HPC Storage – Breaking the I/O Bottleneck

Steve Paulhus
Xyratex
Petascale computing is not just about the FLOPs... It will also require a new breed of HPC storage technology capable of feeding the beast!
HPC Storage I/O Bottleneck - Not just a Top10 Problem

- Look at latest Top500 – HPC Storage is not just a Top10 problem
  - Top10 > 1PFLOP
  - Top120 > 100TFLOPs
  - 500th System > 50TFLOPs

The deployment growth in CPUs and Cores is far outpacing the growth in available storage bandwidth!

Today: At 100GB/s per PFLOP, HPC Storage Bandwidth Requirements for the Top120 is up to 1TB/s
Higher Bandwidth HPC Interconnects Are Helping...
HPC Storage: Traditional vs. Dynamic Scaling

**Traditional Scaling Model**
- Scalable Storage Unit (SSU)

**Dynamic Scaling Model**
- Overall Rack Amounts
  - Network I/O Ports:
  - Compute & RAM:
  - Total HDDs:

**Balanced Performance & Scalability**
- Throughput: >24 GB/sec
- Throughput: 10 GB/sec

© 2012 Xyratex, Ltd
## SSU Rack Aggregates/Totals

<table>
<thead>
<tr>
<th>SSU</th>
<th>3TB HDD TBs: (U/R)</th>
<th>4TB HDD TBs: (U/R)</th>
<th>GB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSU</td>
<td>192 / 240</td>
<td>256 / 320</td>
<td>3.5</td>
</tr>
<tr>
<td>SSU</td>
<td>384 / 480</td>
<td>512 / 640</td>
<td>7</td>
</tr>
<tr>
<td>SSU</td>
<td>576 / 720</td>
<td>768 / 960</td>
<td>10.5</td>
</tr>
<tr>
<td>SSU</td>
<td>768 / 960</td>
<td>1024 / 1280</td>
<td>14</td>
</tr>
<tr>
<td>SSU</td>
<td>960 / 1200</td>
<td>1280 / 1600</td>
<td>17.5</td>
</tr>
<tr>
<td>SSU</td>
<td>1152 / 1440</td>
<td>1536 / 1920</td>
<td>21</td>
</tr>
<tr>
<td>SSU</td>
<td>1344 / 1680</td>
<td>1792 / 2240</td>
<td>24.5</td>
</tr>
</tbody>
</table>
## TCO Analysis – Traditional vs. Dynamic Scaling

<table>
<thead>
<tr>
<th>1TB/second Aggregate Bandwidth</th>
<th>Traditional HPC Storage System</th>
<th>Dynamic HPC Storage System</th>
<th>Dynamic vs. Traditional Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Racks:</td>
<td>55</td>
<td>36</td>
<td>-35%</td>
</tr>
<tr>
<td>Data Center Floor Space:</td>
<td>91.5 (m²)</td>
<td>60 (m²)</td>
<td>-34%</td>
</tr>
<tr>
<td>Number of Enclosures:</td>
<td>550</td>
<td>252</td>
<td>-54%</td>
</tr>
<tr>
<td>Hard Drives:</td>
<td>46,200</td>
<td>17,280</td>
<td>-63%</td>
</tr>
<tr>
<td>Total Power:</td>
<td>~0.859MW</td>
<td>~0.443MW</td>
<td>-48%</td>
</tr>
<tr>
<td>Total Heat Dissipation (BTUs):</td>
<td>~2,728,000</td>
<td>~1,165,600</td>
<td>-57%</td>
</tr>
</tbody>
</table>

### Dynamic HPC Storage Scaling:
*Exponentially less cost, space, cooling and power – less is more!*

© 2012 Xyratex, Ltd
“Never waste the opportunities offered by a good crisis…”

- Niccolò Machiavelli (1469 – 1527)

- Requirements for HPC storage in Petascale systems will translate into an unprecedented increment over existing storage implementations and technologies

- HPC Storage must make significant advances in at least three areas:
  - Reduced Latencies – Increased levels of storage hardware and storage integration
  - Dynamic Scaling – Linear performance
  - Reduced Footprint – Significantly less cost, space and power!
Introduction to ClusterStor CS-3000
ClusterStor 3000 - Scalable Storage Unit (SSU)

- **Ultra HD - CS-2584 SSU**
  - 5U/84 Enclosure – completely H/A
    - Two (2) trays of 42 HDD’s each
    - Dual-ported 3.5” NL-SAS & SSD HDD Support
    - 150MB/s SAS available bandwidth per HDD
    - Data Protection/Integrity (RAID 6, 8+2)
    - 2x SSD OSS journal disks
    - 2x Hot Spare Disks
  - SSU: 2x H/A Embedded Application Servers
    - CS-3000 SSU: =3.5GB/sec
  - Network Links per SSU
    - CS-3000: Dual QDR IB or Dual 10GbE
  - 64 Usable Data Disks per SSU
    - 1TB x 64 – 64TB usable per SSU
    - 2TB x 64 - 128TB usable per SSU
    - 3TB x 64 - 192TB usable per SSU
    - 4TB x 64 – 256TB usable per SSU

Only 5°C delta with drawer open
ClusterStor CS-3000 Overview

CS-3000 Overview
- Targeted >20GB/s per rack
- Overall Performance scalable to >100GB/s bandwidth
- Overall Capacity scalable to >30PBs
- Xyratex delivers a complete ready-to-run ClusterStor solution
  - Sizing and Configuration optimization
    - Performance centric
    - Capacity centric
  - Factory Integration & Staging
    - Complete rack integration & Cabling
    - Entire storage software stack factory pre-installed and pre-configured
    - System Burn-in and benchmark testing area at Xyratex factory
    - Up to 30-day “soak” test
- Up to 560 HDD’s per rack (42RU)
- Up to 2.2PBs per rack (with 4TB HDD’s)
- Up to 14 Embedded Servers per rack
- Up to 14 high bandwidth Network connectivity connections per rack

© 2012 Xyratex, Ltd
## ClusterStor - Enhanced Usability

<table>
<thead>
<tr>
<th>Task</th>
<th>Traditional HPC Storage</th>
<th>ClusterStor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration Optimization</td>
<td>Multiple discreet building blocks, myriad of third-party components and options, interoperability issues</td>
<td>Simplified configurator model with single-vendor performance or capacity centric configurations</td>
</tr>
<tr>
<td>Software Installation</td>
<td>&gt;5 Days</td>
<td>Factory Installed</td>
</tr>
<tr>
<td>Out of Box Experience</td>
<td>&gt;5 Days to get system up and running</td>
<td>Out of box to configured storage cluster in &lt;4 hours</td>
</tr>
<tr>
<td>Performance Tuning</td>
<td>2-4 Weeks balancing OSS, networking, RAID and JBOD systems</td>
<td>Immediate (Factory Set)</td>
</tr>
<tr>
<td>Storage Expansion</td>
<td>3-5 Weeks</td>
<td>&lt;5 Hours</td>
</tr>
<tr>
<td>Acceptance Testing</td>
<td>Standard (weeks/months)</td>
<td>Shortened (Days/Weeks)</td>
</tr>
</tbody>
</table>
Introduction to Xyratex
Leading OEM Provider of Digital Storage Technology

- **Largest WW OEM** Disk Storage System provider*

- ~**19% of worldwide external** storage capacity shipped in 2011

- > **4.1 Exabytes** of storage shipped in 2011

- **Shipped our 1 Millionth** Enclosure in 2011

- ~**50% of WW disk drives** are produced utilizing Xyratex Technology (Company estimates)

- **Over 25 years** of storage & networking R&D experience & innovation

*Source: Internal analysis, IDC, December 2011; Coughlin Associates, 2010

© 2012 Xyratex, Ltd
Global Customer Relationships

Strategic relationships established with both market & technology leaders
Xyratex not New to HPC Storage

- Xyratex has had a successful history in supplying storage building blocks for HPC storage solutions through multiple generations of storage platforms and technologies via our OEM Partner base.

- In 2009 Xyratex decided to develop a new range of products specifically tailored to the needs of HPC.

- Xyratex continues to make strategic investments to strengthen our HPC market position:
  - Development of next generation storage platforms, diagnostic software and controllers.
  - ClusterStor acquisition – 2010 brought deep and extensive Lustre FS expertise:
    - Peter Braam & team
  - Continue to augment HPC core competencies with additional HPC talent.
  - Long term commitment to HPC.

Top 5 Vendors to Watch

© 2012 Xyratex, Ltd
## Roles in the Lustre® Community

<table>
<thead>
<tr>
<th>OpenSFS Board Member &amp; EOFS Member</th>
<th>Active Contributor to Lustre Source &amp; Roadmap</th>
<th>Integration of Lustre into ClusterStor™</th>
<th>Lustre Support Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Direct funding of Lustre tree &amp; roadmap development</td>
<td>- World class Lustre development team on staff</td>
<td>- Industry leading HPC storage solutions</td>
<td>- ClusterStor, Lustre &amp; 3rd party hardware</td>
</tr>
</tbody>
</table>
THANK YOU – QUESTIONS?

The ClusterStor 3000
Xyratex Designed, Developed, Manufactured and Supported Storage Solution

- Architected - Revolutionary ‘scale-out’ storage architecture
- Developed – From the ground up - 150 software engineers
- Integrated - Tight integration at every layer of the stack
- Optimized & Qualified - Proven enterprise class validation process
- Supported - Industry leading diagnostics from the lowest level of the hardware up through the entire software stack

© 2012 Xyratex, Ltd