Overview

- SAN properties
- Infiniband SAN technologies
- SRP overview
- Lustre SAN topologies
- Tuning
- Fabric gripes and wishes
SAN Properties

• Flexibilities
  – Storage can be location independent
  – Provisioning for multiple uses
  – Independent hardware lifecycle from compute hardware

• Scalability
  – Deploy in scalable units
  – Aggregation of storage

• Common Fabrics
  – FC (FCP)
  – IB (SRP)
  – Ethernet (iSCSI)
Infiniband SAN Technologies

• Protocols using RDMA
  – iSCSI RMDA extensions (iSER)
  – SCSI RDMA Protocol (SRP)

• SCSI target implementations
  – SCST
  – LIO
  – Storage arrays (DDN/Netapp)
SRP Overview

• SRP protocol provides a method for sending SCSI commands from initiator to target over Infiniband with RDMA
  – Initial connection with IB CM
• Initiator sees a SCSI block device to use for I/O
• High availability with DM Multipath
• srp_daemon can automatically connect to targets available to the HCA
• Default and only pkey is 0xffffffff
Mixed Fabric/Direct Attached Storage

Filesystem Network
Direct Attached
Fabric SAN
LUN Mapping

FS OSS/RP LUN Assignment

LUN Ownership by RP

Controller 0

LUN Ownership by OSS:

Controller 1

Controller 2

OLCF - Oak Ridge Leadership Computing Facility
Fabric Attached Storage Complications

- Lustre can saturate IB links to object storage, so every switch port must be line rate
  - Full bisection bandwidth fabric
  - Lots of infrastructure!
- Managing SRP target dgids
  - Zoning and identifying storage on fabric
- Removing targets
  - Cleanup is not complete and SRP fails to log back in to host
    - Can’t get rid of /sys/class/scsi_host/hostX entries
    - Improvements in recent versions of ib_srp
Tuning Parameters

- `/etc/modprobe.d/ib_srp.conf`
  - options ib_srp srp_sg_tablesize=255
- `/etc/srp_daemon.conf`
  - a max_sect=65535,max_cmd_per_lun=16
  - Per scsi_host: `/sys/class/scsi_host/hostX`
- `/sys/block/sdX/queue`
  - max_sectors_kb (match max_hw_sectors_kb)
  - nr_requests
  - read_ahead_kb
  - scheduler (don’t use cfq)
I/O Size

- IOs can be broken up at various points between application and disk
- SRP limited to 1M writes, larger reads
- SRP limited to 255 scatter gather entries per I/O
- IO coalescing on target is a good thing
- Verify all the way from application to disk
  - Lustre has a proc file brw_stats
  - Stats from storage array
  - On host from sar –d
    - Calculate sectors per transaction from number of sectors rd/wrt per second / tps
Fabric Wish List

• Monitoring health of fabric
• Identifying congestion on fabric
• Evaluate routing algorithms (DFSSSP)
• Partitions for separate SANs
• Converged fabric
  – PXE booting over IB
  – Management IP traffic (NFS, log collection)
Frequent Issues

• Opensm failures
  – Hosts get wrong P-Key
  – Failure to converge on a master SM
  – Rogue SMs

• Physical errors from bad cables