Tiered Storage

Steve Lyness
VP, Solutions Architecture

HPC Users Forum, Sept 11, 2019
HPC STORAGE SYSTEM PARADIGM SHIFT

Most cost-effective to deliver IOPS (IO operations/sec)

Most cost-effective to deliver THROUGHPUT (GB/sec)

Most cost-effective to deliver CAPACITY (Petabyte)

© 2019 Cray Inc.
> 60% of the cost of any storage system is in the media

SSD Media = PERFORMANCE
+ some initial capacity

HDD Media = CAPACITY
+ some performance

SOFTWARE = Workflow-accelerating data placement
on the right media at the right time

CHANGING ECONOMIES OF STORAGE MEDIA

$0.380 $0.250 $0.180 $0.140 $0.110 $0.081 $0.060 $0.000 $0.050 $0.100 $0.150 $0.200 $0.250 $0.300 $0.350 $0.400
2018 2019 2020 2021 2022 2023 2024

SSD $/GB HDD $/GB

© 2019 Cray Inc.
<table>
<thead>
<tr>
<th></th>
<th>CN Metadata Array</th>
<th>CN Flash Array</th>
<th>CN Disk Array</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form Factor</strong></td>
<td>2U24 PCIe 4.0 NVMe</td>
<td>2U24 PCIe 4.0 NVMe</td>
<td>4U106 SAS HDD</td>
</tr>
<tr>
<td><strong>Servers</strong></td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>SSDs Per Enclosure</strong></td>
<td>24</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td><strong>HDDS Per Enclosure</strong></td>
<td>0</td>
<td>0</td>
<td>106</td>
</tr>
<tr>
<td><strong>Optimal Use</strong></td>
<td>Lustre Metadata</td>
<td>High BW/IOPS</td>
<td>Capacity</td>
</tr>
<tr>
<td><strong>Max Per File System</strong></td>
<td>8 (16 servers)</td>
<td>16 per rack</td>
<td>Configuration Dependent</td>
</tr>
</tbody>
</table>
ClusterStor and Slingshot simplify Cray storage

Benefits:
- Lower cost
- Lower complexity
- Lower latency
- Improved small I/O performance
• ClusterStor Data Services provides similar features
  • Integration with WLM for data promotion/demotion orchestration
  • Compute resources allocated after any requested data movement
  • High-speed flash directly on high-speed network
• DataWarp for the masses with >4-6x speed-up over disk
OPTIMIZING APPLICATION PERFORMANCE

**Placement (FCS 2020)**
- What is this?
  - Bridge capabilities from the L300 & L300F
  - Use flash pool for temp data
  - User can direct Lustre to move data up to the flash tier

**Scheduled (2H 2020)**
- What is this?
  - Bridge from DataWarp
  - User inserts WLM directives into the job script
  - Automated migration is performed outside of run time

**Transparent (1H 2021)**
- What is this?
  - A new class of service
  - Read through, write back
  - Transparent tiering
  - No user impact

**On Demand (roadmap)**
- What is this?
  - Advancement from DataWarp
  - Quality of service enhancement
  - Dynamically provisioned via the WLM
PERLMUTTER
 FIRST CLUSTERSTOR NEXT CUSTOMER

30 PB ALL FLASH

>4 TB/SEC
7M IOPS
3.2M file creates/sec.
Lab-wide Community and Campaign Storage

- 200+ PB
- 1+ TB/s
- 1st Half 2020 delivery
- Integrated with DAOS and Aurora in 2021
Call your local Cray Sales Rep for a Deeper Dive on ClusterStor NEXT