Experiences with HP SFS / Lustre at SSCK

Roland Laifer

Computing Centre (SSCK)
University of Karlsruhe
Germany

Laifer@rz.uni-karlsruhe.de
Outline

» What is HP StorageWorks Scalable File Share (HP SFS)?
  – A Lustre product from HP

» Added values using HP SFS

» Current and planned installations at SSCK

» Experiences with HP SFS
  – at one of the first Lustre production installations in Europe

» Performance measurements and performance monitoring
HP SFS system architecture

Interconnect (Quadrics/Myrinet/GigE/Infiniband)

Client Population

Admin
MDS
OSS
OSS
STO
STO
STO
STO

Connections to site network

Legend
Admin: Administration Server
MDS: Metadata Server
OSS: Object Storage Server
STO: Dual connected storage subsystem:
  - either EVA3000 storage arrays
  - or SFS20 storage arrays
What is HP SFS?

» A Lustre product from HP
   – Available since December 2004

» A Lustre appliance
   – Only dedicated hardware is supported:
     • Servers are Xeon based Proliant systems from HP
     • Storage arrays are SFS20 with SATA disks or EVA3000 with FC disks
     • Restricted number of slots allows only 2 interconnects
   – Special software is delivered:
     • HP supplies a hardened Lustre version
     • Management software implements a single system image
Added values using HP SFS (1)

» Easy installation, configuration and upgrade
  – Server installation of MDS / Admin node from CD
    • OSS get their system images from the Admin node
  – CLI for configuration
    • Complete configuration data is stored in database on shared storage
  – Clean upgrade
    • Upgrade is new installation plus configuration with the existing database

» Software
  – HP runs own tests and puts patches on top of a selected Lustre version
  – HP adds additional software for failover and management
    • All management tasks with CLI on the Admin node
  – HP delivers client build kits and client rpm packages
Added values using HP SFS (2)

» Support
  – HP has an excellent support team
  – Good documentation
    • Includes software implications of all hardware replacements

» Performance monitoring
  – Server performance charts can be displayed with a web browser
  – Client performance data can be listed with HP’s tool collectl

» Problem alerts
  – Automatic problem alerts via email
  – CLI command syscheck verifies the system’s health
  – SFS log database provides fine grained search functions
HP SFS on SSCK's HP XC6000 (phase 1)

120 clients (Itanium)

<table>
<thead>
<tr>
<th></th>
<th>$HOME</th>
<th>$WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>3.8 TB</td>
<td>7.6 TB</td>
</tr>
<tr>
<td>Write performance</td>
<td>240 MB/s</td>
<td>480 MB/s</td>
</tr>
<tr>
<td>Read performance</td>
<td>380 MB/s</td>
<td>760 MB/s</td>
</tr>
</tbody>
</table>
Production experiences with HP SFS (1)

» HP solved all problems and provided patches
  – We still use HP SFS 1.1-1 plus patches
    • based on Lustre version 1.2.6
  – No HP SFS related production problem since 5 months

» Using Lustre for home directories worked well
  – Initially HP provided a patch for memory mapped files
  – Due to POSIX compliance no complaints about failing system calls

» Failover works
  – At the beginning this caused some problems

» Filesystem operations continue after a problem is repaired
  – Usually batch jobs continue to run
Production experiences with HP SFS (2)

» Utilization (capacity and throughput) is steadily increasing
  – Lots of different HPC applications run on the system
  – Highest throughput requirements from
    • using Lustre instead of local disks
    • CAE applications (ISV codes)
    • job restart files

» Understanding Lustre error messages is not easy
  – Some error messages are critical and some are not
    • Error messages when jobs are cancelled or run into timeout
    • Compare time stamps of Lustre errors with job end times

» Performance monitoring is important
  – to understand which applications are doing IO
  – to recognize possible problems
Sequential write / read performance

- 4 OSS
- SFS version 1.1-0
- 400 / 300 MB/s from one process
- 120 / 190 MB/s per OSS
Lustre scalability

- 2 OSS
- SFS version 1.1-0
- no performance degradation with many clients
Performance with new SFS version

- 1 OSS
- SFS version 2.1-0 with ext3 option *extents*
- Write performance 15% better than with version 1.1-0

![Bar chart showing throughput (MB/s) vs number of nodes.

Throughput (MB/s) vs Number of nodes:
- 1 node: 120 MB/s (write), 180 MB/s (read)
- 2 nodes: 110 MB/s (write), 170 MB/s (read)
- 3 nodes: 100 MB/s (write), 160 MB/s (read)
Metadata performance

- SFS version 2.1-0
- Up to 5000 file operations per second
OSS hardware performance with EVA5000 storage

» Quadrics Elan4
  - Internally about 1300 MB/s
  - Only PCI-X adapters exist

» PCI-X bus on servers
  - About 900 MB/s

» Dual-ported FC adapter
  - About 195 MB/s
  - Actually only 1 port is used

» EVA5000 storage array
  - About 140 MB/s for writes
  - Nearly 500 MB/s for reads
OSS hardware performance with SFS20 storage

» Quadrics Elan4
  – Internally about 1300 MB/s
  – Only PCI-X adapters exist

» PCI-X bus on servers
  – About 900 MB/s

» 2x U320 SCSI adapter
  – About 640 MB/s

» 4x SFS20 storage array
  – About 400 MB/s for writes
  – About 600 MB/s for reads
Performance monitoring on one OSS

Statistics for xc1-1s3 between 10:10:00 and 11:10:00 on 20060213

Aggregate Lustre OSTs
Read KB/s —
Write KB/s —
Read Ops —
Write Ops —
Performance monitoring on the MDS
HP SFS on the upcoming HP XC4000 (phase 2)

760 clients (Opteron)

InfiniBand 4X DDR Interconnect

Capacity
- 8 TB
- 48 TB

Write performance
- 400 MB/s
- 2400 MB/s

Read performance
- 600 MB/s
- 3600 MB/s
Plan for a central parallel file system

Parallel file system

Ethernet Switch

NFS Server

CIFS Server

Campus network
Summary

» Lustre provides a stable parallel file system

» Sequential IO in Lustre nearly reaches hardware performance

» HP SFS supplies additional features
  – which make it a real product

» SSCK uses HP SFS successfully since more than one year
  – See http://www.rz.uni-karlsruhe.de/dienste/lustretalks.php