OPENFABRICS ALLIANCE

SRP Update

Bart Van Assche, FUSION-IO®

Overview

- Involvement With SRP
- SRP Protocol Overview
- Recent SRP Driver Changes
- Possible Future Directions

Involvement with SRP

- Maintainer of the open source Linux SRP initiator and the SCST SRP target drivers.
- Member of the Fusion-io ION team. ION is an all-flash H.A. shared storage appliance.
- Flash memory provides low latency and high bandwidth.
- The focus of RDMA is on low latency and high bandwidth.
- In other words, RDMA is well suited for remote access to flash memory.

SRP Protocol Overview

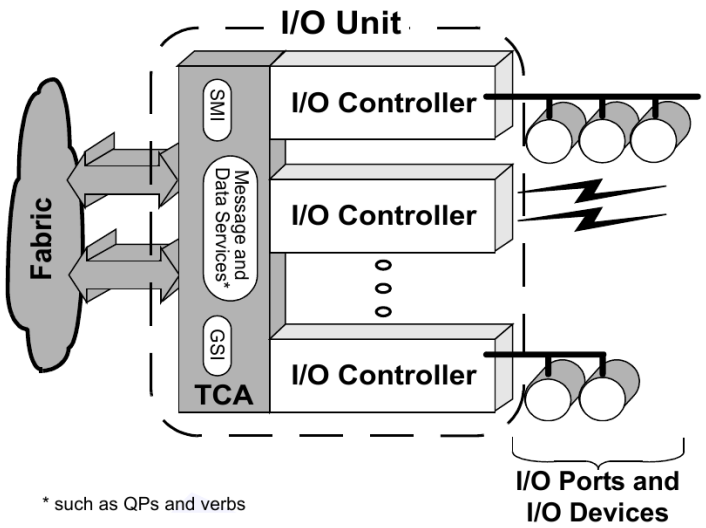
- SRP = SCSI RDMA Protocol.
- Defines how to perform SCSI communication over an RDMA network.
- Defines how to discover InfiniBand SRP targets, how to log in, how to transfer SCSI CDB's and also how to transfer data via RDMA.
- Revision 16a of the SRP protocol has been approved as an official ANSI standard in 2007.

SRP and SCSI

- SRP defines a SCSI transport layer.
- Enables supports for e.g. these SCSI features:
  - Reading and writing data blocks.
  - Read capacity.
  - Command queueing.
  - Multiple LUNs per SCSI host.
  - Inquire LUN information, e.g. volume identification, caching information and thin provisioning support (a.k.a. TRIM / UNMAP).
  - Atomic (vectored) write - helps to make database software faster.
  - VAAI (WRITE SAME, UNMAP, ATS, XCOPY).
  - End-to-end data integrity (a.k.a. T10-PI).
  - Persistent reservations a.k.a. cluster support.
  - Asymmetric Logical Unit Access (ALUA).
- Fusion-io is actively involved in the ANSI T10 committee for standardization of new SCSI commands.

SRP Protocol - Login

- IB spec defines *device management*.
- Initiator sends device management query to subnet manager.
- Subnet manager reports ports with device management capabilities.
- Initiator sends I/O controller query to each port with device management capabilities.
- SRP target reports I/O controllers.
- Initiator sends login request to selected I/O controllers.
- Initiator requests SCSI LUN report and queries capacity and identification of each LUN.
- I/O starts.



Model for an I/O Unit

Linux SRP Initiator Support

- Kernel driver ib_srp - implements SRP protocol.
- User space srptools package.
- srp_daemon and ibsrpdm executables.
  - Target discovery.
  - Target login.
- Interface between kernel and user space
  - /sys/class/infiniband_srp/srp-${port}/add_target
  - /sys/class/srp_remote_ports
  - /sys/class/scsi_host/*/{sgid,dgid,...}
  - /sys/class/scsi_device/*/{state,queue_depth,...}

SRP Login - Example

```
# cat /etc/srp_daemon.conf
a queue_size=128,max_cmd_per_lun=128
# srp_daemon -oaecd/dev/infiniband/umad1
id_ext=0002c90300fc3210,ioc_guid=0002c90300fc3210,dg
id=fe8000000000000002c90300fc3211,pkey=ffff,service
_id=0002c90300fc3210
id_ext=0002c90300a543b0,ioc_guid=0002c90300a543b0,d
gid=fe8000000000000002c90300fc3221,pkey=ffff,servic
e_id=0002c90300a543b0
[ … ]
# ls /sys/class/srp_remote_ports/
port-453:1  port-459:1  [ … ]
# lsscsi
[5:0:0:0]    disk    FUSIONIO ION LUN         3243  /dev/sdc
[5:0:0:1]    disk    FUSIONIO ION LUN         3243  /dev/sdd
```

Recent SRP Initiator Changes

- Queue size is now configurable. Optimal performance for SSDs and hard disk RAID arrays can only be achieved with a large queue size (128 instead of the default 64).
- Support for modifying the queue depth dynamically has been added.
- Path loss detection time has been reduced from about 40s to about 17s. Further reduction is possible by lowering the subnet timeout on the subnet manager. This makes a significant difference in H.A. setups.
- Added support for fast_io_fail_tmo and dev_loss_tmo parameters for multipath.
- P_Key support has been added in srp_daemon.
- Many smaller changes in the srptools package.

OFED and SRP Support

	ib_srp	srptools
Upstream Linux kernel	3.14.0	1.0.2
RHEL 6.5	2.6.32+	0.0.4
SLES 11 SP3	3.0.101	0.0.4
MLNX OFED 2.1	3.13.0	1.0.0
OFED 3.12	3.12.0	1.0.2

Fusion-io is working with Linux distribution vendors to keep SRP support up to date.

SRP Initiator and SCSI Core

- Linux SRP initiator is a SCSI driver.
- Linux SCSI mid-layer builds on block layer.
- SRP initiator relies on SCSI core for LUN scanning, SCSI error handling, ...
- Path removal triggers a call of scsi_remove_host().
- Path removal during I/O works reliably since Linux kernel 3.8.
- Fusion-io contributed several patches to make the Linux SCSI core and block layer handle path removal during I/O reliably.

Possible Future Directions

- Improving Linux SCSI performance via the scsi-mq project.
- Higher bandwidth by using multiple RDMA channels.
- Latency reduction.
- NUMA performance improvements.
- FRWR support - needed e.g. for ConnectIB HCA support.
- End-to-end data integrity (T10-PI) support; supported by Oracle database software. Builds on FRWR support.
- Adding SR-IOV support.
- Support for Ethernet networks (RoCE and/or iWARP).
  - Requires to switch from IB/CM to RDMA/CM.
  - Requires modification of the target discovery software (srptools). The current target discovery software is based on InfiniBand MADs.

OPENFABRICS
ALLIANCE