Release Notes

SCSI RDMA Protocol (SRP) Initiator

Rev 1.7.0

SCSI RDMA Protocol (SRP) Initiator Release Notes

Document Number:

Mellanox Technologies, Inc.
2900 Stender Way
Santa Clara, CA 95054
U.S.A.
www.Mellanox.com

Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies Ltd
PO Box 586 Hermon Building
Yokneam 20692
Israel

Tel: +972-4-909-7200
Fax: +972-4-959-3245

1 Overview

This document describes the contents of the SCSI RDMA Protocol (SRP) Initiator Rev 1.7.0 release. The SRP standard describes the message format and protocol definitions required for transferring commands and data between a SCSI initiator port and a SCSI target port using RDMA communication service. Currently, SRP Initiator supports SRP-1 only.

> Note: For supported platforms and operating systems, please refer to the *Mellanox IB Gold Distribution Release Notes*.

This document includes the following sections:

2 Software Dependencies

SRP Initiator depends on the installation of the IB Gold Distribution stack with OpenSM running.

3 Major Features

This SRP Initiator is based on the *SCSI RDMA Protocol, Doc. no. T10/1415-D.*

(www.t10.org/ftp/t10/drafts/srp2/srp2r00a.pdf)

It supports:

- Basic *SCSI Primary Commands -3 (SPC-3)* (www.t10.org/ftp/t10/drafts/spc3/spc3r21b.pdf)
- Basic *SCSI Block Commands -2 (SBC-2)* (www.t10.org/ftp/t10/drafts/sbc2/sbc2r16.pdf)

3.1 New Features

1. **Persistent target binding** - Since Linux assigns SCSI device nodes dynamically as each additional SCSI logical unit is detected, the mapping from device nodes (e.g., /dev/sda or /dev/sdb) to SRP SCSI targets and logical units may vary. Variations in process scheduling and network delay may result in SRP SCSI targets being mapped to different SCSI device nodes each time the driver is started. To provide a more reliable name-space, the SRP SCSI driver uses its own package to create persistent device naming for SRP SCSI devices.

   This method of device naming results in persistent device mapping. To avoid errors in mapping device names, the device names it creates should be used by applications and fstab files, and not those of direct referencing of particular SCSI device nodes. The Persistent Target Binding feature is activated automatically upon loading the ib_srp module.
2. **Network detection for new Target devices** - This features allows detection of new Target device in the network. This feature also remove the dependency of the Initiator to be loaded only after the Target is already initialized.

After adding new devices in the network run *rescan-scsi-bus.sh* and the Initiator will detect any new/removed device in the network.

3. **Direct and indirect addressing** - A direct data buffer contains a single memory descriptor, which is a single memory segment within a memory-region's virtual address space.An indirect data buffer is comprised of one or more memory segments that may be discontiguous. The previous and current versions of the SRP Initiator support indirect addressing. Under some conditions, the current Initiator attempts direct addressing; however, if it fails, it defaults back to indirect addressing. Note that direct addressing yields better performance.
4. **Performance enhancement** - The following system configuration was used to compare the performance of this new release with the previous 1.6.0 release: A system with dual Xeon 2.8 GHz, 3 GB memory, a K 2.65smp kernel, an MHXL-CF128-T InfiniHost Adapter card, and an MHEL-CF128-T InfiniHost III Ex Adapter card. The following results were received:

   - For raw sequential read operations: Throughput went up from 450 MB/s to 650 MB/s on the InfiniHost card, and from 450MB/s to 760MB/s on the InfiniHost III Ex card.
   - For raw sequential write operations: Throughput went up from 390 MB/s to 569 MB/s on the InfiniHost card, and from 390MB/s to 660 MB/s on the InfiniHost III Ex card.

4 KNOWN ISSUES AND UNSUPPORTED FEATURES

The following table lists currently unsupported compliancy features.

Table 1 - Known Issues And Unsupported Compliancy Features

Issue/Limitation	Compliancy	Description
SRP CRED REQ	6.10	The Initiator may use this request to adjust an SRP Initiator REQUEST LIMIT value
LOSOLNT bit	6.2	Logout solicited notification bit: Indicates whether an SRP_T_LOGOUT request should use normal or solicited message reception.
UCSOLNT bit	6.7 and 6.8	Unsuccessful completion bit: Indicates whether an SRP RSP response reporting an unsuccessful completion of task management should use normal or solicited message notification.
SCSOLNT	6.7 and 6.8	Successful completion bit: Indicates whether an SRP RSP response reporting a successful completion of task management should use normal or solicited message notification.
No support for additional CDB	6.8	Contains the length (in units of four-byte words) of the Additional CDB field.
No support for Task Management	6.7	The Initiator does not issue any specific SRP_TASK_MGNT. Instead, it disconnects / destroys the rdma channel.
No support for multiple Service ID of the same or different IOC(s) of the same or different IOUnit(s)	Annex B	The Initiator only issues SRP_LOGIN_CMD to a specific Service ID. There is no way to distinguish/connect to multiple SRP targets (different IOUnits) now.
Initiator does not detect a newly added SRP target		In a live system, a new SRP target is not detected by the Initiator. Workaround: run - rescan-scsi-scan-bus.sh

5 FIXED BUGS

Table 2 - Bugs fixes summery

Index	Description	Details
1	Ctrl-C during initialization can cause a kernel oops	In a system where the SRP Initiator does not detect any target in the network and a Ctrl-C is performed by the user, the Initiator will hang during module removal.
2	Module removal can cause a race in closing the QP	During module removal, if the module exit call and a CM idle event detection occur simultaneously, then a race may occur as both try to destroy the QP.
3	Heavy 1M ioreqs running on a machine with kernel 2.4 cause a kernel oops	A race occurs during the 1M ioreqs causes the oops
4	For kernel 2.6, removal of the SRP Target before that of the Initiator caused a kernel oops	If ‘modprobe -r ib_srp_target’ is run before ‘modprobe -r ib_srp’ on a machine running with kernel 2.6, the Initiator hangs.
5	During scsi_scan a kernel oops occurred for an SLES9 kernel running on an x86_64 machine	Fixed a wrong 32/64-bit casting of sg_dma_address

6 MAIN VERIFICATION FLOWS

In order to verify the correctness of the SRP Initiator, the following tests and parameters were run.

Table 3 - Verification Tests

Test	Description	Flags
dd	Convert and copy a file	-bs=512 | 16k | 64k | 1M -seek= | skip=
bonnie++ http://www.textuality.com/bonnie/	bonnie is a benchmark that measures the performance of Unix file-systems operations. It performs a series of tests on a file of known size. If the size is not specified, bonnie assumes 100 Mb.	-s | b
fdisk	Show disk partitions	-l
hdparm	Get/Set hard-disk parameters	-a | f | g | r | T | t | Y
iotest	This package of benchmarks, utilities, and exercisers should run on any SCSI disk.	-a | r | w | l
dt http://www.bit-net.com/~rmiller/dt.html	The data test (dt) program is modeled after the dd syntax but dt can do a lot more than sequential copies. It is a comprehensive data test program for SCSI devices such as disks, tapes and CDROM/DVDs.	bs=1k | 2k | 4k | 8k | 64k | 250k max | limit | incr | pattern | iotype
iometer http://www.iometer.org/	iometer is an I/O subsystem measurement and characterization tool for single and clustered systems. (It was originally developed by Intel Corporation.)	bs | limit | incr | records | patterns | procs | lbdata
iogen http://beta.acnc.com/04_02.html	iogen is a package of programs designed to measure I/O performance.	
iozone http://www.iozone.org	iozone is a filesystem benchmark tool. The benchmark tests file I/O performance for the following operations: read, write, re-read, re-write, read backwards, fread, fwrite, random read/write, aio_read, aio_write	
Oracle Orion test suite	This test suite includes: small-random, large-random, sequential-matrix reads and writes	
RMAN backup and recovery	Bring up Oracle's 10g database and run RMAN backup validate, backup and recovery.	
pldd http://members.aol.com/plscsi/tools/pldd	A re-write of dd, the standard Unix tool for copying bytes of files and block devices. Pldd is a layer built on top of gccscsi.	-bs=512 | 16k | 64k | 1M -seek= | skip=
bad flow	Checks the following bad flow scenarios during run-time: • Removal of an SRP Target module • Disconnecting a cable/switch between the Initiator and the Target. • Killing running tests using Ctrl-C/Z • Changing IB links layer states (moving from Port Active to Port Down).	