified of the arrival of TCP RESET is an example of when the Connection_Terminate_Notify Operational Primitive is invoked.

10.3 Data_Completion_Notify

Input qualifiers: Connection_Handle, ITT, SN

Return Results: Not specified.

A Datamover layer notifies its local iSCSI layer on completing the retrieval of the data or upon sending the data, as requested in a prior iSCSI data-type PDU, from/to the peer Datamover layer on the remote iSCSI node via this Operational Primitive. The iSCSI layer processes the operation as defined in [RFC3720].

SN may be either the DataSN associated with the SCSI Data-In PDU or R2TSN associated with the R2T PDU depending on the SCSI operation. Note that, for targets, a TTT (see [RFC3720]) could have been specified instead of an SN. However, the considered choice was to leave the SN to be the

qualifier for two reasons - a) it is generic and applicable to initiators and targets as well as Data-in and Data-out, and b) having both SN and TTT qualifiers for the notification was considered onerous on the Datamover layer, in terms of state maintenance for each completion notification. The implication of this choice is that iSCSI target implementations will have to adapt to using the ITT-SN tuple in associating the solicited data to the appropriate task, rather than the ITT-TTT tuple for doing the same.

If `Notify_Enable` was set in either a `Put_Data` or a `Get_Data` invocation, the Datamover layer MUST invoke the `Data_Completion_Notify` Operational Primitive upon completing that requested data transfer. If the `Notify_Enable` was cleared in either a `Put_Data` or a `Get_Data` invocation, the Datamover layer MUST NOT invoke the `Data_Completion_Notify` Operational Primitive upon completing that requested data transfer.

A `Data_Completion_Notify` invocation serves to notify the iSCSI layer of the `Put_Data` or `Get_Data` completion respectively. As earlier noted in sections 9.2 and 9.3, specific Datamover protocol definitions may restrict the usage scope of `Put_Data` and `Get_Data`, and thus implicitly the usage scope of `Data_Completion_Notify`.

A target iSCSI layer being notified of the retrieval of a write data sequence is an example of invoking the `Data_Completion_Notify` Operational Primitive.

10.4 `Data_ACK_Notify`

Input qualifiers: `Connection_Handle`, `ITT`, `DataSN`

Return Results: Not specified.

A target Datamover layer notifies its local iSCSI layer of the arrival of a previously requested data acknowledgement from the peer Datamover layer on the remote (initiator) iSCSI node via this Operational Primitive. The iSCSI layer processes the data acknowledgement notification as defined in [RFC3720].

A target iSCSI layer being notified of the arrival of a data acknowledgement for a certain SCSI Read data PDU is the only example of invoking the Data_ACK_Notify Operational Primitive.

11 Datamover Interface (DI)

11.1 Overview

This chapter describes the interactions model between iSCSI and Datamover layers when the iSCSI connection is Datamover-assisted so the iSCSI layer may carry out the following -

- send iSCSI data-type PDUs and exchange iSCSI control-type PDUs, and
- handle asynchronous notifications such as completion of data sequence transfer, and connection failure.

This chapter relies on the notion of Operational Primitives (section 7.4) to define DI.

11.2 Interactions for handling asynchronous notifications

11.2.1 Connection termination

As stated in section 10.2, the Datamover layer notifies the iSCSI layer of a failed or terminated connection via the `Connection_Terminate_Notify` Operational Primitive. The iSCSI layer MUST consider the connection as unusable upon the invocation of this Primitive and handle the connection termination as specified in [RFC3720].

11.2.2 Data transfer completion

As stated in section 10.3, the Datamover layer notifies the iSCSI layer of a completed data transfer operation via the `Data_Completion_Notify` Operational Primitive. The iSCSI layer processes the transfer completion as specified in [RFC3720].

11.2.2.1 Completion of a requested SCSI Data transfer

The Datamover layer, to notify the iSCSI layer of the completion of a requested iSCSI data-type PDU transfer, uses the `Data_Completion_Notify` Operational Primitive with the following input qualifiers.

a) `Connection_Handle`

- b) ITT: Initiator Task Tag semantics as defined in [RFC3720]
- c) SN: DataSN for a SCSI Data-in/Data-out PDU, and R2TSN for an iSCSI R2T PDU. The semantics for both types of sequence numbers are as defined in [RFC3720].

The rationale for choosing SN is explained in section 10.3.

Every invocation of the Data_Completion_Notify Operational Primitive MUST be preceded by an invocation of the Put_Data or Get_Data Operational Primitive with the Notify_Enable qualifier set by the iSCSI layer at an earlier point in time.

11.2.3 Data acknowledgement

[RFC3720] allows the iSCSI targets to optionally solicit data acknowledgement from the initiator for one or more Data-in PDUs, via setting of the A-bit on a Data-in PDU. The Data_ACK_Notify Operational Primitive with the following input qualifiers is used by the target Datamover layer to notify the local iSCSI layer of the arrival of data acknowledgement of a previously solicited iSCSI read data acknowledgement. This Operational Primitive thus is applicable only to iSCSI targets.

- a) Connection_Handle
- b) ITT: Initiator Task Tag semantics as defined in [RFC3720]
- c) DataSN: of the next SCSI Data-in PDU which immediately follows the SCSI Data-in PDU with the A-bit set to which this notification corresponds, with semantics as defined in [RFC3720].

Every invocation of the Data_ACK_Notify Operational Primitive MUST be preceded by an invocation of the Put_Data Operational Primitive by the iSCSI target layer with the A-bit set to 1 at an earlier point in time.

11.3 Interactions for sending an iSCSI PDU

This section discusses the interactions model for sending each of the iSCSI PDUs defined in [RFC3720]. A Connection_Handle (see section 7.3) is assumed to qualify each of these interactions so that the Datamover layer can route it to the appropriate Transport Connection. The qualifying Connection_Handle is not explicitly listed in the subsequent sections.

11.3.1 SCSI Command

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a SCSI Command PDU.

- a) BHS and AHS, if any, of the SCSI Command PDU as defined in [RFC3720]
- b) DataDescriptorOut: that defines the I/O Buffer meant for Data-out for the entire command, in the case of a write or bidirectional command
- c) DataDescriptorIn: that defines the I/O Buffer meant for Data-in for the entire command, in the case of a read or bidirectional command
- d) ImmediateDataSize: that defines the number of octets of immediate unsolicited data for a write/bidirectional command
- e) UnsolicitedDataSize: that defines the number of octets of immediate and non-immediate unsolicited data for a write/bidirectional command.

Note that some implementations may choose to ascertain the ImmediateDataSize input qualifier from the DataSegmentLength field of the SCSI Command PDU. The defined list of input qualifiers represents the semantically required set for the Datamover layer to consider in implementing the Primitive.

11.3.2 SCSI Response

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a SCSI Response PDU.

- a) BHS of the SCSI Response PDU as defined in [RFC3720]
- b) DataDescriptorStatus: that defines the iSCSI buffer which contains the sense and response information for the command

11.3.3 Task Management Function Request

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a Task Management Function Request PDU.

- a) BHS of the Task Management Function Request PDU as defined in [RFC3720]
- b) DataDescriptorOut: that defines the I/O Buffer meant for Data-out for the entire command, in the case of a write or bidirectional command (Only valid if Function="TASK REASSIGN" - [RFC3720])
- c) DataDescriptorIn: that defines the I/O Buffer meant for Data-in for the entire command, in the case of a read or bidirectional command (Only valid if Function="TASK REASSIGN" - [RFC3720])

11.3.4 Task Management Function Response

The Send_Control Operational Primitive with the following input qualifier is used for requesting the transmission of a Task Management Function Response PDU.

- a) BHS of the Task Management Function Response PDU as defined in [RFC3720]

11.3.5 SCSI Data-out & SCSI Data-in

11.3.5.1 SCSI Data-out

The Send_Control Operational Primitive with the following input qualifiers is used by the initiator iSCSI layer for requesting the transmission of a SCSI Data-out PDU carrying the non-immediate unsolicited data.

- a) BHS of the SCSI Data-out PDU as defined in [RFC3720]
- b) DataDescriptorOut: that defines the I/O Buffer with the Data-out to be carried in the iSCSI data segment of the PDU

11.3.5.2 SCSI Data-in

The Put_Data Operational Primitive with the following input qualifiers is used by the target iSCSI layer for requesting the transmission of the data carried by a SCSI Data-in PDU.

- a) BHS of the SCSI Data-in PDU as defined in [RFC3720]
- b) DataDescriptorIn: that defines the I/O Buffer with the Data-in being requested for transmission

11.3.6 Ready To Transfer (R2T)

The Get_Data Operational Primitive with the following input qualifiers is used by the target iSCSI layer for requesting the retrieval of the data as specified by the semantic content of an R2T PDU.

- a) BHS of the Ready To Transfer PDU as defined in [RFC3720]
- b) DataDescriptorOut: that defines the I/O Buffer for the Data-out being requested for retrieval

11.3.7 Asynchronous Message

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of an Asynchronous Message PDU.

- a) BHS of the Asynchronous Message PDU as defined in [RFC3720]
- b) DataDescriptorSense: that defines an iSCSI buffer which contains the sense and iSCSI Event information.

11.3.8 Text Request

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a Text Request PDU.

- a) BHS of the Text Request PDU as defined in [RFC3720]
- b) DataDescriptorTextOut: that defines the iSCSI Text Request buffer

11.3.9 Text Response

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a Text Response PDU.

- a) BHS of the Text Response PDU as defined in [RFC3720]
- b) DataDescriptorTextIn: that defines the iSCSI Text Response buffer

11.3.10 Login Request

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a Login Request PDU.

- a) BHS of the Login Request PDU as defined in [RFC3720]
- b) DataDescriptorLoginRequest: that defines the iSCSI Login Request buffer

Note that specific Datamover protocols may choose to disallow the standard DA Primitives from being used for the iSCSI Login phase. When used in conjunction with such Datamover protocols, an attempt to send a Login Request via the Send_Control Operational Primitive invocation is clearly an error scenario, as the Login Request PDU is being sent while the connection is in the iSCSI full feature phase. It is outside the scope of this document to specify the resulting implementation behavior in this case - [RFC3720] already defines the error handling for this error scenario.

11.3.11 Login Response

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a Login Response PDU.

- a) BHS of the Login Response PDU as defined in [RFC3720]
- b) DataDescriptorLoginResponse: that defines the iSCSI Login Response buffer

Note that specific Datamover protocols may choose to disallow the standard DA Primitives from being used for the iSCSI Login phase. When used in conjunction with such Datamover protocols, an attempt to send a Login Response via the Send_Control Operational Primitive invocation is clearly an error scenario, as the Login Response PDU is being sent while in the iSCSI full feature phase. It is outside the scope of this document to specify the resulting implementation behavior in this case - [RFC3720] already defines the error handling for this error scenario.

11.3.12 Logout Command

The Send_Control_Operational_Primitive with the following input_qualifier is used for requesting the transmission of a Logout Command PDU.

- a) BHS of the Logout Command PDU as defined in [RFC3720]

11.3.13 Logout Response

The Send_Control_Operational_Primitive with the following input_qualifier is used for requesting the transmission of a Logout Response PDU.

- a) BHS of the Logout Response PDU as defined in [RFC3720]

11.3.14 SNACK Request

The Send_Control_Operational_Primitive with the following input_qualifier is used for requesting the transmission of a SNACK Request PDU.

- a) BHS of the SNACK Request PDU as defined in [RFC3720]

11.3.15 Reject

The Send_Control_Operational_Primitive with the following input_qualifiers is used for requesting the transmission of a Reject PDU.

- a) BHS of the Reject PDU as defined in [RFC3720]
- b) DataDescriptorReject: that defines the iSCSI Reject buffer

11.3.16 NOP-Out

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a NOP-Out PDU.

- a) BHS of the NOP-Out PDU as defined in [RFC3720]
- b) DataDescriptorNOPOut: that defines the iSCSI Ping data buffer

11.3.17 NOP-In

The Send_Control Operational Primitive with the following input qualifiers is used for requesting the transmission of a NOP-In PDU.

- a) BHS of the NOP-In PDU as defined in [RFC3720]
- b) DataDescriptorNOPIn: that defines the iSCSI Return Ping data buffer

11.4 Interactions for receiving an iSCSI PDU

The only PDUs that are received by an iSCSI layer operating on a Datamover layer are the iSCSI control-type PDUs. The Datamover layer delivers the iSCSI control-type PDUs as they arrive, qualifying each with the Connection_Handle (see section 7.3) that identifies the iSCSI connection the PDU is meant for. The subsequent processing of the iSCSI control-type PDUs proceeds as defined in [RFC3720].

11.4.1 SCSI Command

The Control_Notify Operational Primitive is used for notifying the arrival of a SCSI Command PDU.

11.4.2 SCSI Response

The Control_Notify_Operational_Primitive is used for notifying the arrival of a SCSI Response PDU.

11.4.3 Task Management Function Request

The Control_Notify_Operational_Primitive is used for notifying the arrival of a Task Management Function Request PDU.

11.4.4 Task Management Function Response

The Control_Notify_Operational_Primitive is used for notifying the arrival of a Task Management Function Response PDU.

11.4.5 SCSI Data-out & SCSI Data-in

11.4.5.1 SCSI Data-out

The Control_Notify_Operational_Primitive is used for notifying the iSCSI layer of the arrival of a SCSI Data-out PDU carrying the non-immediate unsolicited data. Note however that the solicited SCSI Data-out arriving on the target is not notified to the iSCSI layer using the Control_Notify_Primitive because the solicited SCSI Data-out was not sent by the initiator iSCSI layer as control-type PDUs.

11.4.5.2 SCSI Data-in

The arrival of the SCSI Data-in is not notified to the iSCSI layer by the Datamover layer at the initiator, because SCSI Data-in is an iSCSI data-type PDU (see section 7.1). The iSCSI layer at the initiator however may infer the arrival of the SCSI Data-in when it receives a subsequent notification of the SCSI Response PDU via a Control_Notify invocation.

While this document does not contemplate the possibility of a Data-in PDU being received at the initiator iSCSI layer, specific Datamover protocols may define how to deal with an unexpected inbound SCSI Data-in PDU that may result in the initiator iSCSI layer receiving the Data-in PDU. This document leaves the details of handling this error scenario to the specific Datamover protocols, so each may define the appropriate error handling specific to the Datamover environment.

11.4.6 Ready To Transfer (R2T)

Because an R2T PDU is an iSCSI data-type PDU (see section 7.1) that is not delivered as-is to the initiator iSCSI layer, the arrival of an R2T PDU is not notified to the iSCSI layer by the Datamover layer. When an iSCSI node sends an R2T PDU to its local Datamover layer, the local and remote Datamover layers transparently bring about the data transfer requested by the R2T PDU.

While this document does not contemplate the possibility of an R2T PDU being received at the initiator iSCSI layer, specific Datamover protocols may define how to deal with an unexpected inbound R2T PDU that may result in the initiator iSCSI layer receiving the R2T PDU. This document leaves the details of handling this error scenario to the specific Datamover protocols, so each may define the appropriate error handling specific to the Datamover environment.

11.4.7 Asynchronous Message

The Control Notify Operational Primitive is used for notifying the arrival of an Asynchronous Message PDU.

11.4.8 Text Request

The Control Notify Operational Primitive is used for notifying the arrival of a Text Request PDU.

11.4.9 Text Response

The Control_Notify_Operational_Primitive is used for notifying the arrival of a Text Response PDU.

11.4.10 Login Request

The Control_Notify_Operational_Primitive is used for notifying the target iSCSI layer of the arrival of a Login Request PDU. Note that specific Datamover protocols may choose to disallow the standard DA Primitives from being used for the iSCSI Login phase. When used in conjunction with such Datamover protocols, the arrival of a Login Request necessitating the Control_Notify_Operational_Primitive invocation is clearly an error scenario, as the Login Request PDU is arriving in the iSCSI full feature phase. It is outside the scope of this document to specify the resulting implementation behavior in this case - [RFC3720] already defines the error handling in this error scenario.

11.4.11 Login Response

The Control_Notify_Operational_Primitive is used for notifying the initiator iSCSI layer of the arrival of a Login Response PDU. Note that specific Datamover protocols may choose to disallow the standard DA Primitives from being used for the iSCSI Login phase. When used in conjunction with such Datamover protocols, the arrival of a Login Response necessitating the Control_Notify_Operational_Primitive invocation is clearly an error scenario, as the Login Response PDU is arriving in the iSCSI full feature phase. It is outside the scope of this document to specify the resulting implementation behavior in this case - [RFC3720] already defines the error handling in this error scenario.

11.4.12 Logout Command

The Control_Notify_Operational_Primitive is used for notifying the arrival of a Logout Command PDU.

11.4.13 Logout Response

The Control Notify Operational Primitive is used for notifying the arrival of a Logout Response PDU.

11.4.14 SNACK Request

The Control Notify Operational Primitive is used for notifying the arrival of a SNACK Request PDU.

11.4.15 Reject

The Control Notify Operational Primitive is used for notifying the arrival of a Reject PDU.

11.4.16 NOP-Out

The Control Notify Operational Primitive is used for notifying the arrival of a NOP-Out PDU.

11.4.17 NOP-In

The Control Notify Operational Primitive is used for notifying the arrival of a NOP-In PDU.

12 Security Considerations

In view of the fact that DA does not define any new wire protocol nor propose modifications to the existing protocols, there are no additional security considerations in employing DA, in addition to that of using the iSCSI protocol itself. Any additional security considerations resulting from the use of any Datamover protocol must be identified by the specific Datamover protocol specification as appropriate.

13 IANA Considerations

If a well-known port is chosen as the mechanism to identify a Datamover protocol on TCP, the well-known port must be registered with IANA. Because the use of the well-known port is specific to the Datamover protocol in such a case, the resulting IANA considerations from such use must be specified by the specific Datamover protocol. DA itself does not have any specific IANA considerations.

14 References and Bibliography

14.1 Normative References

[RFC3720] J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, E. Zeidner, "Internet Small Computer Systems Interface (iSCSI)", RFC 3720, April 2004.

14.2 Informative References

[DDP] H. Shah et al., "Direct Data Placement over Reliable Transports", IETF Internet Draft draft-ietf-rddp-ddp-00.txt (work in progress), February 2003.

[iSER] M. Ko et al., "iSCSI Extensions for RDMA", IETF Internet Draft draft-ko-iwarp-iser-02.txt (work in progress), July 2004.

[MPA] P. Culley et al., "Marker PDU Aligned Framing for TCP Specification", IETF Internet Draft draft-culley-iwarp-mpa-02.txt (work in progress), February 2003.

[RDMAP] R. Recio et al., "An RDMA Protocol Specification", IETF Internet Draft draft-ietf-rddp-rdmap-00.txt (work in progress), February 2003.

[RFC2119] Bradner, S. "Key Words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[SAM] ANSI X3.270-1998, SCSI-3 Architecture Model (SAM).

[SCTP] R. Stewart et al., "Stream Control Transmission Protocol", RFC 2960, October 2000.

[SPC3]T10/1416-D, SCSI Primary Commands-3.

15 Authors' Addresses

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

John L. Hufferd
IBM
San Jose CA, USA
Phone: +1-408-256-0403
E-mail: hufferd@us.ibm.com

Julian Satran
IBM, Haifa Research Lab
Haifa University Campus - Mount Carmel
Haifa 31905, Israel
Phone +972-4-829-6264
E-mail: Julian.Satran@il.ibm.com

Hemal Shah
Intel Corporation
MS PTL1
1501 South Mopac Expressway, #400
Austin, TX 78746 USA
Phone: +1 (512) 732-3963
Email: hemal.shah@intel.com

Comments may be sent to Mallikarjun Chadalapaka.

16 Acknowledgements

The IPS Working group in the Transport Area of IETF is responsible for defining the iSCSI protocol (apart from a host of other relevant IP Storage protocols). The authors are grateful to the entire working group, whose work allowed this document to build on the concepts and details of the iSCSI protocol.

In addition, the following individuals had reviewed and contributed to the improvement of this document. The authors are grateful for their contribution.

John Carrier
Adaptec, Inc.
691 S. Milpitas Blvd.
Milpitas, CA 95035 USA
Phone: +1 (360) 378-8526
Email: john_carrier@adaptec.com

Hari Ghadia
Adaptec, Inc.
691 S. Milpitas Blvd.,
Milpitas, CA 95035 USA
Phone: +1 (408) 957-5608
Email: hari_ghadia@adaptec.com

Hari Mudaliar
Adaptec, Inc.
691 S. Milpitas Blvd.,
Milpitas, CA 95035 USA
Phone: +1 (408) 957-6012
Email: hari_mudaliar@adaptec.com

Patricia Thaler
Agilent Technologies, Inc.
1101 Creekside Ridge Drive, #100
M/S-RG10
Roseville, CA 95678
Phone: +1-916-788-5662

email: pat_thaler@agilent.com

Uri Elzur
Broadcom Corporation
16215 Alton Parkway
Irvine, California 92619-7013 USA
Phone: +1 (949) 585-6432
Email: Uri@Broadcom.com

Mike Penna
Broadcom Corporation
16215 Alton Parkway
Irvine, California 92619-7013 USA
Phone: +1 (949) 926-7149
Email: MPenna@Broadcom.com

Ted Compton
EMC Corporation
Research Triangle Park, NC 27709, USA
Phone: 919-248-6075
Email: compton_ted@emc.com

Dwight Barron
Hewlett-Packard Company
20555 SH 249
Houston, TX 77070-2698 USA
Phone: +1 (281) 514-2769
Email: Dwight.Barron@Hp.com

Paul R. Culley
Hewlett-Packard Company
20555 SH 249
Houston, TX 77070-2698 USA
Phone: +1 (281) 514-5543
Email: paul.culley@hp.com

Dave Garcia
Hewlett-Packard Company
19333 Vallco Parkway
Cupertino, Ca. 95014 USA

Phone: +1 (408) 285-6116
Email: dave.garcia@hp.com

Randy Haagens
Hewlett-Packard Company
8000 Foothills Blvd, MS 5668
Roseville CA
Phone: +1-916-785-4578
email: randy_haagens@hp.com

Jeff Hilland
Hewlett-Packard Company
20555 SH 249
Houston, Tx. 77070-2698 USA
Phone: +1 (281) 514-9489
Email: jeff.hilland@hp.com

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

Jim Wendt
Hewlett-Packard Company
8000 Foothills Blvd, MS 5668
Roseville CA
Phone: +1-916-785-5198
email: jim_wendt@hp.com

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

Renato Recio
IBM Corporation
11501 Burnett Road
Austin, TX 78758 USA
Phone: +1 (512) 838-1365
Email: recio@us.ibm.com

Howard C. Herbert
Intel Corporation
MS CH7-404
5000 West Chandler Blvd.
Chandler, AZ 85226 USA
Phone: +1 (480) 554-3116
Email: howard.c.herbert@intel.com

Dave Minturn
Intel Corporation
MS JF1-210
5200 North East Elam Young Parkway
Hillsboro, OR 97124 USA
Phone: +1 (503) 712-4106
Email: dave.b.minturn@intel.com

James Pinkerton
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
Phone: +1 (425) 705-5442
Email: jpink@microsoft.com

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

17 Appendix

17.1 Design considerations for a Datamover protocol

This section discusses the specific considerations for RDMA-based and RDDP-based Datamover protocols, and is only informational.

- a) Note that the modeling of interactions for SCSI Data-Out (section 11.3.5.1) is only used for unsolicited data transfer.
- b) The modeling of interactions for SNACK (section 11.3.14, and section 11.4.14) is not expected to be used given that one of the design requirements on the Datamover is that it "guarantees an error-free, reliable, in-order transport mechanism" (section 8). The interactions for sending and receiving a SNACK are nevertheless modeled in this document because the receiving iSCSI layer can deterministically deal with an inadvertent SNACK. This also shows the DA designers' intent that DI is not meant to filter certain types of PDUs.
- c) The onus is on a reliable Datamover (per requirements stated in section 8) to realize end-to-end data acknowledgements via Datamover-specific means. In view of this, even data-ACK-type SNACKs are unnecessary to be used. Consequently, an initiator may never request sending a SNACK Request in this model assuming that the proactive (timeout-driven) SNACK functionality is turned off in the legacy iSCSI code.
- d) Note that the current DA model for bootstrapping a Connection Handle into service - i.e. associating a new iSCSI connection with a Connection Handle - clearly implies that the iSCSI connection must already be in full feature phase when the Datamover layer comes into the stack. This further implies that the iSCSI login phase must be carried out in the traditional "Byte streaming mode" with no assistance or involvement from the Datamover layer.

17.2 Examples of Datamover interactions

The figures described in this section provide some examples of the usage of Operational Primitives in interactions between the iSCSI layer and the Datamover layer. The following abbreviations are used in this section.

Avail - Available

Abted - Aborted

Buf - I/O Buffer

Cmd - Command

Compl - Complete

Conn - Connection

Ctrl_Ntfy - Control_Notify

Dal_Tk_Res - Deallocate_Task_Resources

Data_Cmp_Nfy - Data_Completion_Notify

Data_ACK_Nfy - Data_ACK_Notify

DM - Datamover

Imm - Immediate

Snd_Ctrl - Send_Control

Msg - Message

Resp - Response

Sol - Solicited

TMF Req - Task Management Function Request

TMF Res - Task Management Function Response

Trans - Transfer

Unsol - Unsolicited

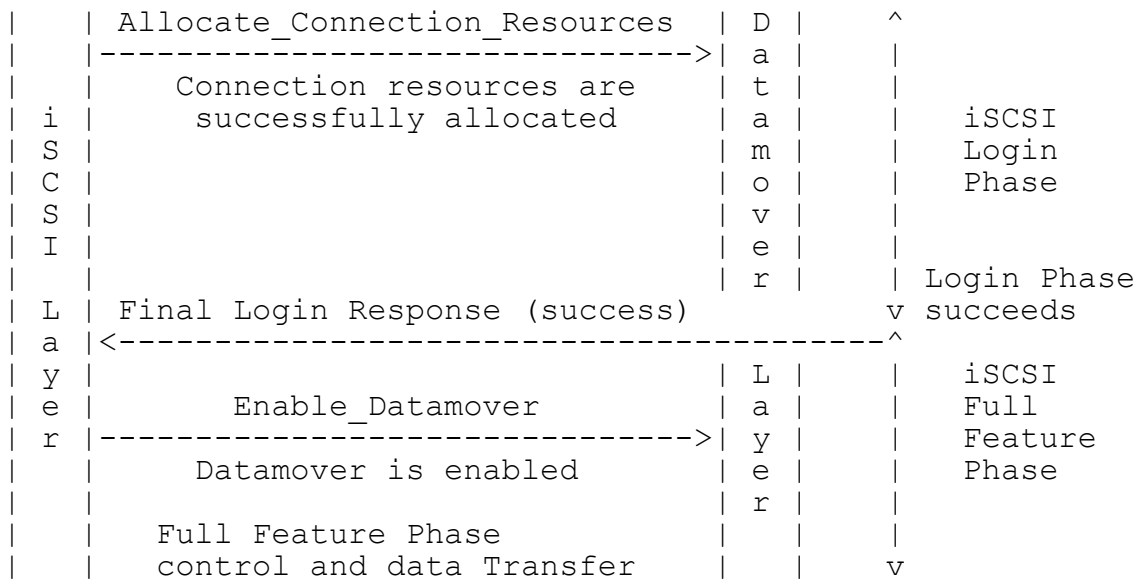


Figure 2 A successful iSCSI login on initiator

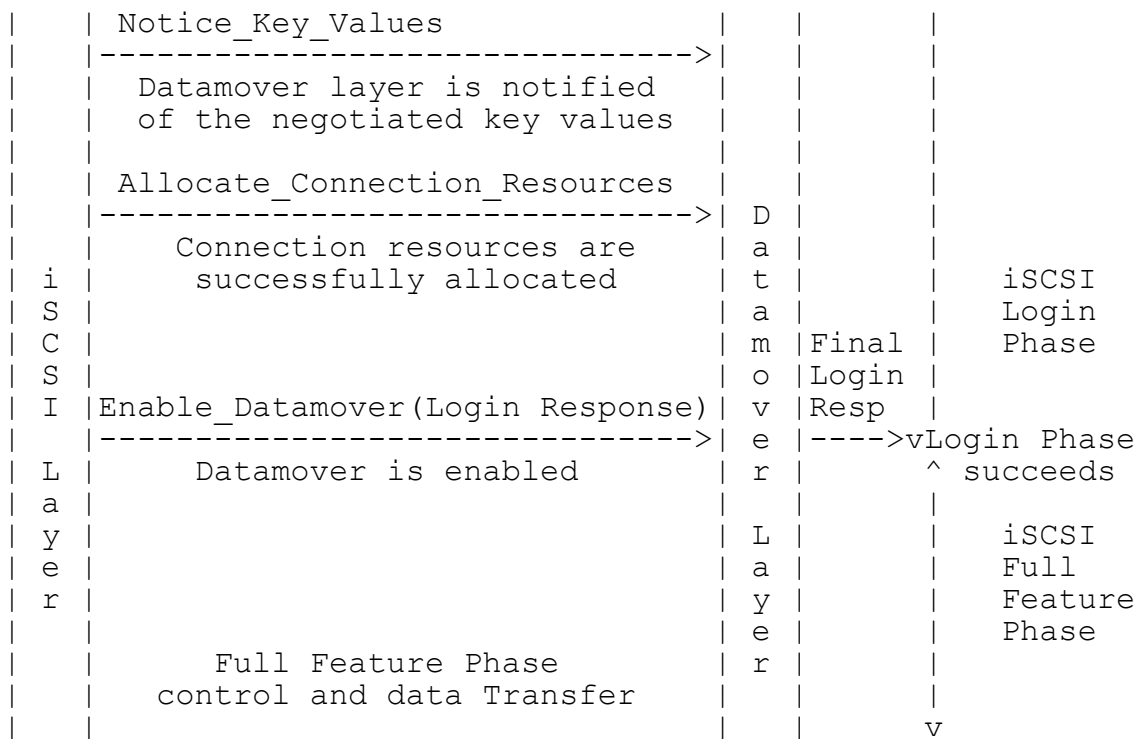


Figure 3 A successful iSCSI login on target

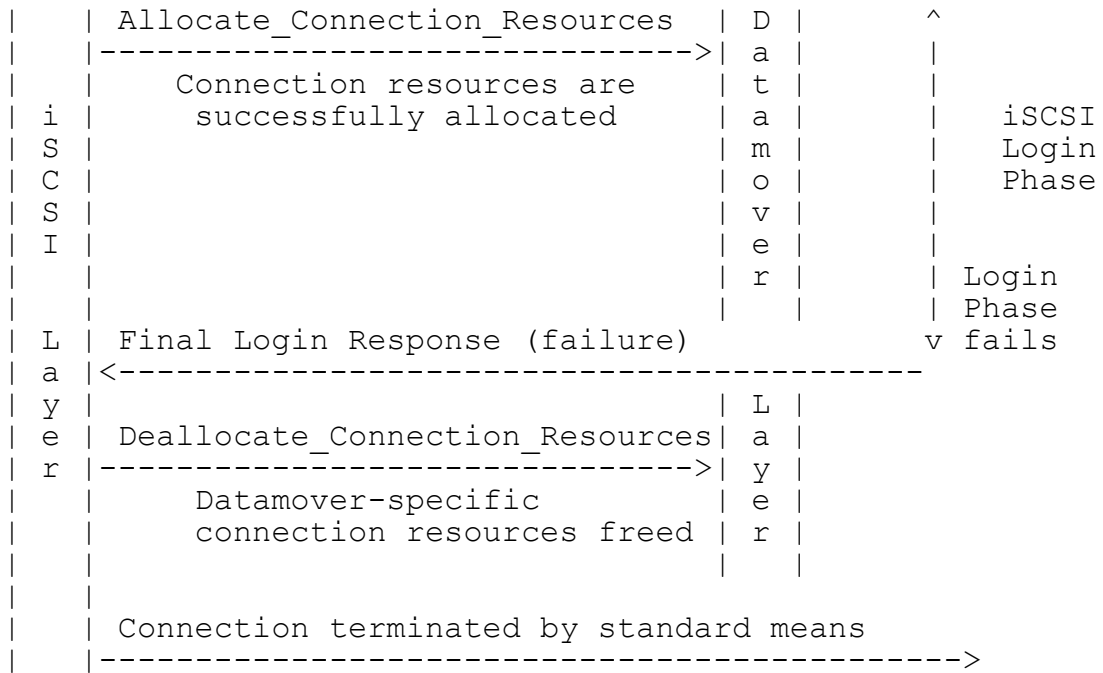


Figure 4 A failed iSCSI login on initiator

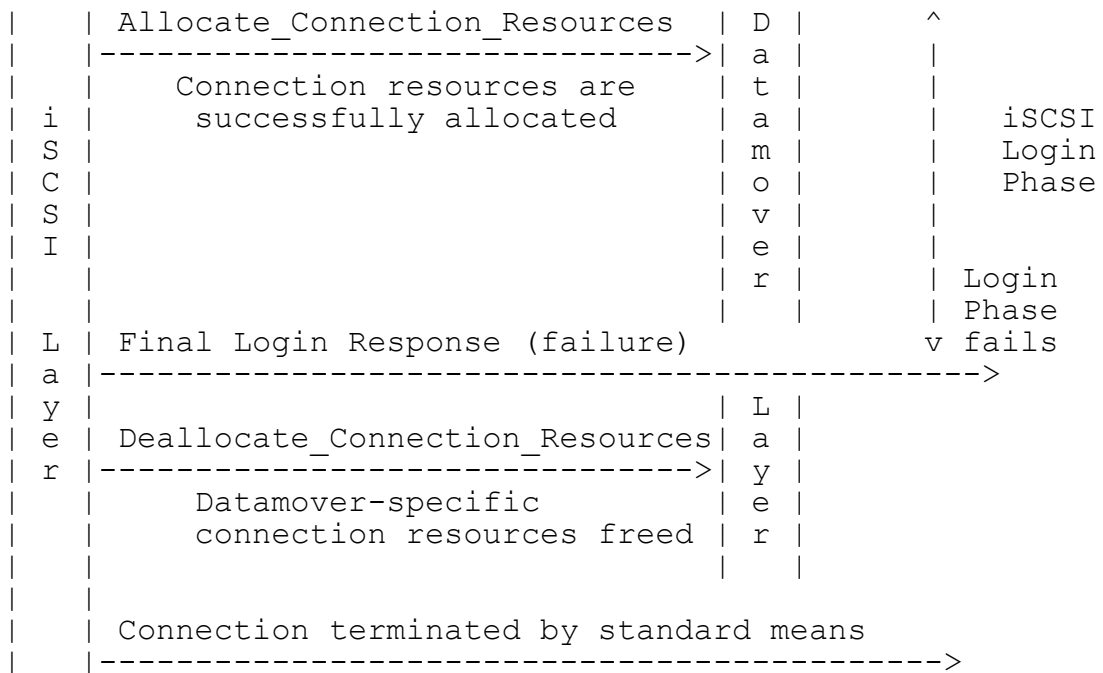


Figure 5 A failed iSCSI login on target

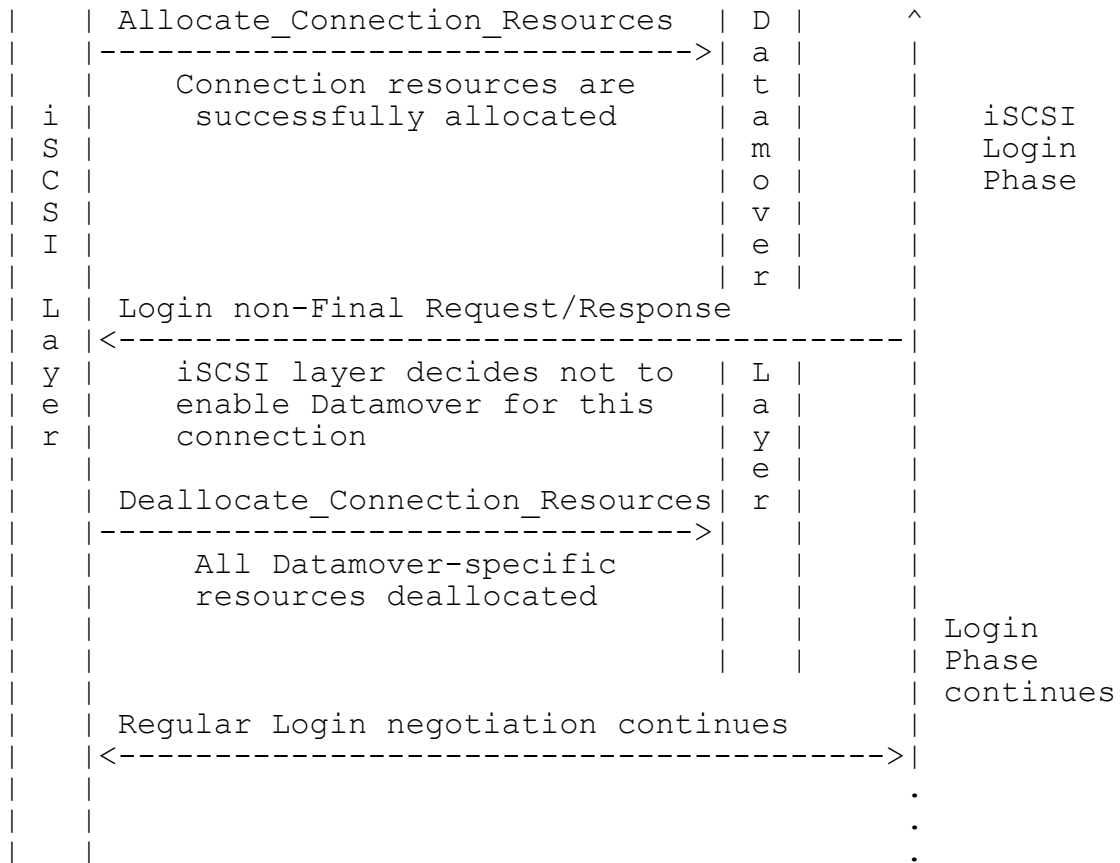


Figure 6 iSCSI does not enable the Datamover

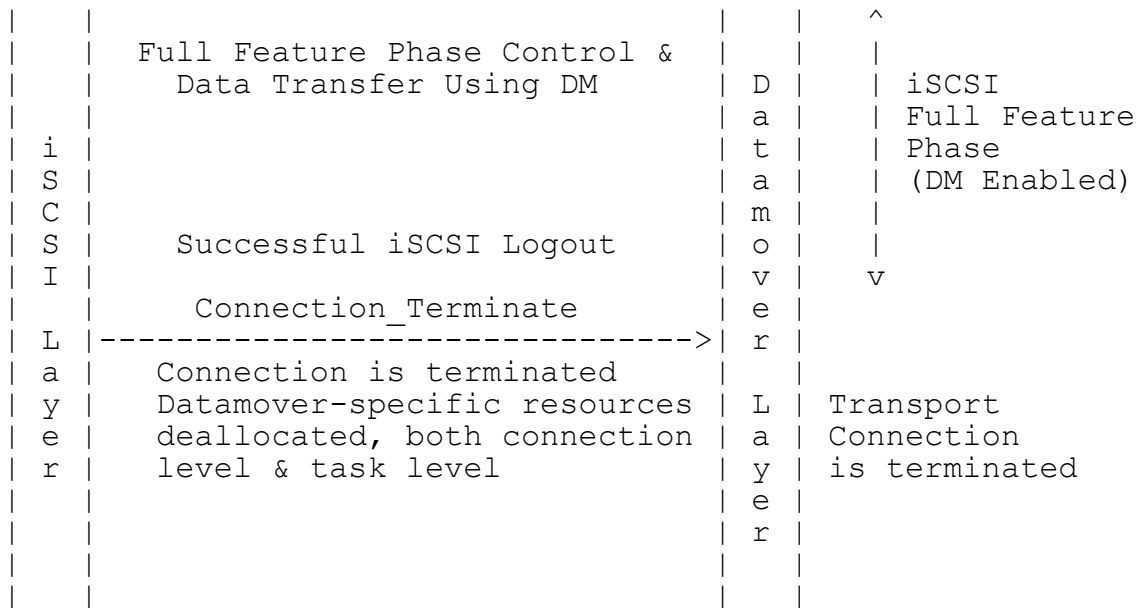


Figure 7 A normal iSCSI connection termination

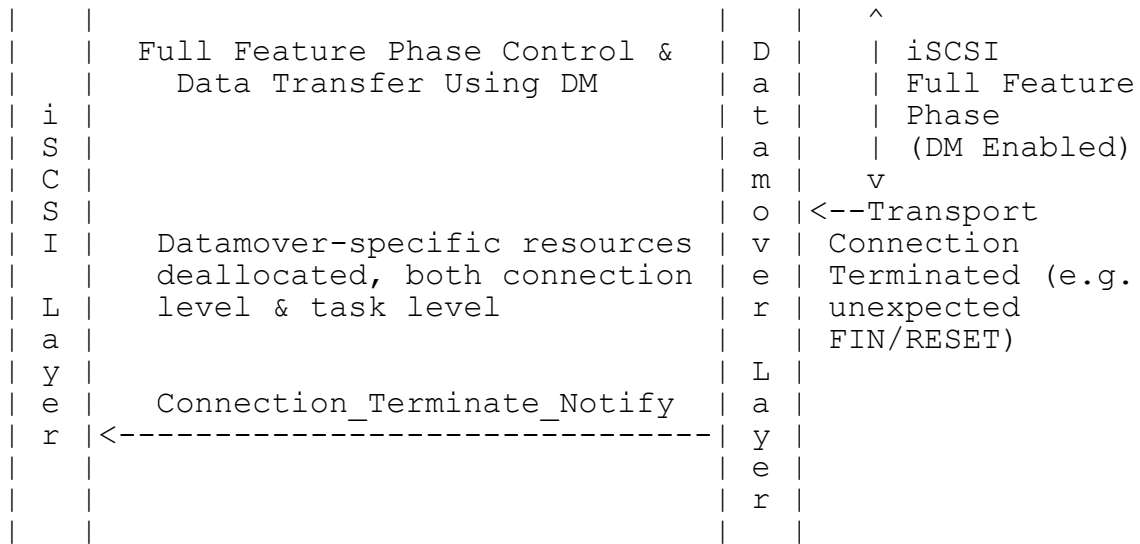


Figure 8 An abnormal iSCSI connection termination

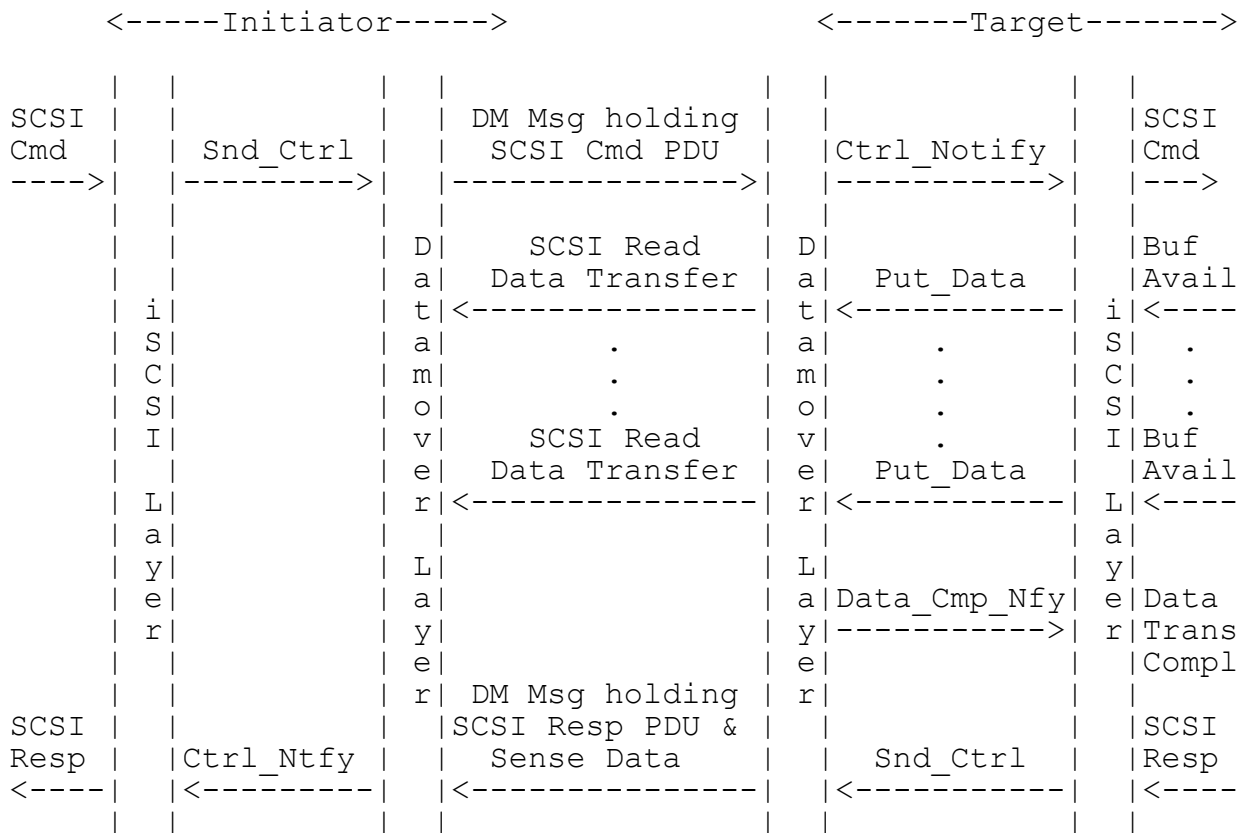


Figure 10 A SCSI Read data transfer

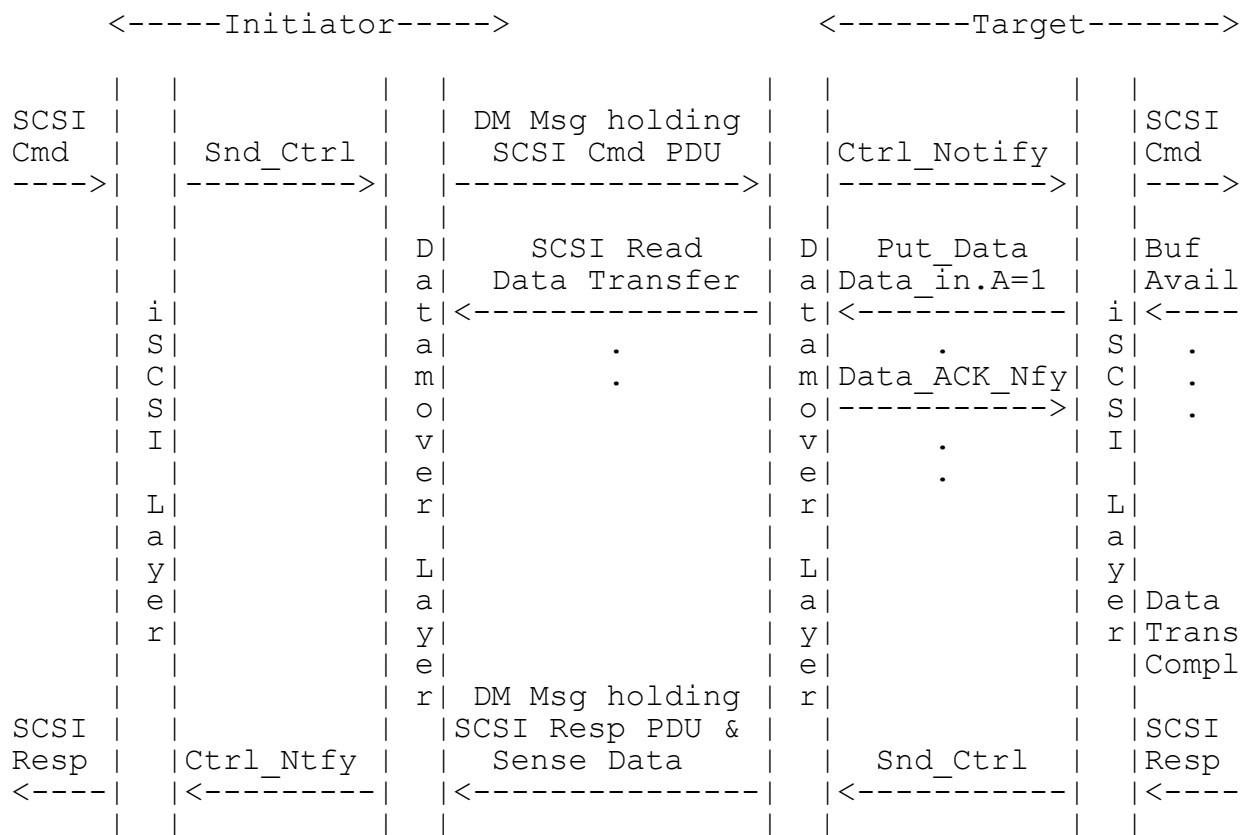


Figure 11 A SCSI Read data acknowledgement

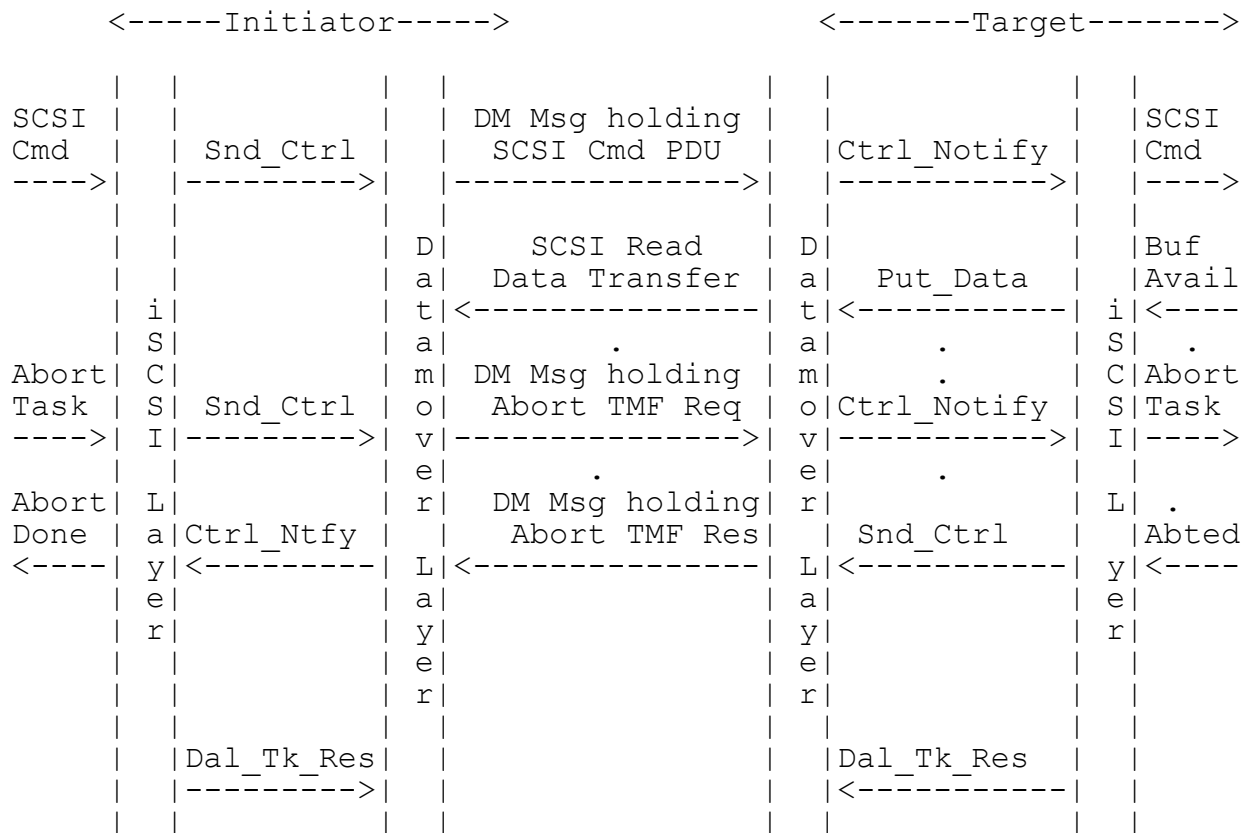


Figure 12 Task resource cleanup on abort

18 Full Copyright Statement

Copyright (C) The Internet Society (2004). This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

19 Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.