Highly-Available Lustre with SRP-Mirrored LUNs

UF HPC Center
Research Computing
University of Florida
HA Lustre

✈ Design Goals
✈ Minimize Cost per TB
✈ Maximize Availability
✈ Good Performance (within cost constraints)
✈ Avoid External SAS/Fibre Attached JBOD
✈ Avoid External RAID Controllers
✈ Support Ethernet and InfiniBand clients
✈ Standard Components
✈ Open Source Software
HA Lustre

♫ To Minimize Cost
♫ Commodity storage chassis
♫ Internal PCIe RAID controllers
♫ Inexpensive, high-capacity 7200 rpm drives

♫ Problem: How do we enable failover?

♫ Solution: InfiniBand + SRP
♫ SCSI RDMA Protocol
HA Lustre

Problem

- All storage is internal to each chassis
- No way for one server to take over the storage of the other server in the event of a server failure
- Without dual-ported storage and external RAID controllers how can one server take over the other’s storage?

Solution

- InfiniBand
- SCSI Remote/RDMA Protocol (SRP)
HA Lustre

- **InfiniBand**
  - Low-latency, high-bandwidth interconnect
  - Used natively for distributed memory applications (MPI)
  - Encapsulation layer for other protocols (IP, SCSI, FC, etc.)

- **SCSI Remote Protocol (SRP)**
  - Think of it as SCSI over IB
  - Provides a host with block-level access to storage devices in another host.
  - Via SRP host A can see host B’s drives and vice-versa
Host A can see host B’s storage and host B can see host A’s storage but there’s a catch…

If host A fails completely, host B still won’t be able to access host A’s storage since host A will be down and all the storage is internal.

So SRP/IB doesn’t solve the whole problem.

But… what if host B had a local copy of Host A’s storage and vice-versa (pictures coming – stay tuned).

Think of a RAID-1 mirror, where the mirrored volume is comprised of one local drive and one remote (via SRP) drive
HA Lustre

☞ Mirrored (RAID-1) Volumes

☞ Two (or more) drives

☞ Data is kept consistent across both/all drives

☞ Writes are duplicated to each disk

☞ Reads can take place from either/all disk(s)
Remote Mirrors

Not Possible?

Host A

Host A Disk 1  Host A Disk 2

Host B

Host B Disk 1  Host B Disk 2

QDR IB
Remote Mirrors

Remote targets exposed via SRP
Remote Mirrors

Mirroring Possibilities

Host A

Mirror A

Host A Disk 1

Host B Disk 1

Mirror B

Host A Disk 2

Host B Disk 2

Host B

Mirror A

Host A Disk 1

Host B Disk 1

Mirror B

Host A Disk 2

Host B Disk 2

SRP

QDR IB
Remote Mirrors

Normal Operating Conditions

Host A
Mirror A
- Host A Disk 1
- Host B Disk 1
- Host A Disk 2
- Host B Disk 2

Host B
Mirror B
- Host A Disk 1
- Host B Disk 1
- Host A Disk 2
- Host B Disk 2

SRP
QDR IB
Remote Mirrors

Host A is down
Remote Mirrors

★ Degraded mirrors on host B

[Diagram showing the layout of mirrored disks between Host A and Host B, with connections through SRP and QDR IB.]
HA Lustre

Hardware Configuration

- Chenbro RM91250 Chassis (50 Drives, 9U)
- SuperMicro X8DAH System Board
  - PCIe Slots: 2 x16, 4 x8, 1 x4
- Intel E5620 Processors (2)
- 24 GB RAM
- Adaptec 51245 PCI-E RAID Controller (4) (x8 slots)
- Mellanox MT26428 ConnectX QDR IB HCA (2) (x16 slot)
- Mellanox MT25204 InfiniHost III SDR IB HCA (1) (x4 slot)
HA Lustre

RAID Configuration

- Adaptic 51245 (4)
- RAID-6 (4+2) (to stay below 8 TB LUN)
- 7.6 TiB per LUN
- 2 LUNs per controller
- 8 LUNs per OSS
- 60.8 TiB per OSS
HA Lustre

LVM2 Configuration

Encapsulate each LUN in an LV

- Identification
- Convenience

LVs named by host, controller, LUN

- h<L>c<M>v<N>
  - h1c1v0, h1c1v1
  - h1c2v0, h1c2v1
  - h1c3v0, h1c3v1
  - h1c4v0, h1c4v1
HA Lustre

MD (Mirror) Configuration

Mirror consists of 1 local and 1 remote LUN

Host 1

/dev/<vg>/<lv>: /dev/h1c1v0/h1c1v0 (local)
/dev/h2c1v0/h2c1v0 (remote)

Device: /dev/md/ost0000

Host 2

/dev/<vg>/<lv>: /dev/h1c1v1/h1c1v1 (remote)
/dev/h2c1v1/h2c1v1 (local)

Device: /dev/md/ost0004
HA Lustre

Host 1

LVs
md100 = h1c1v0 + h2c1v0
md101 = h1c2v0 + h2c2v0
md102 = h1c3v0 + h2c3v0
md103 = h1c4v0 + h2c4v0

OSTs
ost0000 = md100
ost0001 = md101
ost0002 = md102
ost0003 = md103

Host 2

LVs
md104 = h1c1v1 + h2c1v1
md105 = h1c2v1 + h2c2v1
md106 = h1c3v1 + h2c3v1
md107 = h1c4v1 + h2c4v1

OSTs
ost0004 = md104
ost0005 = md105
ost0006 = md106
ost0007 = md107
SRP Mirrored Lustre HA OSS Pair

h1c1v0 (local)  h1c1v1 (remote)
h1c2v0 (local)  h1c2v1 (remote)
h1c3v0 (local)  h1c3v1 (remote)
h1c4v0 (local)  h1c4v1 (remote)
h2c1v0 (remote) h2c1v1 (local)
h2c2v0 (remote) h2c2v1 (local)
h2c3v0 (remote) h2c3v1 (local)
h2c4v0 (remote) h2c4v1 (local)

Each OSS:
Chenbro RM91250
SuperMicro X8DAH
2 x Intel E5620
24 GB RAM
2 x QDR IB
1 x SDR IB
4 x Adaptec 51245
RAID-6 (4+2)
8 x 8TB LUNs

OST0000 => MD100 => h1c1v0 + h2c1v0
OST0001 => MD101 => h1c2v0 + h2c2v0
OST0002 => MD102 => h1c3v0 + h2c3v0
OST0003 => MD103 => h1c4v0 + h2c4v0
OST0004 => MD104 => h1c1v1 + h2c1v1
OST0005 => MD105 => h1c2v1 + h2c2v1
OST0006 => MD106 => h1c3v1 + h2c3v1
OST0007 => MD107 => h1c4v1 + h2c4v1
HA Lustre

High-Availability Software (Open Source)

- Corosync
- Pacemaker

Corosync

- Membership
- Messaging

Pacemaker

- Resource monitoring and management framework
- Extensible via Resource agent templates
- Policy Engine
HA Lustre

Corosync Configuration

Dual Rings

- Back-to-Back ethernet
- IPoIB via SRP IB Interface

- clear_node_high_bit: yes
- rrp_mode: passive
- rrp_problem_count_threshold: 20
- retransmits_before_loss: 6
HA Lustre

Pacemaker Configuration

Resources

- Stonith (modified to control multiple smart pdus)
- MD (custom)
- Filesystem (stock)

Resource Groups (managed together)

- One per OST (grp_ostNNNN)
- MD + File system
- Not LVs – some disappear if a node goes down
HA Lustre

Performance

- 4 PCI-E RAID Controllers per Server
  - 2 RAID-6 (4+2) Logical Disk per Controller
  - 8 Logical Disks per Server (4 local, 4 remote)
  - 490 MB/sec per Logical Disk
  - 650 MB/sec per Controller (parity limited)

- Three IB Interfaces per Server
  - IB Clients (QDR, Dedicated)
  - IPoIB Clients (SDR, Dedicated)
  - SRP Mirror Traffic (QDR, Dedicated)
HA Lustre

Performance (continued)

Per Server Throughput

- 1.1 GB/sec per server (writes – as seen by clients)
- 1.7 GB/sec per server (reads – as seen by clients)

Actual server throughput is 2x for writing (mirrors!)

That’s 2.2 GB/s per Server

85% of the 2.6 GB/s for the raw storage
HA Lustre

Performance – Didn’t come easy

- Defaults for everything, no mirroring
  - Default PV alignment (??)
  - RAID stripe unit size (256 KB)
  - aacraid `max_hw_sectors_kb` (256 KB, controlled by `acbsize`)
  - MD device `max_sectors_kb` (128 KB)
  - Lustre max RPC size (1024 KB)

- Per-OST streaming throughput, no mirroring

  - Ugh!
    - Reads: ~253 MB/s
    - Writes: ~173 MB/s
HA Lustre

- Performance – Didn’t come easy
  - Alignment PVs to RAID stripe boundary
    - Streaming reads: ~333 MB/s
    - Streaming writes: ~280 MB/s

- Increase MD max I/O = RAID stripe size = aacraid max I/O
  - Required patch to MD RAID1 module (hardwired)
  - Only improved streaming reads: ~360 MB/s

- Increase max I/O size (MD + aacraid) => 512KB
  - aacraid acbsize=4096 (driver unstable beyond 4096)
  - Streaming writes: ~305MB/s
  - Could not reach a 1MB max I/O size
Performance – Didn’t come easy

Introduce SRP Mirrors…

Lustre RPC size = aacraid max I/O = SRP target RDMA size = MD max I/O = 512 KB

Per-OST streaming reads: ~433 MB/s

Improvement via MD read balancing

Per-OST streaming writes: ~280 MB/s

Slight penalty with SRP – can be CPU-bound on the core that handles the SRP HCA interrupts

Slightly faster OSS CPU would presumably help this
HA Lustre

Performance – Summary

- HA OSS (4 SRP-mirrored OSTs total)
- Streaming writes: 1.1 GB/s (i.e. 2.2 GB/s)
- 85% of sgpdd-survey result
- Reads: 3.4 GB/s (per pair)
  - 1.7 GB/s observed from each HA OSS
- Considerable improvement over defaults
HA Lustre

Keeping the data safe

- Mirrors enable failover
- Provide a second copy of the data
- Each Mirror
  - Hardware RAID
  - RAID-6 (4+2), two copies of parity data

Servers protected by UPS

- Orderly shutdown of servers in the event of a sudden power outage.
- 3+1 Redundant power supplies each to a different UPS.
HA Lustre

Problems Encountered

- Unstable SRP Target: OFED SRP target proved unstable
  - Used SCST SRP target (started w/ pre 2.0 release)

- MD Mirror Assembly
  - May choose wrong mirror under corosync.
  - Could not duplicate outside of corosync control
  - Requires deactivating the out-of-sync volume, assembling the degraded mirror, then adding the out-of-sync volume. Not ideal

- Poor Initial Performance
  - Resolved through tuning (described previously)
HA Lustre

Problems Encountered (continued)

- Zone Allocator killed us
- Blocked monitoring agents led to many needless remounts and sometimes STONITH events
- Could not pinpoint the problem which often but not always seemed correlated with load
- Seems we were the last to know about the long delays caused by the zone allocator
- Many timeout parameters unnecessarily adjusted to be very loooong.
- `vm.zone_reclaim_mode = 0`
- 100% stable now
HA Lustre

Future Improvements

- SSD cache (i.e. Adaptec maxCache)
- External journal device
- 6 Gbps RAID cards capable of > 512KB I/Os
- Faster processor (for SRP interrupt handling)
- 8+2 RAID-6 OSTs
  - More efficient disk utilization (4/5 vs 2/3)
  - Affects chassis and backplane choices
HA Lustre

Thank You

Questions or Comments?