Lustre User Group 2009

Sun Lustre Storage System
Cluster Best Practices

Joey Jablonski
Sun Microsystems, Professional Services
Agenda

• What defines the Sun Lustre Storage System?
• HA-MDS RAID Layout
• HA-OSS RAID Layout
• External Journals/Bitmaps
• Standup Processes
• Failover Testing Process
What defines the Sun Lustre Storage System?

- Known/Consistent HW Platform
- Known/Consistent RAID Layout
- Known performance characteristics
- Simplified deployment
- Streamlined Support
Key Modules & Architecture

HA MDS Module
- Shared Storage
  - MDS 1 (Active)
  - MDS 2 (Standby)

Interconnect
- Clients

Standard OSS Module
- OSS
- CPU
- Storage

HA OSS Module
- OSS 1 (Active)
- Shared Storage
- OSS 2 (Active)
- Shared Storage

= Failover
MDS = Meta Data Server
OSS = Object Storage Server
HA-MDS RAID Layout

- Allows easy separation of MGS and MDT functionality
- Allows easy additions of MDTs and new filesystems
- Allows use of both systems within MGS failover-pair
HA-OSS RAID Layout

- 2 OSTs per J4400 (RAID6 8+2)
- 1 External Journal + 1 External Bitmap per OST
  > Separate RAID 1 Devices
HA-OSS OST Association
Standup Process

• Stress testing all disks
  > `dd if=/dev/zero of=/dev/md10 bs=512k count 500k`
  > `dd of=/dev/null if=/dev/md10 bs=512k count 500k`

• Stress testing the SAS connections
  > Failed SAS cables show as continuous HD rebuilds

• Stress testing the nodes

• Rebuild Bandwidth
  > `/proc/sys/dev/raid/speed_limit_min`
  > `/proc/sys/dev/raid/speed_limit_max`
External Journals/Bitmaps

- Example Command Assumptions
  > 3 MD devices per OST
    > /dev/md10 – OST
    > /dev/md11 – External Journal
    > /dev/md12 – External Bitmap

- Bitmaps
  > Significantly decreases SW RAID recovery time
  > mdadm -C /dev/md10 –bitmap-/dev/md12/md10.bitmap
    /dev/dsk/c0t0d{0,1,2,3,4,5,6,7,8,9}

- Journals
  > mkfs.lustre --ost –mkfsoptions “-J /dev/md11” /dev/md10
Failover Testing Process

- Completed for each HA-pair
- Functional Testing
  > `hb_takeover`
- "Real" Testing
  > `echo 1 > /proc/sys/kernel/sysrq`
  > `echo c > /proc/sysrq-trigger`
HA-OSS Module Performance

IOZone Benchmark Results

- Sustained write performance observed to be ~2.1 GB/sec
- 16GB file size used to defeat client side caching
- Peak performance was reached at 96 threads; plateau effect observed at ~72 threads
- Varying block sizes 256K, 512K, 1MB did not significantly alter results
Additional Resources

- Sun Lustre Storage System Blueprint

- External Product Page
  > http://www.sun.com/servers/hpc/storagecluster/
THANK YOU

joeyj@sun.com
+1 505.407.8001