ORNL OLCF Facilities Plans

Jack Wells
Director of Science
Oak Ridge Leadership Computing Facility
Oak Ridge National Laboratory

HEP-ASCR Requirements Workshop
Bethesda
10 June 2015
DOE’s Office of Science Computation User Facilities

- DOE is leader in open High-Performance Computing
- Provide the world’s most powerful computational tools for open science
- Access is free to researchers who publish
- Boost US competitiveness
- Attract the best and brightest researchers

NERSC
Edison is 2.57 PF

ALCF
Mira is 10 PF

OLCF
Titan is 27 PF
What is the Leadership Computing Facility (LCF)?

- Collaborative DOE Office of Science user-facility program at ORNL and ANL

- Mission: Provide the computational and data resources required to solve the most challenging problems.

- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community

- Highly competitive user allocation programs (INCITE, ALCC).

- Projects receive 10x to 100x more resource than at other generally available centers.

- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).
What is CORAL (Partnership for 2017 System)

• CORAL is a Collaboration of Oak Ridge, Argonne, and Lawrence Livermore Labs to acquire three systems for delivery in 2017.

• DOE’s Office of Science (DOE/SC) and National Nuclear Security Administration (NNSA) signed an MOU agreeing to collaborate on HPC research and acquisitions

• Collaboration grouping of DOE labs was done based on common acquisition timings. Collaboration is a win-win for all parties.
  – It reduces the number of RFPs vendors have to respond to
  – It improves the number and quality of proposals
  – It allows pooling of R&D funds
  – It strengthens the alliance between SC/NNSA on road to exascale
  – It encourages sharing technical expertise between Labs
Accelerating Future DOE Leadership Systems ("CORAL")

"Summit" System

"Sierra" System

5X – 10X Higher Application Performance

IBM POWER CPUs, NVIDIA Tesla GPUs, Mellanox EDR 100Gb/s InfiniBand

Paving The Road to Exascale Performance
2017 OLCF Leadership System
Hybrid CPU/GPU architecture

Vendor: IBM (Prime) / NVIDIA™ / Mellanox Technologies®

At least 5X Titan’s Application Performance

Approximately 3,400 nodes, each with:
• Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta architecture
• CPUs and GPUs completely connected with high speed NVLink
• Large coherent memory: over 512 GB (HBM + DDR4)
  – all directly addressable from the CPUs and GPUs
• An additional 800 GB of NVRAM, which can be configured as either a burst buffer or as extended memory
• over 40 TF peak performance

Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect

IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.
### How does Summit compare to Titan?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Summit</th>
<th>Titan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Performance</td>
<td>5-10x Titan</td>
<td>Baseline</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>~3,400</td>
<td>18,688</td>
</tr>
<tr>
<td>Node performance</td>
<td>&gt; 40 TF</td>
<td>1.4 TF</td>
</tr>
<tr>
<td>Memory per Node</td>
<td>&gt;512 GB (HBM + DDR4)</td>
<td>38GB (GDDR5+DDR3)</td>
</tr>
<tr>
<td>NVRAM per Node</td>
<td>800 GB</td>
<td>0</td>
</tr>
<tr>
<td>Node Interconnect</td>
<td>NVLink (5-12x PCIe 3)</td>
<td>PCIe 2</td>
</tr>
<tr>
<td>System Interconnect (node injection bandwidth)</td>
<td>Dual Rail EDR-IB (23 GB/s)</td>
<td>Gemini (6.4 GB/s)</td>
</tr>
<tr>
<td>Interconnect Topology</td>
<td>Non-blocking Fat Tree</td>
<td>3D Torus</td>
</tr>
<tr>
<td>Processors</td>
<td>IBM POWER9</td>
<td>AMD Opteron™</td>
</tr>
<tr>
<td></td>
<td>NVIDIA Volta™</td>
<td>NVIDIA Kepler™</td>
</tr>
<tr>
<td>File System</td>
<td>120 PB, 1 TB/s, GPFS™</td>
<td>32 PB, 1 TB/s, Lustre®</td>
</tr>
<tr>
<td>Peak power consumption</td>
<td>10 MW</td>
<td>9 MW</td>
</tr>
</tbody>
</table>
Two Tracks for Future Large Systems

Many Core
- 10’s of thousands of nodes with millions of cores
- Homogeneous cores
- Multiple levels of memory – on package, DDR, and non-volatile
- Unlike prior generations, future products are likely to be self hosted

Hybrid Multi-Core
- CPU / GPU Hybrid systems
- Likely to have multiple CPUs and GPUs per node
- Small number of very fat nodes
- Expect data movement issues to be much easier than previous systems – coherent shared memory within a node
- Multiple levels of memory – on package, DDR, and non-volatile

Cori at NERSC
- Self-hosted many-core system
- Intel/Cray
- 9300 single-socket nodes
- Intel® Xeon Phi™ Knights Landing (KNL)
- 16GB HBM, 64-128 GB DDR4
- Target delivery date: June, 2016

Summit at OLCF
- Hybrid CPU/GPU system
- IBM/NVIDIA
- 3400 multi-socket nodes
- POWER9/Volta
- More than 512 GB coherent memory per node
- Target delivery date: 2017

ALCF-3 at ALCF
- 3rd Generation Intel Xeon Phi (Knights Hill (KNH))
- > 50,000 compute nodes
- Target delivery date: 2018
### ASCR Computing Upgrades At a Glance

<table>
<thead>
<tr>
<th>System attributes</th>
<th>NERSC Now</th>
<th>OLCF Now</th>
<th>ALCF Now</th>
<th>NERSC Upgrade</th>
<th>OLCF Upgrade</th>
<th>ALCF Upgrade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Edison</td>
<td>TITAN</td>
<td>MIRA</td>
<td>Cori 2016</td>
<td>Summit 2017-2018</td>
<td>Aurora 2018-2019</td>
</tr>
<tr>
<td>Planned Installation</td>
<td>TITAN</td>
<td>TITAN</td>
<td>MIRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System peak (PF)</td>
<td>2.6</td>
<td>27</td>
<td>10</td>
<td>&gt; 30</td>
<td>150</td>
<td>180</td>
</tr>
<tr>
<td>Peak Power (MW)</td>
<td>2</td>
<td>9</td>
<td>4.8</td>
<td>&lt; 3.7</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Total system memory</td>
<td>357 TB</td>
<td>710TB</td>
<td>768TB</td>
<td>~1 PB DDR4 + High Bandwidth Memory (HBM) + 1.5PB persistent memory</td>
<td>&gt; 1.74 PB DDR4 + HBM + 2.8 PB persistent memory</td>
<td>&gt; 7 PB DRAM and persistent memory</td>
</tr>
<tr>
<td>Node performance (TF)</td>
<td>0.460</td>
<td>1.452</td>
<td>0.204</td>
<td>&gt; 3</td>
<td>&gt; 40</td>
<td>&gt; 15 times Mira</td>
</tr>
<tr>
<td>Node processors</td>
<td>Intel Ivy Bridge</td>
<td>AMD Opteron</td>
<td>Nvidia Kepler</td>
<td>Intel Knights Landing many core CPUs Intel Haswell CPU in data partition</td>
<td>Multiple IBM Power9 CPUs &amp; multiple Nvidia Voltas GPUs</td>
<td>Intel Knights Hill many core CPUs</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>5,600 nodes</td>
<td>18,688 nodes</td>
<td>49,152</td>
<td>9,300 nodes 1,900 nodes in data partition</td>
<td>~3,500 nodes</td>
<td>&gt;50,000 nodes</td>
</tr>
<tr>
<td>System Interconnect</td>
<td>Aries</td>
<td>Gemini</td>
<td>5D Torus</td>
<td>Aries</td>
<td>Dual Rail EDR-IB</td>
<td>Intel Omni-Path Architecture</td>
</tr>
<tr>
<td>File System</td>
<td>7.6 PB 168 GB/s, Lustre®</td>
<td>32 PB 1 TB/s, Lustre®</td>
<td>26 PB 300 GB/s, Lustre®</td>
<td>28 PB 744 GB/s, Lustre®</td>
<td>120 PB 1 TB/s, Lustre®</td>
<td>150 PB 1 TB/s, Lustre®</td>
</tr>
</tbody>
</table>
Center for Accelerated Application Readiness: Summit

OLCF-4 issued a call for proposals in FY2015 for application development partnerships between community developers, OLCF staff and the OLCF Vendor Center of Excellence.

Center for Accelerated Application Readiness (CAAR)
• Performance analysis of community applications
• Technical plan for code restructuring and optimization
• Deployment on OLCF-4
## New Application Readiness Activities CAAR

<table>
<thead>
<tr>
<th>Application</th>
<th>Domain</th>
<th>Principal Investigator</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME (N)</td>
<td>Climate Science</td>
<td>David Bader</td>
<td>Lawrence Livermore National Laboratory</td>
</tr>
<tr>
<td>DIRAC</td>
<td>Relativistic Chemistry</td>
<td>Lucas Visscher</td>
<td>Free University of Amsterdam</td>
</tr>
<tr>
<td>FLASH</td>
<td>Astrophysics</td>
<td>Bronson Messer</td>
<td>Oak Ridge National Laboratory</td>
</tr>
<tr>
<td>GTC (NE)</td>
<td>Plasma Physics</td>
<td>Zhihong Lin</td>
<td>University of California – Irvine</td>
</tr>
<tr>
<td>HACC(N)</td>
<td>Cosmology</td>
<td>Salman Habib</td>
<td>Argonne National Laboratory</td>
</tr>
<tr>
<td>LSDALTON</td>
<td>Chemistry</td>
<td>Poul Jørgensen</td>
<td>Aarhus University</td>
</tr>
<tr>
<td>NAMD (NE)</td>
<td>Biophysics</td>
<td>Klaus Schulten</td>
<td>University of Illinois – Urbana Champaign</td>
</tr>
<tr>
<td>NUCCOR</td>
<td>Nuclear Physics</td>
<td>Gaute Hagen</td>
<td>Oak Ridge National Laboratory</td>
</tr>
<tr>
<td>NWCHEM (N)</td>
<td>Chemistry</td>
<td>Karol Kowalski</td>
<td>Pacific Northwest National Laboratory</td>
</tr>
<tr>
<td>QMCPACK</td>
<td>Materials Science</td>
<td>Paul Kent</td>
<td>Oak Ridge National Laboratory</td>
</tr>
<tr>
<td>RAPTOR</td>
<td>Engineering</td>
<td>Joseph Oefelein</td>
<td>Sandia National Laboratory</td>
</tr>
<tr>
<td>SPECFEM</td>
<td>Seismic Science</td>
<td>Jeroen Tromp</td>
<td>Princeton University</td>
</tr>
<tr>
<td>XGC (N)</td>
<td>Plasma Physics</td>
<td>CS Chang</td>
<td>Princeton Plasma Physics Laboratory</td>
</tr>
</tbody>
</table>
# CAAR Timeline

<table>
<thead>
<tr>
<th>FY</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>TITAN</td>
<td>CAAR I</td>
<td>P8+</td>
<td>P9</td>
<td>PHASE I</td>
</tr>
<tr>
<td>Q2</td>
<td>CFP</td>
<td>CAAR II</td>
<td></td>
<td></td>
<td>ES</td>
</tr>
<tr>
<td>Q3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TRAINING</td>
</tr>
<tr>
<td>Q4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>POSTDOCS</td>
</tr>
</tbody>
</table>

- 1. November 2014: Call for CAAR applications
- 2. February 20, 2015: CAAR proposal deadline
- 3. March 2015: Selection of CAAR application teams
- 4. **April 2015: CAAR application training workshop**
- 5. April 2015: CAAR application teams start
- 6. June 2016: CAAR project review
- 7. October 2017: Call for Early Science projects
- 10. October 2018: Early Science project ends
CAAR in Preparation of Summit

Application Developer Team involvement
• Knowledge of the application
• Work on application in development “moving target”
• Optimizations included in application release

Early Science Project
• Demonstration of application on real problems at scale
• Shake-down on the new system hardware and software
• Large-scale science project is strong incentive to participate

Vendor technical support through the IBM/NVIDIA Center of Excellence is crucial
• Programming environment often not mature
• Best source of information on new hardware features

Access to multiple resources, including early hardware
Joint training activities
Portability is a critical concern
PanDA Tool Provides Titan with Next-Gen Workflow for Big Data

• Researchers with the ATLAS experiment in Europe have been integrating its scheduling and analysis tool, PanDa, with Titan.

• Global PanDA workflow includes 1.8 million jobs each day distributed among 100 or so computing centers spread across the globe.

• PanDA’s ability to efficiently match available computing time with high-priority tasks holds great promise for Titan.

• Team developers redesigned parts of the PanDA system on Titan responsible for job submission on remote sites and gave PanDA new capability to collect information about unused worker nodes on Titan.

• Deployment of the tool could lead to a higher utilization of available hours on Titan.
  – Three day test in July 2014 increased Titan utilization by 2.5%.

PanDA manages all of ATLAS’s data tasks from a server located at CERN, the European Organization for Nuclear Research.
PanDA architecture for Titan

- Pilot(s) executes on HPC interactive node
- Pilot interact with local job scheduler to manage job
- Data, produced on HPC automatically moves to external storage
OLCF & ESNet are implementing the Science DMZ to enable high-performance access to ESNet WAN.
A diverse ecosystem of compute and data infrastructure and services integrated with extreme scale systems.
ALICE-USA Project Plan:
New ALICE T2 facility at Oak Ridge National Laboratory

- ORNL CADES Facility:
  - Compute And Data Environment for Science

- LBNL NERSC & ORNL CADES
  - Both with Scientific Computing strength
  - High-bandwidth, monitored uplink to ESnet
  - Proximity to DOE HPC Resources with strategic alignment to $O^2$ project
    - Oak Ridge Leadership Computing Facility (OLCF)
    - NERSC DOE SC Flagship facility

- 2014 Project Proposal:
  - Establish new distributed ALICE T2 facility
    - Continue deployment & operation at LBNL NERSC
    - Establish new T2 site early 2015 at **ORNL CADES**
    - Transition operations from LLNL to ORNL in 2015

- Project Proposal Reviewed, June 2014
  - Proposal endorsed with modest recommendations