Lustre failover experience

Lustre Administrators and Developers Workshop
Paris
September 25, 2012
• Who we are
• Our Lustre experience: the environment
• Deployment
• Benchmarks
• What's next
Who we are

Company for technology transfer

- HPC services
  - Cluster deployment
  - Storage solution
- Training
  - Sys admin and user oriented programs
- On-demand HPC
• Primary research institute in Italy
  • medical research
• Translational Genomic and Bioinformatics
  • personalized medicine: customization of healthcare by use of genetic information
DNA sequencing

- High-throughput DNA sequencing
- The $1000 genome meme

Next Generation Sequencing is a big data problem!
Customer needs analysis
Customer needs analysis

- Lot of genomic data from Illumina Hi-Seq 2000
  - To backup (~20k € per run)
  - To post-process
  - Always available

Data from the sequencer need to be served to the computational infrastructure

Need for a fast, high performance, highly scalable file system, with robust failover and recovery mechanisms
Infrastructure

InfiniBand QDR
Fibre Channel
Gigabit
Lustre filesystem

2 Lustre filesystems
- 2 MDSs, 3 OSSs
- ~50 clients
- 60 terabytes from SAN

always available!

MDS SERVER
2 x Intel E5645@2.4 Ghz
24 GB RAM

OSS SERVER
2 x Intel E5645@2.4 Ghz
48 GB RAM

OST
Lustre high availability

Lustre clients

Lustre servers

Lustre 1 30 TB

Lustre 2 30 TB

active

standby

active

active

active

active

standby

active

active

active

active

HIGH AVAILABILITY STACK

Hardware

MDS1

MDS2

OSS1

OSS2

OSS3
Lustre high availability

Lustre clients
- Lustre 1: 30 TB
  - Lustre 2: 30 TB

Lustre servers
- active
- standby

DRBD
- primary
- secondary
- sync

Resource manager
- crmd
- stonithd
- cib

Messaging layer
- PACEMAKER

Hardware
- MDS1
- MDS2
- OSS1
- OSS2
- OSS3

COROSYNC
SAN provisioning for Lustre OSSs

HP P2000 G3
2 controllers, 2xFC 8Gb

2 TB nearline SAS
7200rpm
High availability on OSSs

- Failures
  - Power*
  - Fibre channel*
  - InfiniBand*

- Weights distribution, scoring mechanism
  - Each OST has a score with respect to each OSS
  - A OSS mounts an OST when that OST has the highest score on that OSS

*both the links!
High availability on OSSs

<table>
<thead>
<tr>
<th></th>
<th>OST1</th>
<th>OST1</th>
<th>OST2</th>
<th>OST2</th>
<th>OST3</th>
<th>OST3</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSS1</td>
<td>1000</td>
<td>1000</td>
<td>600</td>
<td>800</td>
<td>800</td>
<td>600</td>
</tr>
<tr>
<td>OSS2</td>
<td>800</td>
<td>600</td>
<td>1000</td>
<td>1000</td>
<td>600</td>
<td>800</td>
</tr>
<tr>
<td>OSS3</td>
<td>600</td>
<td>800</td>
<td>800</td>
<td>600</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>
High availability on OSSs

- OSS3 fails
  - OST3LUSTRE1, OST3LUSTRE2 → -INF score
- OSS2, OSS1 receive a new OST
Metadata target

HP 100GB 3G SATA MLC LFF (3.5-inch) SC Enterprise Mainstream Solid State Drive – PCI-e attached
High availability on MDSs

- Failures
  - Power*
  - InfiniBand*

*both the links!
Metadata integrity
High availability on MDSs

- STONITH!
  - MDS1 failovers
  - MDS2 takeovers
Lustre network
Lustre HA: SW stack

- CentOS 6.2
- Lustre 2.1.1
  - Lustre-kernel RPM on I/O server
  - Lustre patchless client RPM modules on clients
  - Lustre iokit
  - Shine
- Pacemaker/corosync
High availability tests

- **Unplug → failover**
  - Power
  - InfiniBand
  - Fibre Channel (on OSS)
  - InfiniBand + Fibre Channel

- **Replug → failback**
  - Automatic on OSSs
  - Manual on MDSs
  - No automatic split-brain resolution!

DOWNTIME = ~120s
## Performance on OSTs

<table>
<thead>
<tr>
<th></th>
<th>XDD 1 thread</th>
<th>Sgpdd-survey 16 threads</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>READ</td>
<td>WRITE</td>
</tr>
<tr>
<td>2 TB SINGLE DISK</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>RAID6 (7 DISKS)</td>
<td>400</td>
<td>400</td>
</tr>
</tbody>
</table>

Results in MB/s
## Performance on MDTs

<table>
<thead>
<tr>
<th></th>
<th>FILE CREATION</th>
<th>OPERATION ON DIRECTORIES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16 threads</td>
<td>64 threads</td>
</tr>
<tr>
<td>MDT on SSD</td>
<td>1000</td>
<td>2600</td>
</tr>
<tr>
<td>MDT on HD</td>
<td>800</td>
<td>1200</td>
</tr>
</tbody>
</table>

Results in operations per second
What's next

- Upgrade to 2.1.3
  - (almost) no downtime thanks to HA
- Monitoring the HA software stack
  - DRBD on MDTs

www.exact-lab.it
info@exact-lab.it