Rock-Hard Lustre
Trends in Scalability and Quality
nathan_rutman@xyratex.com
Big Iron

• Lustre dominates the top end today

2010 numbers:

• 100% of the top 3
• 70% of the top 10
• 66% of the top 100
  – 59/100 Lustre + 8 suspected
  – 22/100 GPFS
  – 3/100 PanFS
  – 1/100 CXFS
  – 6/100 unknown, suspect non-Lustre
## Lustre Installations

<table>
<thead>
<tr>
<th>Rank</th>
<th>Machine</th>
<th>Storage</th>
<th>Speed</th>
<th>Clients</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Riken “K”</td>
<td>4PB</td>
<td>1TB/s</td>
<td>64,512</td>
<td>Fujitsu Exabyte File System</td>
</tr>
<tr>
<td>2</td>
<td>Tianhe-1A</td>
<td>2PB</td>
<td></td>
<td>7168</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ORNL Jaguar</td>
<td>10PB</td>
<td>240GB/s</td>
<td>18,688</td>
<td>Biggest U.S.</td>
</tr>
<tr>
<td>5</td>
<td>Tsubame 2</td>
<td>11PB</td>
<td></td>
<td>1408</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>NASA Pleiades</td>
<td>5.1PB</td>
<td></td>
<td>11776</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>CEA Tera 100</td>
<td>20PB</td>
<td>500GB/s</td>
<td>4324</td>
<td>Biggest EU, Lustre 2.0</td>
</tr>
<tr>
<td></td>
<td>Riken HPCI</td>
<td>30PB</td>
<td>720GB/s</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Riken “K”</td>
<td>100PB</td>
<td>1TB/s</td>
<td></td>
<td>upgrade</td>
</tr>
<tr>
<td></td>
<td>LLNL Sequoia</td>
<td>50PB</td>
<td>500GB/s</td>
<td>98,304</td>
<td>expect 1TB/s</td>
</tr>
</tbody>
</table>
Continuing Systems Growth

• Biggest systems get bigger
  – Lustre scale barriers continue to fall
  – Nothing new

• Larger systems become more affordable
  – More customers need Lustre
New Scaling Features in Lustre 2.1

- Complete rework of MDS and Client IO stacks
- EXT4
  - 16TB file size (=stripe object limit)
  - 4B files
  - unlimited files/dir
  - faster fsck
    - skip uninitialized bitmaps
    - skip unused inodes
    - checksum group descriptors
  - faster mkfs
- MMP
- MDRAID improvements
- Large LUN
  - 128TB LUNs
Future Community Scaling Projects

- Wide striping
  - 1350+ OSTs
- SMP scaling
  - Vastly improved MD rates
- Simplified SOM
  - ‘ls -l’
- Flash cache
- DNE
Quality
Lustre Quality

• Early perceptions of Lustre: “it’s a science project”
• Sun Microsystems purchased CFS in 2007
• Needed a solid foundation to make $
Stability / Quality Improvements

• Testing
  – LBATS, YALA automated build and test
  – Hyperion
  – feature-specific tests
  – scale tests

• Process
  – Short development cycles
  – Strict inspections
  – Branch gatekeepers
  – Train model

• Documentation
  – LID
  – Oak Ridge’s Lustre Internals
  – Subsystem map
  – Doxygen
• The quality focus remains post-Oracle
  – Xyratex, Whamcloud both follow the same quality processes
  – OpenSFS has added community inspectors to the SOW acceptance criteria
  – Community-based testing

• Recognition that the quality initiatives have paid off
  – Among developers first
  – Among users
  – Growing user base
Growth of Lustre Installations and Users

- Many significant contributors
  - Whamcloud, Xyratex, Cray, LLNL, ORNL, CEA, Bull, TACC, DDN
Positive Feedback Loop

• Two facts
  – Top end hardware moves down
  – Lustre is open source

• Imply two trends
  – Community base will continue to broaden
    • Community contributions will increase
  – Quality will continue to improve
Future Quality Features

• T10-DIF
  – Prevents server-to-drive corruption

• End-to-End Data Integrity
  – Prevents client-to-drive corruption, including network

• On-line LFSCK
  – Continuous verification and repair of metadata

• Imperative Recovery
  – Accelerate recovery for big systems
Improved High Availability in Integrated Solutions

- Integrated solutions like the ClusterStor 3000
  - Reduce the complexity
  - Encapsulate the HA
- Data access always provided for any single point failure
  - Switch, OSS controller, RAID, management server
- Eliminated controller-drive cabling
- Separate management network
- Integrated monitoring
- Integrated HA software
Thank You

nathan_rutman@xyratex.com