NERSC’s KNL System: Cori

Richard Gerber
NERSC Senior Science Advisor
High Performance Computing Department Head
NERSC: the Mission HPC Facility for DOE Office of Science Research

Largest funder of physical science research in the U.S.

Bio Energy, Environment

Computing

Materials, Chemistry, Geophysics

Particle Physics, Astrophysics

Nuclear Physics

Fusion Energy, Plasma Physics

6,000 users, 700 projects, 700 codes, 48 states, 40 countries, universities & national labs
Allocation of Computing Time 2017

- **DOE Mission Science 80%**
  Distributed by DOE Office of Science program managers

- **ALCC 10%**
  Competitive awards run by DOE Advanced Scientific Computing Research Office

- **Directors Discretionary 10%**
  Strategic awards from NERSC

Pre-production time available on KNL nodes
Through June 30, 2017
Email Richard Gerber and/or Jack Deslippe
Initial Allocation Distribution Among Offices for 2016

2016 NERSC Allocations by Reserve (Millions)

- ASCR Specials, 11.6, 0.5%
- SBIR, 20.4, 0.9%
- ASCR, 142, 5.9%
- BER/Climate, 365, 15.2%
- BER/Bio, 90, 3.8%
- HEP, 366, 15.3%
- BES/MatSci, 299, 12.5%
- FES, 429, 17.9%
- BES/ChemSci, 301, 12.5%
- BES/Geo, 49, 2.0%
- BES/Facility, 63, 2.6%
<table>
<thead>
<tr>
<th>Cori</th>
<th>2,000 Haswell nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cray XC40 system</td>
<td>Dual-socket 16 core @ 2.3 GHz</td>
</tr>
<tr>
<td>9,300 Intel Xeon Phi (KNL 7250) @ 1.4 GHz</td>
<td>128 GB DDR4 @ 2133 MHz</td>
</tr>
<tr>
<td>Single socket, self-hosted nodes</td>
<td>Cray Aries 3-level dragonfly network connects KNL and</td>
</tr>
<tr>
<td>68-core KNL, each with 4 HW threads</td>
<td>Haswell nodes</td>
</tr>
<tr>
<td>16 GB MCDRAM, 450 GB/s BW</td>
<td>NVRAM Burst Buffer 1.8 PB, 1.5 TB/sec</td>
</tr>
<tr>
<td>96 GB DDR4 @ 2400 MHz</td>
<td>30 PB Lustre scratch, &gt;700 GB/sec I/O</td>
</tr>
</tbody>
</table>

#5 on Top 500 November 2016, 31 PF peak
Knights Landing Overview

**Chip:** 36 Tiles interconnected by 2D Mesh
**Tile:** 2 Cores + 2 VPU/core + 1 MB L2

**Memory:** MCDRAM: 16 GB on-package; High BW DDR4: 6 channels @ 2400 up to 384GB
**IO:** 36 lanes PCIe Gen3, 4 lanes of DMI for chipset
**Node:** 1-Socket only
**Fabric:** Omni-Path on-package (not shown)

**Vector Peak Perf:** 3+TF DP and 6+TF SP Flops
**Scalar Perf:** ~3x over Knights Corner
**Streams Triad (GB/s):** MCDRAM: 400+; DDR: 90+

Source: Intel. All products, computer systems, dates and figures specified are preliminary based on current expectations and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. BENCHMARKS: The Intel Xeon processors product benchmarks represent performance results measured by Intel. The benchmark results are intended for comparison purposes only. Any references to system performance are provided as illustrative examples. Specific platform configurations or software applications may achieve different performance results.
Cray XC40 KNL Blade

Aries
~10 GB/s sustained 1 direction
0.3 – 2.3 μs latency
NERSC Exascale Scientific Application Program (NESAP)

Goal: Prepare DOE Office of Science users for Cori’s manycore CPUs
Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community

NESAP activities include:

- Close interactions with vendors
- Developer Workshops
- Early engagement with code teams
- Postdoc Program
- Engage in the Broad Community
- Training and online modules
- Early access to KNL
NERSC at a Glance

A U.S. Department of Energy Office of Science User Facility
Provides High Performance Computing and Data Systems and Services
Unclassified Basic and Applied Research in Energy-Related Fields
6,000 users, 700 different scientific projects
Located at Lawrence Berkeley National Lab, Berkeley, CA
Permanent Staff of about 70
Production High Performance Computing Systems

Cori
9,300 Intel Xeon Phi “KNL” manycore nodes
2,000 Intel Xeon “Haswell” nodes
700,000 processor cores, 1.2 PB memory
Cray XC40 / Aries Dragonfly interconnect
30 PB Lustre Cray Sonexion scratch FS
1.5 PB Burst Buffer

#5 on list of Top 500 supercomputers in the world

Edison
5,560 Ivy Bridge Nodes / 24 cores/node
133 K cores, 64 GB memory/node
Cray XC30 / Aries Dragonfly interconnect
6 PB Lustre Cray Sonexion scratch FS
Thread-Level Parallelism for Xeon Phi Manycore

Xeon Phi “Knights Landing”

68 Cores with 1-4 threads

Commonly using OpenMP to express threaded parallelism
On-Chip Parallelism – Vectorization (SIMD)

- Single instruction to execute up to 16 DP floating point operations per cycle per VPU.
- 32 Flop / cycle / core
- 44 Gflops / core
- 3 TFlops / node

![Diagram showing comparison between scalar and SIMD processing]
Let the hardware automatically manage the integrated on-package memory as an “L3” cache between KNL CPU and external DDR.

Manually manage how your application uses the integrated on-package memory and external DDR for peak performance.

Harness the benefits of both cache and flat models by segmenting the integrated on-package memory.

Maximum performance through higher memory bandwidth and flexibility.
Data layout crucial for performance

Enables efficient vectorization

Cache “blocking”

Fit important data structures in 16 GB of MCDRAM

MCDRAM memory/core = 235 MB

DDR4 memory/core = 1.4 GB