Cray Storage Update LUG 2018
Nathan Schumann, Sr. Product Manager
This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray’s documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.
### Cray ClusterStor Platforms

**ClusterStor L300**
- Lustre 2.7
- 12 Gbit SAS enclosures
- Broadwell based ESMs
- EDR IB
- OPA
- 100 GbE
- 2x 40 GbE
- 6/8/10 TB HDDs

**ClusterStor L300N**
- NXD I/O Manager
- 3.2 TB SSDs
- Advanced MMU

**Large sequential I/O workloads**

**Many applications with mixed I/O patterns**
ClusterStor L300N: Any Workload, Any Time

Goal: Support Widest Application I/O Variety Within Budget

Challenge: Lost Productivity Due To Poor I/O Predictability

Solution: Smart I/O Management Provides Consistent Performance

Result: I/O Optimized For Each Application I/O Profile

Compute Cluster Applications
Sensor Data Capture & Processing
Data mining - Simulations
Modelling - Software development
Visualization of complex data - Rapid mathematical calculations

Mixed I/O Patterns=
Unpredictable Application Performance

ClusterStor NXD =
Transparent Redirection of I/O to Appropriate SSD or HDD Medium

ClusterStor NXD =
Predictable and Fastest Compute Application Time-to-Solution

Random
Un - Aligned
Small Block
Sequential
Three Horizons of ClusterStor

- **DEFINE** HPC Storage with engineered rack-scale Lustre systems
- **REFINE** HPC Storage with hybrid Lustre systems
- **REDEFINE** HPC Storage (again 😁) with All-Flash Lustre systems

Customer Value


Primary HPC Storage

Colors: All HDD, Hybrid, All Flash
## Lustre Improvements for Flash

<table>
<thead>
<tr>
<th>Reduce client code overhead</th>
<th>Reduce network RPCs</th>
<th>Reduce MDS overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small IO improvements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Tiny writes [2.11]*</td>
<td>• Lock Ahead [2.11]*</td>
<td>• DNE.2 [2.8] (reduce MDS bottlenecks)</td>
</tr>
<tr>
<td>• Fast reads [2.11]*</td>
<td>• Immediate short IO [2.11]*</td>
<td>• Progressive file layouts [2.10] (more compact layouts)</td>
</tr>
<tr>
<td></td>
<td>• Small overwrites [2.11]*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Data on MDT [2.11]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Size on MDT*</td>
<td></td>
</tr>
</tbody>
</table>

Other improvements:
- OST/MDT server tunings for SSD
- Multi-channel PTLRPC (QoS)
ClusterStor Lustre Release Adoption

- New Lustre community release
- Contribute work upstream to community master
- Re-baseline to community release
- Implement bug fixes and features for Cray customers
- Curate bug fixes from community master
## Lustre Features Coming to ClusterStor

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Lustre Release</th>
<th>ClusterStor Release</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lock ahead</td>
<td>2.11</td>
<td>2.x</td>
</tr>
<tr>
<td>Multiple MD modify RPCs per client</td>
<td>IEEL 3.0, 2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>DNE phase IIb (async commit and recovery)</td>
<td>2.8</td>
<td>3.next</td>
</tr>
<tr>
<td>UID/GID mapping</td>
<td>2.9</td>
<td>3.next</td>
</tr>
<tr>
<td>Subdirectory mounts</td>
<td>2.9</td>
<td>3.next</td>
</tr>
<tr>
<td>Large bulk I/O</td>
<td>2.9</td>
<td>3.next</td>
</tr>
<tr>
<td>Multirail LNet</td>
<td>2.10</td>
<td>3.next</td>
</tr>
<tr>
<td>Progressive file layouts</td>
<td>2.10</td>
<td>3.next</td>
</tr>
<tr>
<td>Project quotas</td>
<td>2.10</td>
<td>3.next</td>
</tr>
<tr>
<td>Data on MDT (DoM)</td>
<td>2.11</td>
<td>3.next</td>
</tr>
<tr>
<td>File-level redundancy (FLR) - delayed resync</td>
<td>2.11</td>
<td>3.next</td>
</tr>
</tbody>
</table>
Optimization

If you can’t measure it, you can’t improve it.
Job Performance Details and Scoring

Jobs
Searchable fields for the most important data to the job step level

Scoring
Know which jobs might be causing issues with visual cues
Performance Visualization and Comparison

Visualize
Performance graphs over the life of the job for write, read, and metadata operations

Compare
Compare this job to the rest of the system at a glance
Questions?