Another Year, Another Petabyte

A Look Into the Laureate Institute for Brain Research’s CephFS Deployment
Research At LIBR

• Neuroscience-based, clinical and developmental research:
  1. To develop neuroscience-based individually predictive assessments.
  2. To develop novel brain-body based interventions
     • Focus: mood, anxiety, addiction, or eating disorders.
  3. To use experimental systems to quickly test assessments and interventions.
The LIBR Facilities
My Inheritance
November 2015

- Foundry Big Iron RX-16
- EMC Isilon
- Oracle ZFS Appliance
- Spectra Logic Black Pearl
- Zmanda
- Spectra Logic T950e
The Problem

- Depletion of compute/storage resources
- Neglected infrastructure
- $$$
- Inadequate performance
- Unreliable
Network Diagram
Options Presented to Stakeholders
January 2016

• Scale existing Isilon with used hardware ($1.3M)
• Panasas ($630k)
• Oracle ZFS ($450k)
• DiY scale-out + Network Refresh ($450k)
  – Lustre
  – Gluster
  – Ceph
  – OrangeFS
Slow Progress

- February – Approval to develop pilot
- June – pilot approval
- July – Pilot Equipment arrival
- July – CEO approval to proceed
Board of Directors Approval
August 2016

• Placed Order

• Brings configuration to:
  – 8x OSD (24 disk + 2 journal NVMe per node)
    • 1.1PB raw
  – 3x MON
  – 2x MDS
Ceph Design

• 8x OSD nodes
  – 256 GB RAM
  – 2x Intel S3610 for OS
  – 24x 6TB Enterprise SATA (24 slot chassis)
  – 2x Intel P3700
  – 2x Xeon 2660 V4
    • 28x2.0GHz
  – Dual 40GB Ethernet Adapter (Mellanox ConnectX-4)
Ceph Design (Cont...)

- 2x MDS + 3x MON node
  - 128 GB RAM
  - 2x Intel S3610 for OS
  - 2x Xeon 2643 V4
    - 12x3.4GHz
  - Dual 40GB Ethernet Adapter (Mellanox ConnectX-4)
Other Hardware and Software

• Brocade VDX
• Spectra Logic
• Zabbix
• ELK
• Nfs-ganesha
• Samba
<table>
<thead>
<tr>
<th>Blocksize</th>
<th>MB/sec Read</th>
<th>MB/sec Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>4kB</td>
<td>6.845013</td>
<td>3.221325</td>
</tr>
<tr>
<td>16kB</td>
<td>27.43865</td>
<td>13.15497</td>
</tr>
<tr>
<td>64kB</td>
<td>103.5336</td>
<td>51.65482</td>
</tr>
<tr>
<td>256kB</td>
<td>424.7271</td>
<td>193.5268</td>
</tr>
<tr>
<td>1MB</td>
<td>1576.252</td>
<td>782.4383</td>
</tr>
<tr>
<td>4MB</td>
<td>2619.374</td>
<td>3095.016</td>
</tr>
</tbody>
</table>

Random I/O 100% Read vs 100% Write
<table>
<thead>
<tr>
<th>Blocksize</th>
<th>MB/sec Read</th>
<th>MB/sec Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>4kB</td>
<td>3.806028</td>
<td>3.780054</td>
</tr>
<tr>
<td>16kB</td>
<td>16.70582</td>
<td>13.7856</td>
</tr>
<tr>
<td>64kB</td>
<td>78.28159</td>
<td>57.00872</td>
</tr>
<tr>
<td>256kB</td>
<td>449.4757</td>
<td>203.128</td>
</tr>
<tr>
<td>1MB</td>
<td>1790.493</td>
<td>791.1742</td>
</tr>
<tr>
<td>4MB</td>
<td>6158.911</td>
<td>3112.754</td>
</tr>
<tr>
<td>Blocksize</td>
<td>MB/sec Read</td>
<td>MB/sec Write</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td>4kB</td>
<td>3.873204</td>
<td>1.331676</td>
</tr>
<tr>
<td>16kB</td>
<td>14.99433</td>
<td>4.84942</td>
</tr>
<tr>
<td>64kB</td>
<td>60.45827</td>
<td>20.30537</td>
</tr>
<tr>
<td>256kB</td>
<td>233.5852</td>
<td>81.27407</td>
</tr>
<tr>
<td>1MB</td>
<td>771.8829</td>
<td>327.8226</td>
</tr>
<tr>
<td>4MB</td>
<td>1773.824</td>
<td>1301.71</td>
</tr>
</tbody>
</table>

Random I/O 60% Read/40% Write

- **MB/sec Read**
- **MB/sec Write**

![Bar chart showing MB/sec Read and MB/sec Write for various block sizes ranging from 4kB to 4MB.](chart.png)
<table>
<thead>
<tr>
<th>Blocksize</th>
<th>MB/sec Read</th>
<th>MB/sec Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>4kB</td>
<td>3.082823</td>
<td>1.739016</td>
</tr>
<tr>
<td>16kB</td>
<td>12.86438</td>
<td>5.733841</td>
</tr>
<tr>
<td>64kB</td>
<td>56.4047</td>
<td>23.77207</td>
</tr>
<tr>
<td>256kB</td>
<td>280.6739</td>
<td>81.68536</td>
</tr>
<tr>
<td>1MB</td>
<td>820.0511</td>
<td>320.4626</td>
</tr>
<tr>
<td>4MB</td>
<td>2427.747</td>
<td>1260.187</td>
</tr>
</tbody>
</table>
Current Status

• The Good
  – Power users and their projects migrated
    • No more “file not found” errors
    • Computation > 100% faster
  – Backups running and performing nicely
  – Everybody is fairly happy
  – Very resilient to many types of failure

• The Bad
  – Snapshots