Logistics

- Final exam period is Friday, June 10, 10am-12pm
- No final exam
- Project presentations
  - Discuss what has been done on the project
  - Talk about design approach, development efforts, testing, …
- Graduate students term paper
  - Due 5pm, Friday, June 10
  - Give 10 minute synopsis of what you learned
Heterogeneous Processing Approaches

- There are 2 leading approaches for building heterogeneous processing architectures currently
  - Both based on manycore processors

- Accelerator approach
  - Dominated by GPGPUs (NVIDIA, AMD)
  - Mostly see GPUs on high-performance I/O bus
    - PCIexpress
  - DSPs also fit in this approach

- Coprocessor
  - Not a GPU, but a manycore device
  - Also see residing on high-performance I/O bus
  - Intel MIC architecture (Xeon Phi) is leading device
Tianhe-1A uses 7,000 NVIDIA GPUs (2010)

- Tianhe-1A uses
  - 7,168 NVIDIA Tesla M2050 GPUs
  - 14,336 Intel Westmeres

- Performance
  - 4.7 PF peak
  - 2.5 PF sustained on HPL

- 4.04 MW
  - If Tesla GPU’s were not used in the system, the whole machine could have needed 12 megawatts of energy to run with the same performance, which is equivalent to 5000 homes

- Custom fat-tree interconnect
  - 2x bandwidth of Infiniband QDR
#1: NUDT Tiahne-2 (Milkyway-2)

- **Compute Nodes** have 3.432 Tflop/s per node
  - 16,000 nodes
  - 32,000 Intel Xeon CPU
  - 48,000 Intel Xeon Phi

- **Operations Nodes**
  - 4,096 FT CPUs

- **Proprietary interconnect**
  - TH2 express

- **1PB memory**
  - Host memory only

- **Global shared parallel storage is 12.4 PB**

- **Cabinets**: 125+13+24 =162
  - Compute, communication, storage
  - ~750 m²
#2: ORNL Titan Hybrid System (Cray XK7)

- Peak performance of 27.1 PF
  - 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla “K20x” GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak power
Next-generation DOE Procurements

- DOE (Department of Energy)
- NERSC + ACES
  - Collaboration of 3 DOE national labs
  - Los Alamos, Sandia, NERSC
- CORAL
  - Collaboration of 3 DOE national labs
  - Oak Ridge, Argonne, Lawrence Livermore
  - Deliver a machine 5x times as powerful to each lab
NERSC + ACES

- National Energy Research Scientific Computing (NERSC) Center and the Alliance for Computing at Extreme Scale (ACES), a collaboration between Los Alamos National Laboratory and Sandia National Laboratory

- Two next generation systems:
  - Cori (NERSC-8) ([https://www.nersc.gov/users/computational-systems/cori/](https://www.nersc.gov/users/computational-systems/cori/))
  - Both to be delivered in the 2015/2016 time frame
  - Intention is to choose a single vendor to deliver two systems of similar technology
# LANL Trinity

<table>
<thead>
<tr>
<th>Trinity High-level Technical Specifications</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational Lifetime</td>
<td>2015 to 2020</td>
</tr>
<tr>
<td>Capability</td>
<td>8x to 12x improvement over Cielo in fidelity, physics, and performance capabilities</td>
</tr>
<tr>
<td>Architecture</td>
<td>Cray XC30</td>
</tr>
<tr>
<td>Memory capacity</td>
<td>&gt;2 PB of DDR4 DRAM</td>
</tr>
<tr>
<td>Peak performance</td>
<td>&gt;40 PF</td>
</tr>
<tr>
<td>Number of compute nodes</td>
<td>&gt;19,000</td>
</tr>
<tr>
<td>Processor architecture</td>
<td>Intel Haswell &amp; Knights Landing</td>
</tr>
<tr>
<td>Parallel file system capacity (usable)</td>
<td>&gt;80 PB</td>
</tr>
<tr>
<td>Parallel file system bandwidth (sustained)</td>
<td>1.45 TB/s</td>
</tr>
<tr>
<td>Burst buffer storage capacity (usable)</td>
<td>3.7 PB</td>
</tr>
<tr>
<td>Burst buffer bandwidth (sustained)</td>
<td>3.3 TB/s</td>
</tr>
<tr>
<td>Footprint</td>
<td>&lt;5,200 sq ft</td>
</tr>
<tr>
<td>Power requirement</td>
<td>&lt;10 MW</td>
</tr>
</tbody>
</table>
LLNL Sierra

Sierra System

Compute Rack
- Standard 19"
- Warm water cooling

Compute System
- 2.1 – 2.7 PB Memory
- 120 -150 PFLOPS
- 10 MW

Components

Compute Node
- POWER® Architecture Processor
- NVIDIA®Volta™
- NVMe-compatible PCIe 800GB SSD
- > 512 GB DDR4 + HBM
- Coherent Shared Memory

IBM POWER
- NVLink™

NVIDIA Volta
- HBM
- NVLink

Mellanox® Interconnect
- Dual-rail EDR Infiniband®
- 1.0 TB/s bandwidth

GPFS™ File System
- 120 PB usable storage
ORNL Summit

- Uses the same type of system as Sierra
- See movie
- https://www.olcf.ornl.gov/summit
- Delivery in 2018
Argonne Aurora (http://aurora.alcf.anl.gov)

AURORA: HOW IT COMPARES

Aurora's revolutionary architecture features Intel’s HPC scalable system framework and 2nd generation Intel® Omni-Path Fabric. The system will have a combined total of over 8 Petabytes of on-package high bandwidth memory and persistent memory, connected and communicating via a high-performance system fabric to achieve landmark throughput. The nodes will be linked to a dedicated burst buffer and a high-performance parallel storage solution.

A second system, named Theta, will be delivered in 2016. Theta will be based on Intel’s second-generation Xeon Phi processor and will serve as an early production system for the ALCF.

<table>
<thead>
<tr>
<th>System Features</th>
<th>Mira</th>
<th>Aurora</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Nodes</td>
<td>49,152</td>
<td>&gt;50,000</td>
</tr>
<tr>
<td>Processor</td>
<td>PowerPC A2 1600 MHz</td>
<td>3rd Generation Intel Xeon Phi</td>
</tr>
<tr>
<td>System Memory</td>
<td>768 TB</td>
<td>&gt;7 PB DRAM and persistent memory</td>
</tr>
<tr>
<td>System Interconnect</td>
<td>IBM 6D torus interconnect with VCSEL photonics</td>
<td>2nd Generation Intel Omni-Path Architecture with silicon photonics</td>
</tr>
<tr>
<td>File System Capacity</td>
<td>26 PB GFS</td>
<td>&gt;150 PB Lustre</td>
</tr>
<tr>
<td>Intel Architecture (x86-64) Compatibility</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Peak Power Consumption</td>
<td>4.8 MW</td>
<td>13 MW</td>
</tr>
</tbody>
</table>

- Delivery in 2019