INTEGRATED HPC/AI STORAGE ROADMAP & CRAY CLUSTERSTOR E1000

John Howarth, Product Management, Cray, a Hewlett Packard Enterprise company
Uli Plechschmidt, Product Marketing, Cray, a Hewlett Packard Enterprise company

November 16, 2019
## We are converging on ClusterStor E1000

<table>
<thead>
<tr>
<th></th>
<th>HPE Scalable Storage for Lustre</th>
<th>Cray ClusterStor L300</th>
<th>Cray ClusterStor E1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast NVMe Flash for ML/DL workloads</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Aligning the data flow with the workflow</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Strong position in high-end HPC</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Single system image management</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Standard factory integration/validation</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Leading HDD performance efficiency</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ZFS data path option</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Lustre LTS release 2.12</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>High density 4U106 enclosures</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Benefits of being a HPE product</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>CY2020</td>
<td>CY2021</td>
<td>CY2022</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Q4CY19</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Storage Systems**

- **Cray ClusterStor L300**
- **HPE SS4L***

**Transition Information**
ClusterStor L300 available throughout CY2020
Last ship date HPE SS$L February 2020

**Cray ClusterStor E1000**

*Mid Life Kicker in 2H-CY2021*

**Storage Software**

- **ClusterStor Neo 3.x**
- **ClusterStor Neo 4.0**
- **ClusterStor System Management Services 1.x**
- **ClusterStor SMS 2.x**
- **Clay View for ClusterStor 1.x**
- **Cray View for ClusterStor 2.x**
- **Cray View for ClusterStor 3.x**
- **HPE Data Management Framework V7.x**
- **ClusterStor Data Services**

**ClusterStor Data Services Tiering Functionality:**
- Scripted Tiering
- Scheduled Tiering (mid-CY2020)
- Transparent Tiering (1H-CY2021)

**Merged Suite of Data Services**

*HPE Scalable Storage for Lustre*

**Notes:**

- Shipping
- POR
- Investigation
A NEW ERA NEEDS NEW STORAGE: CRAY CLUSTERSTOR E1000

Uli Plechschmidt (uplechschtm@cray.com)
MODELING & SIMULATION + ARTIFICIAL INTELLIGENCE + BIG DATA ANALYTICS = EXASCALE ERA

RUNNING ON ONE MACHINE IN MISSION-CRITICAL WORKFLOWS
The majority (61%) of the HPC users in every economic sector are running machine learning programs today.

Intersect360 HPC User Budget Map Survey: Machine Learning’s Impact on HPC Environments, October 2019
Machine learning will increasingly be used to replace portions (sub-models) of traditional simulations based on partial differential equations. This is already being done by developers in areas like quantum chemistry, high energy physics, and material science. Even early attempts at this have seen sub-model performance speedups as high as 1,000X and even 10,000X.

Rick Stevens, ANL Associate Lab Director

https://www.nextplatform.com/2019/10/22/exascale-is-not-your-grandfathers-hpc
**Key Challenge: Disruptive Workload Convergence**

<table>
<thead>
<tr>
<th>Primary workload</th>
<th>Primary compute</th>
<th>Primary storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling &amp; Simulation</td>
<td>Converged</td>
<td>HPC/AI ML/DL</td>
</tr>
<tr>
<td>CPU nodes</td>
<td>CPU/GPU nodes</td>
<td>GPU nodes</td>
</tr>
<tr>
<td>HDD-based parallel storage (ClusterStor L300, DDN EXAScaler, IBM ESS, Lenovo DSS-G etc.)</td>
<td>?</td>
<td>All Flash enterprise storage (NetApp AFF, Pure Storage FlashBlade, Dell EMC Isilon, WekaIO Matrix etc.)</td>
</tr>
<tr>
<td>Petabyte</td>
<td>Petabyte</td>
<td>Terabyte</td>
</tr>
<tr>
<td>Large</td>
<td>Mixed</td>
<td>Mixed</td>
</tr>
<tr>
<td>Sequential</td>
<td>Mixed</td>
<td>Mixed</td>
</tr>
<tr>
<td>Extensive</td>
<td>Extensive</td>
<td>Moderate</td>
</tr>
<tr>
<td>Write-intensive</td>
<td>Both</td>
<td>Read-intensive</td>
</tr>
<tr>
<td>HDD</td>
<td>Both</td>
<td>SSD</td>
</tr>
</tbody>
</table>

- **Capacity measured in**: Petabyte, Petabyte, Terabyte
- **Typical file size**: Large, Mixed, Mixed
- **Typical file access**: Sequential, Mixed, Extensive
- **Data movement**: Extensive, Extensive, Moderate
- **Primary I/O pattern**: Write-intensive, Both, Read-intensive

*HDD do not serve mixed, random I/O well*

*Price-prohibitive to scale to multi PB needs*
CAN YOU GUESS WHAT THOSE NUMBERS MEAN?

Up to 1,600 gigabyte per second throughput* PER FLASH RACK

Up to 341 SSDs needed to drive 1,000 GIGABYTE per second* PER DISK RACK

Up to 120 gigabyte per second throughput PER DISK RACK

*sequential read  ** sequential write - all performance numbers for expansion racks
CRAY CLUSTERSTOR E1000 “AT A GLANCE”

THE CORE:
NEW ClusterStor Data Services:
• Intelligent orchestration of the data flow with the workflow
• Three modes to match specific customer preferences

NEW Cray Slingshot™ 200 Gbps 64 port TOR switches
• Advanced congestion management and separate storage traffic classes
• 100/200 GbE and IB EDR/HDR to connect to non-Cray Shasta compute

NEW dense rack-scale packaging
• Up to 1.6 TB/sec (read) and 1.2 TB/sec (write) per SSD rack
• Up to 120 GB/sec and 10 PB usable per HDD rack
• Entry config starts at 6 rack units

NEW Cray designed, end-to-end PCIe 4.0 storage controllers
• Serving up to 3.3 GB/sec from each SSD to the file system clients
• Serving more than 140 MB/sec from each HDD to the file system clients

Lustre 2.12 with NEW functionality
• Highest performance/lowest cost parallel file system
• Open source software with enterprise-grade support
• Hardened by Cray for reliability @ scale
NEW ZERO BOTTLENECK PCIe 4.0 STORAGE CONTROLLERS

Two embedded HA application modules with single socket AMD “Rome” PCIe 4.0 CPU each

Up to 6 x 100/200 Gbps PCIe 4.0 NICs (Cray Slingshot, GbE, IB)

Flexible building block for ClusterStor E1000

System Mgmt Unit (5 SSD)

Metadata Unit (24 SSD)

All Flash Unit (24 SSD)

Disk Controller (2 SSD)

One per file system

Up to 10 per file system

Up to 24 x dual ported NVMe PCIe 4.0 SSD in the 2U enclosure

Up to 424 x 7.2K RPM HDD SAS-attached to the 2U enclosure

Supported storage drives

NVMe Gen 4

7.2K RPM

3 DWPD:

1.6/3.2/6.4/12.8 TB

4/6/10/12/14/16 TB

1 DWPD:

1.92/3.84/7.68/15.36 TB

SAS-attached 4U106 HPE D8000 HDD enclosures
“RACK SCALE” SYSTEM ARCHITECTURE

Base rack
- Networking infrastructure
- System Management Unit (SMU)
- MetaData Unit (MDU)
- Initial Scalable Storage Units (SSU)

Scale out (nearly) linear with Scalable Storage Units (SSU) in expansion racks*

Flash SSU
(All Flash)
- Each SSU adds
  - up to 80 GB/s read
  - up to 60 GB/s write
  - up to 253 TB usable capacity (15.3 TB SSD)
- in a 2U form factor
  - up to 20 SSU per rack

Disk SSU
(Balanced)
- Each SSU adds
  - up to 30 GB/s read
  - up to 30 GB/s write
  - up to 2.5 PB usable capacity (16 TB HDD)
- in a 10U form factor
  - up to 4 SSU per rack

Disk SSU
($/GB optimized)
- Each SSU adds
  - up to 30 GB/s read
  - up to 30 GB/s write
  - up to 5 PB usable capacity (16 TB HDD)
- in a 18U form factor
  - up to 2 SSU per rack

*add MDUs if necessary to scale metadata performance
### Powerful Rack Level Specifications

<table>
<thead>
<tr>
<th>HDD Configuration</th>
<th>Hybrid Configuration</th>
<th>Flash Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lots of small, random read I/O</td>
<td>Mixed I/O, Converged workloads</td>
<td>Lots of small, random read I/O</td>
</tr>
<tr>
<td>AI/ML/DL</td>
<td>Single digit PB to EB</td>
<td>AI/ML/DL</td>
</tr>
<tr>
<td>Terabytes to single digit PB</td>
<td>Converged workloads</td>
<td>Terabytes to single digit PB</td>
</tr>
<tr>
<td>Mixed I/O</td>
<td>Single digit PB to EB</td>
<td>Mixed I/O</td>
</tr>
<tr>
<td>Converged workloads</td>
<td>Converged workloads</td>
<td>Converged workloads</td>
</tr>
<tr>
<td>Single digit PB to EB</td>
<td>Single digit PB to EB</td>
<td>Single digit PB to EB</td>
</tr>
</tbody>
</table>

#### Base Rack Specs:
- 90 GB/sec (read/write)
- 7.5 PB usable capacity

#### Expansion Rack Specs:
- 120 GB/sec (read/write)
- 10 PB usable capacity

Mix and match Flash and Disk building blocks tailored to your workloads with Cray ClusterStor Data Services.

**HDD/Disk**

**SSD/Flash**

#### Base Rack Specs:
- 1,440 GB/sec (read)
- 1,080 GB/sec (write)

#### Expansion Rack Specs:
- 1,600 GB/sec (read)
- 1,200 GB/sec (write)
NEW CRAY CLUSTERSTOR DATA SERVICES

• Intelligent software to exploit the strengths of SSD and HDD without incurring their weaknesses
• Aligns the data flow with the workflow by promoting data to flash (and demoting it to HDD) at the right time

**Scripted Tiering (Day 1)**
- Use flash pool for temp data
- User can direct Lustre to move data up to the flash tier

**Scheduled Tiering (Mid 2020*)**
- User inserts WLM directives into the job script
- Automated migration is performed outside of run time

**Transparent Tiering (1H 2021*)**
- Read through, write back
- Less efficient, but no user action required/transparent for the user

*Current plan of record. Subject to change without notice
• Aligns the dataflow with the workflow - workload manager-driven
• Promotes the right data at the right time to fast Flash
• Challenge: Only available in Cray XC supercomputers

• Aligns the dataflow with the workflow - workload manager-driven
• Promotes the right data at the right time to fast Flash

• External parallel storage system
• Connects to any HPC compute of any vendor
• Challenge: No orchestration of the dataflow with the workflow

• Software intelligence in the external parallel storage system
• Connects to any HPC compute of any vendor

Cray DataWarp™

Cray ClusterStor L300

Cray ClusterStor Data Services for Cray ClusterStor E1000

*Since years in production at leadership sites like LANL, NERSC, KAUST etc.
A FLEXIBLE HPC STORAGE PLATFORM

- Delivered as soak tested, fully integrated storage system in Cray ClusterStor racks
- Or shipped soak tested “rack less” for installation in customer supplied racks

- 200 Gbps Cray Slingshot
- 100/200 GbE
- Infiniband EDR/HDR

Data flow: Via user command, via workload manager or policy-driven

I/O profile: Performs well with:
- large and small I/O
- sequential and random I/O
- write and read-intensive I/O

File system: Lustre today but architecturally open for other emerging file systems

Data path: LDISKFS or ZFS
START ANYWHERE, SCALE TO WHEREVER YOU NEED

All Flash entry point (6U)
Up to 80/60 GB/sec read/write and 115 TB capacity*

Disk entry point (14U):
30 GB/sec and 630 TB capacity*

Nearly 10 TB/sec and > 700 PB capacity*

> 4 TB/sec and 30PB (All Flash) capacity*

2.5 TB/sec and ~400 PB capacity*

Up to 1.3 TB/sec and ~200 PB capacity*

*usable capacity
HPE and Cray

HPC/AI compute for the exascale era

HPC/AI storage for the exascale era

Cray

ClusterStor

E1000

We design and implement complete end-to-end systems and support them L1 through L3 for:

- Software
- Compute
- Interconnect (if Cray Slingshot™)
- Storage (including Lustre file system)
CRAY CLUSTERSTOR E1000: NEW STORAGE FOR A NEW ERA

**Powerful**
- Highest file system performance in the industry for
  - All Flash
  - All Disk
  - Hybrid system

**Intelligent**
- Intelligent software that aligns the data flow with the workflow.
- Intelligent hardware design that extracts the maximum performance from each storage drive.

**Flexible**
- All Flash, All Disk and Hybrid configurations
- Modeling & Simulation and ML/DL and High performance data analytics
- Attaches to HPC/AI compute of any vendor

**Scalable**
- Start anywhere, scale to wherever you need.
- Selected by 100% of the U.S. DOE exascale sites.

**Easy**
- Easy to buy with HPE and Cray HPC/AI compute.
- Easy to implement as full integrated, factory-tested system.
- “One hand to shake” from procurement to customer support.
EXCLUSIVE OFFER FOR HP-CAST ATTENDEES

Deep dive storage briefing of your experts by our principle storage architects after the show in your premises

John Fragalla
Principal Storage Architect
jfragalla@cray.com
Greater Los Angeles Area, U.S.A

Torben Kling-Petersen
Principal Storage Architect
tpetersen@cray.com
Gothenburg, Sweden
HPC/AI STORAGE BLUEPRINT FOR THE NEW ERA

Cray/HPE HPC/AI Compute Systems

Cray ClusterStor E1000
Parallel storage for modeling & simulation, AI and high performance data analytics

HPE Data Management Framework
Tiered data management and data protection

Tape
On-premise or in Public Cloud

Object Storage
On-premise or in Public Cloud

*Storage solutions with alternative file systems available on request if needed.
STORE RIGHT, FUTURE BRIGHT!
jhowarth@cray.com
uplechschm@cray.com