INTERNET DRAFT
draft-ietf-ips-iscsi-impl-guide-00.txt

Mallikarjun Chadalapaka
Hewlett-Packard Co.
Editor

Expires January 2006

iSCSI Implementer's Guide

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

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/1id-abstracts.html

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

Abstract

iSCSI is a SCSI transport protocol and maps the SCSI family of application protocols onto TCP/IP. RFC 3720 defines the iSCSI protocol. This document compiles the clarifications to the original protocol definition in RFC 3720 to serve as a companion document for the iSCSI implementers. This document updates RFC 3720 and the text in this document supersedes the text in RFC 3720 when the two differ.

Table of Contents

1 Definitions and acronyms 3
1.1 Definitions 3
1.2 Acronyms 3
2 Introduction 5
3 iSCSI semantics for SCSI tasks 6
3.1 SCSI REPORT LUNS and Residual Overflow 6
4 Task Management 8
4.1 Requests Affecting Multiple Tasks 8
4.1.1 Scope of affected tasks....8
4.1.2 Updated semantics....8
4.1.3 Rationale behind the new semantics....9
5 iSCSI Error Handling and Recovery 11
5.1 ITT 11
5.2 Format Errors 11
5.3 Digest Errors 11
6 Security Considerations 13
7 IANA Considerations 14
8 References and Bibliography 15
8.1 Normative References 15
8.2 Informative References 15
9 Editor's Address 16
10 Acknowledgements 17
11 Full Copyright Statement 18
12 Intellectual Property Statement 19

1 Definitions and acronyms

1.1 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.

1.2 Acronyms

Acronym	Definition
EDTL	Expected Data Transfer Length
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
LU	Logical Unit
LUN	Logical Unit Number
PDU	Protocol Data Unit
RDMA	Remote Direct Memory Access
R2T	Ready To Transfer
R2TSN	Ready To Transfer Sequence Number
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
TMF	Task Management Function
TTT	Target Transfer Tag

2 Introduction

Several iSCSI implementations had been built after [RFC3720] was published and the iSCSI community is now richer by the resulting implementation expertise. The goal of this document is to leverage this expertise both to offer clarifications to the [RFC3720] semantics and to address defects in [RFC3720] as appropriate. This document intends to offer critical guidance to implementers with regard to non-obvious iSCSI implementation aspects so as to improve interoperability and accelerate iSCSI adoption. This document, however, does not purport to be an all-encompassing iSCSI how-to guide for implementers, nor a complete revision of [RFC3720]. This document instead is intended as a companion document to [RFC3720] for the iSCSI implementers.

iSCSI implementers are required to reference [RFC3722] and [RFC3723] in addition to [RFC3720] for mandatory requirements. In addition, [RFC3721] also contains useful information for iSCSI implementers. The text in this document, however, updates and supersedes the text in all the noted RFCs whenever there is such a question.

3 iSCSI semantics for SCSI tasks

3.1 SCSI REPORT LUNS and Residual Overflow

The specification of the SCSI REPORT LUNS command requires that SCSI target limit the amount of data transferred to a maximum size (ALLOCATION LENGTH) provided by the initiator in the REPORT LUNS CDB. If the Expected Data Transfer Length (EDTL) in the iSCSI header of the SCSI Command PDU for a REPORT LUNS command is set to at least as large as that ALLOCATION LENGTH, the SCSI layer truncation prevents an iSCSI Residual Overflow from occurring. A SCSI initiator can detect that such truncation has occurred via other information at the SCSI layer. The rest of the section elaborates this required behavior.

iSCSI uses the (O) bit (bit 5) in the Flags field of the SCSI Response and SCSI Data-Out PDUs to indicate that that an iSCSI target was unable to transfer all of the SCSI data for a command to the initiator because the amount of data to be transferred exceeded the EDTL in the corresponding SCSI Command PDU (see Section 10.4.1 of [RFC 3720]).

The SCSI REPORT LUNS command requests a target SCSI layer to return a logical unit inventory (LUN list) to the initiator SCSI layer (see section 6.21 of SPC-3 [SPC3]). The size of this LUN list may not be known to the initiator SCSI layer when it issues the REPORT LUNS command; to avoid transfer of more LUN list data than the initiator is prepared for, the REPORT LUNS CDB contains an ALLOCATION LENGTH field to specify the maximum amount of data to be transferred to the initiator for this command. If the initiator SCSI layer has under-estimated the number of logical units at the target, it is possible that the complete logical unit inventory does not fit in the specified ALLOCATION LENGTH. In this situation, section 4.3.3.6 in [SPC3] requires that the target SCSI layer "shall terminate transfers to the Data-In Buffer" when the number of bytes specified by the ALLOCATION LENGTH field have been transferred.

Therefore, in response to a REPORT LUNS command, the SCSI layer at the target presents at most ALLOCATION LENGTH bytes of data (logical unit inventory) to iSCSI for transfer to the initiator. For a REPORT LUNS command, if the iSCSI EDTL is at least as large as the ALLOCATION LENGTH, the SCSI truncation ensures that the EDTL will accommodate all of the data to be transferred. If

all of the logical unit inventory data presented to the iSCSI layer ¡ i.e. the data remaining after any SCSI truncation - is transferred to the initiator by the iSCSI layer, an iSCSI Residual Overflow has not occurred and the iSCSI (O) bit MUST NOT be set in the SCSI Response or final SCSI Data-Out PDU. This is not a new requirement but is already required by the combination of [RFC 3720] with the specification of the REPORT LUNS command in [SPC3].

The LUN LIST LENGTH field in the logical unit inventory (first field in the inventory) is not affected by truncation of the inventory to fit in ALLOCATION LENGTH; this enables a SCSI initiator to determine that the received inventory is incomplete by noticing that the LUN LIST LENGTH in the inventory is larger than the ALLOCATION LENGTH that was sent in the REPORT LUNS CDB. A common initiator behavior in this situation is to re-issue the REPORT LUNS command with a larger ALLOCATION LENGTH.

4 Task Management

4.1 Requests Affecting Multiple Tasks

This section updates the original text in section 10.6.2 of [RFC3720]. The clarified semantics are a superset of the semantics of the original text in it the new text covers all TMFs that can impact multiple tasks.

4.1.1 Scope of affected tasks

ABORT TASK SET: All outstanding tasks for the I_T_L nexus identified by the LUN field in the ABORT TASK SET TMF Request PDU.
CLEAR TASK SET: All outstanding tasks in the task set for the LU identified by the LUN field in the CLEAR TASK SET TMF Request PDU. See [SPC3] for the definition of a "task set".

LOGICAL UNIT RESET: All outstanding tasks from all initiators for the LU identified by the LUN field in the LOGICAL UNIT RESET Request PDU.

TARGET WARM RESET/TARGET COLD RESET: All outstanding tasks from all initiators across all LUs that the TMF-issuing session has access to on the SCSI target device hosting the iSCSI session.

Usage example: an "ABORT TASK SET TMF Request PDU" in the preceding text is an iSCSI TMF Request PDU with the "Function" field set to "ABORT TASK SET" as defined in [RFC3720]. Similar usage is employed for other descriptions.

4.1.2 Updated semantics

The execution of ABORT TASK SET, CLEAR TASK SET, LOGICAL UNIT RESET, TARGET WARM RESET, and TARGET COLD RESET TMF Requests consists of the following sequence of actions in the specified order on each of the entities.

The initiator:

a) Issues ABORT TASK SET/CLEAR TASK SET/LOGICAL UNIT RESET/TARGET WARM RESET/TARGET COLD RESET request.

b) Continues to respond to each TTT received for the affected tasks.

c) Receives any responses that the target may provide for some tasks among the affected tasks (may process them as usual because they are guaranteed to be valid).

d) Receives the task management response concluding all the tasks in the set of affected tasks.

The Target:

a) Receives the ABORT TASK SET/CLEAR TASK SET/LOGICAL UNIT RESET/TARGET WARM RESET/TARGET COLD RESET request.

b) Waits for all currently valid target transfer tags of the affected tasks to be responded.

c) Based on the CmdSN ordering, waits (concurrent with the wait in step (b)) for all commands of the affected tasks to be received. In the case of target-scoped requests (i.e. TARGET WARM RESET and TARGET COLD RESET), all the commands that are not received, as at the end of step (b), in the command stream however can be considered to have been received with no command waiting period - i.e. the entire CmdSN space upto the CmdSN of the task management function can be "plugged" (refer section 6.9 on how aborting a specific task can implicitly plug the CmdSN of the task being aborted) at the end of step (b).

d) Propagates the TMF request to and receives the response from the target SCSI layer.

e) Takes note of last-sent StatSN on each of the connections in the iSCSI session(s) (one or more) sharing the affected tasks, and waits for acknowledgement of each StatSN (may solicit for acknowledgement by way of a NOP-In). If some tasks originate from non-iSCSI I_T_L nexuses then the means by which the target insures that all affected tasks have returned their status to the initiator are defined by the specific non-iSCSI transport protocol(s).

f) Sends the task set management response to the issuing initiator.

4.1.3 Rationale behind the new semantics

There are fundamentally three basic objectives behind the semantics specified in section 4.1.2.

1. Maintaining an ordered command flow I_T nexus abstraction to the target SCSI layer even with multi-connection sessions.

   - Target iSCSI processing of a TMF request must maintain the single flow illusion - steps c & d of the target behavior correspond to this objective.

2. Maintaining a single ordered response flow I_T nexus abstraction to the initiator SCSI layer even with multi-connection sessions.

   - Target must ensure that the initiator does not see "old" task responses (that were placed on the wire chronologically earlier than the TMF response) after seeing the TMF response - step e of the target behavior corresponds to this objective.

3. Draining all active TTTs corresponding to affected tasks before the TMF is acted on.

   - Targets are better off if the TTTs are deterministically retired before the affected tasks are terminated because that eliminates the possibility of large-sized Data-out PDUs with stale TTTs arriving after the tasks are terminated. Step b of the target behavior corresponds to this objective.

The only other notable thing in step c of the target behavior is the "plugging" part - it is an optimization that says if all tasks on the I_T nexus will be aborted anyway (as with a target reset), there is no need to wait, the target can simply plug all missing CmdSN slots and move on with TMF processing. The first objective (maintaining a single ordered command flow) is still met with this optimization because target SCSI layer only sees ordered commands.

5 iSCSI Error Handling and Recovery

5.1 ITT

Section 10.19 in [RFC3720] mentions this in passing but noted here again for making it obvious. An ITT value of 0xffffffff is reserved and MUST NOT be used by the initiator. The only instance it may be seen on the wire is in a target-initiated NOP-In PDU (and the initiator response to that PDU if necessary).

5.2 Format Errors

Section 6.6 of [RFC3720] discusses format error handling. This section elaborates on the "inconsistent" PDU field contents noted in [RFC3720].

All initiator-detected PDU construction errors MUST be considered as format errors. Some examples of such errors are:

- NOP-In with a valid TTT but an invalid LUN

- NOP-In with a valid ITT (i.e. a NOP-In response) and also a valid TTT

- SCSI Response PDU with Status=CHECK CONDITION, but DataSegmentLength = 0

5.3 Digest Errors

Section 6.7 of [RFC3720] discusses digest error handling. It states that "No further action is necessary for initiators if the discarded PDU is an unsolicited PDU (e.g., Async, Reject)" on detecting a payload digest error. This is incorrect.

An Asynchronous Message PDU or a Reject PDU carries the next StatSN value on an iSCSI connection, advancing the StatSN. When an initiator discards one of these PDUs due to a payload digest error, the entire PDU including the header MUST be discarded. Consequently, the initiator MUST treat the exception like a loss of any other solicited response PDU ¡ i.e. it MUST use one of the following options noted in [RFC3720]:

a) Request PDU retransmission with a status SNACK.

b) Logout the connection for recovery and continue the tasks on a different connection instance.

c) Logout to close the connection (abort all the commands associated with the connection).

6 Security Considerations

This document does not introduce any new security considerations other than those already noted in [RFC3720]. Consequently, all the iSCSI-related security text in [RFC3723] is also directly applicable to this document.

7 IANA Considerations

This draft does not have any specific IANA considerations other than those already noted in [RFC3720].

8 References and Bibliography

8.1 Normative References

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

[RFC3722] Bakke, M., "String Profile for Internet Small Computer Systems Interface (iSCSI) Names", RFC 3722, April 2004.

[RFC3723] Aboba, B., Tseng, J., Walker, J., Rangan, V., and F. Travostino, "Securing Block Storage Protocols over IP", RFC 3723, April 2004.

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

8.2 Informative References

[RFC3721] Bakke, M., Hafner, J., Hufferd, J., Voruganti, K., and M. Krueger, "Internet Small Computer Systems Interface (iSCSI) Naming and Discovery", RFC 3721, April 2004.

[iSER] Ko, M., Chadalapaka, M., Elzur, U., Shah, H., Thaler, P., J. Hufferd, "iSCSI Extensions for RDMA", IETF Internet Draft draft-ietf-ips-iser-04.txt (work in progress), June 2005.

[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).

9 Editor's Address

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

10 Acknowledgements

The IP Storage (ips) Working Group in the Transport Area of IETF has been responsible for defining the iSCSI protocol (apart from a host of other relevant IP Storage protocols). The editor acknowledges the contributions of the entire working group.

The following individuals directly contributed to identifying [RFC3720] issues and/or suggesting resolutions to the issues clarified in this document: David Black (REPORT LUNS/overflow semantics), Gwendal Grignou (TMF scope), Mike Ko (digest error handling for Asynchronous Message), Dmitry Fomichev (reserved ITT). This document benefited from all these contributions.

11 Full Copyright Statement

Copyright (C) The Internet Society (2005). 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.

12 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.