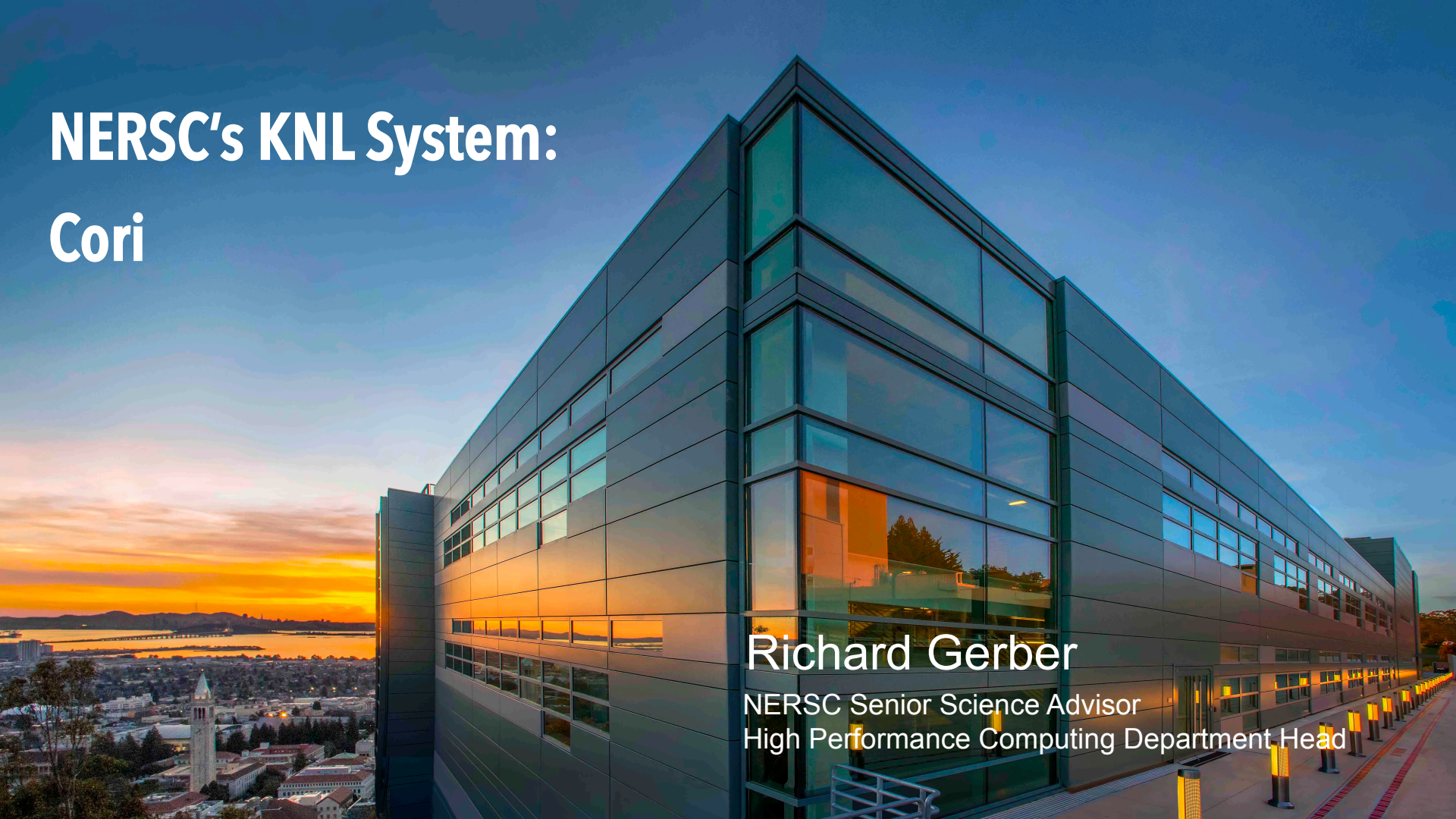
NERSC's KNL System: Cori

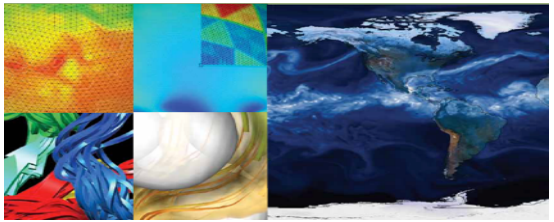


Richard Gerber
NERSC Senior Science Advisor
High Performance Computing Department Head

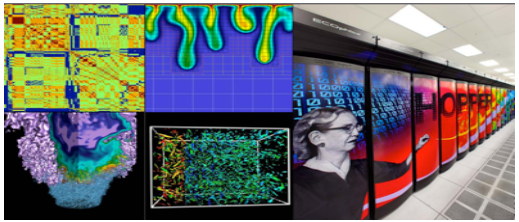
NERSC: the Mission HPC Facility for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

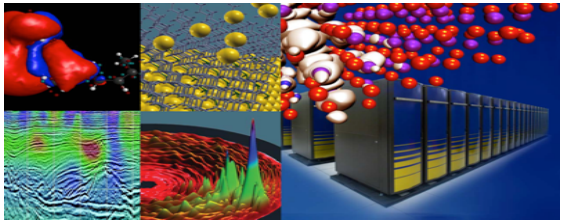
Largest funder of physical science research in the U.S.



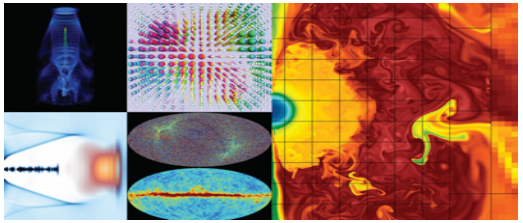
Bio Energy, Environment



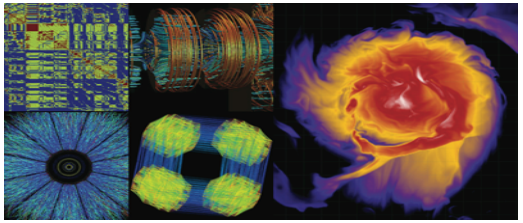
Computing



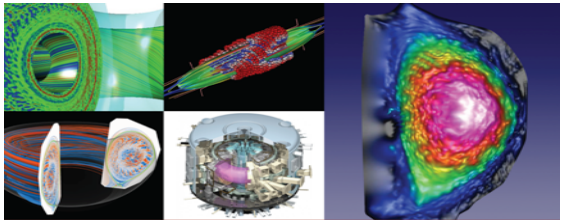
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics

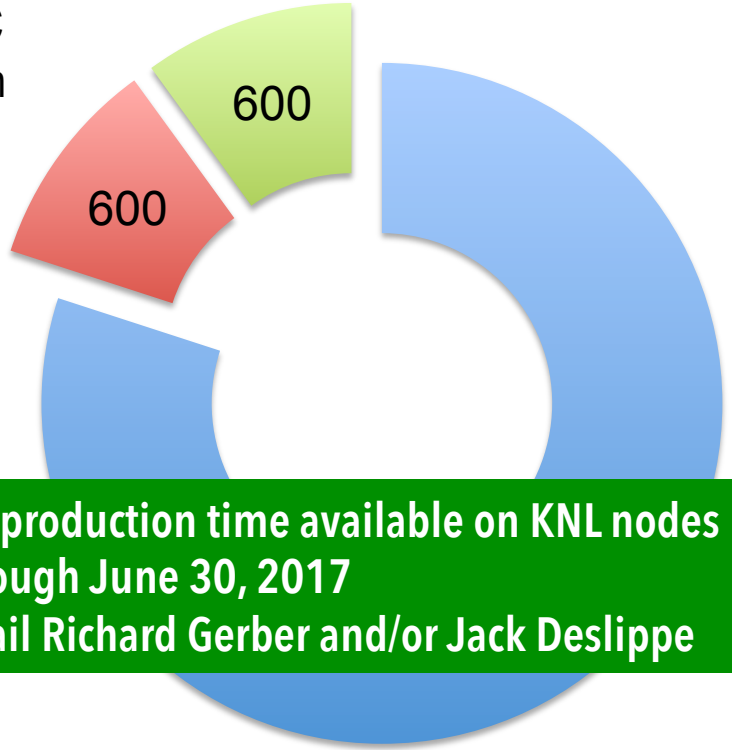


Fusion Energy, Plasma Physics

6,000 users, 700 projects, 700 codes, 48 states, 40 countries, universities & national labs

Allocation of Computing Time 2017

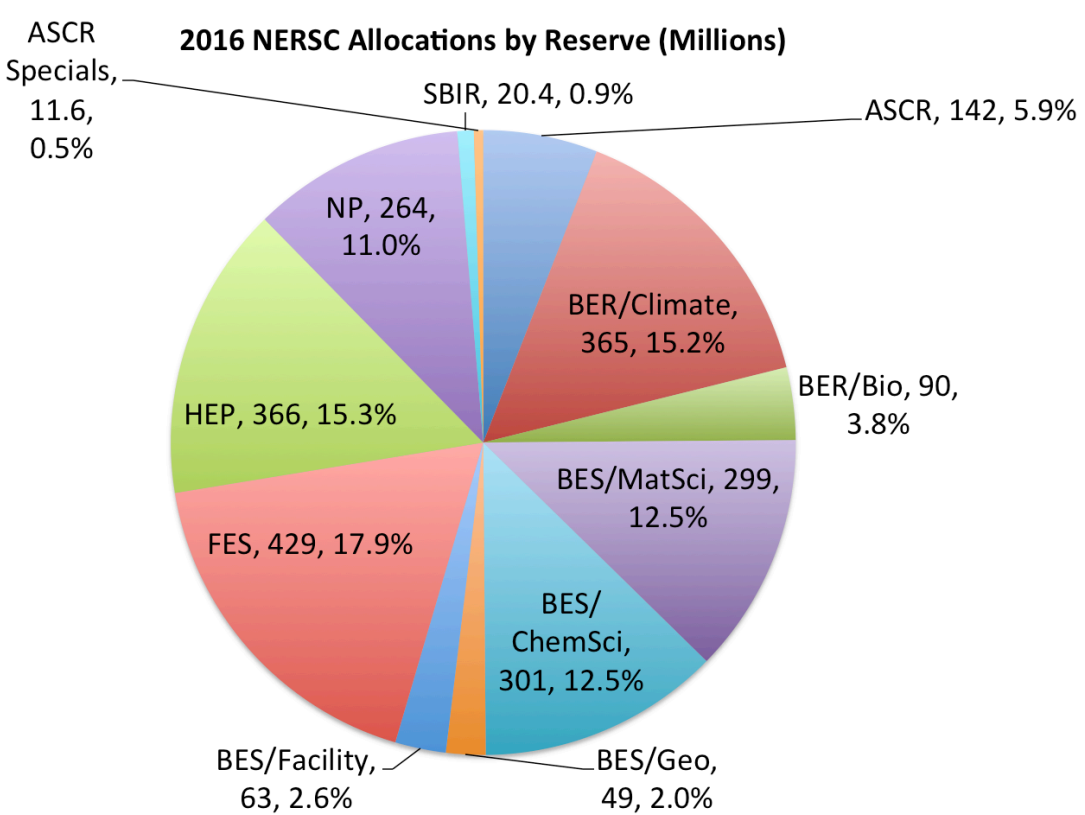
NERSC hours in millions



Pre-production time available on KNL nodes
Through June 30, 2017
Email Richard Gerber and/or Jack Deslippe

- 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

Initial Allocation Distribution Among Offices for 2016



Cori

Cray XC40 system

9,300 Intel Xeon Phi (KNL 7250) @ 1.4 GHz

Single socket, self-hosted nodes

68-core KNL, each with 4 HW threads

16 GB MCDRAM, 450 GB/s BW

96 GB DDR4 @ 2400 MHz

2,000 Haswell nodes

Dual-socket 16 core @ 2.3 GHz

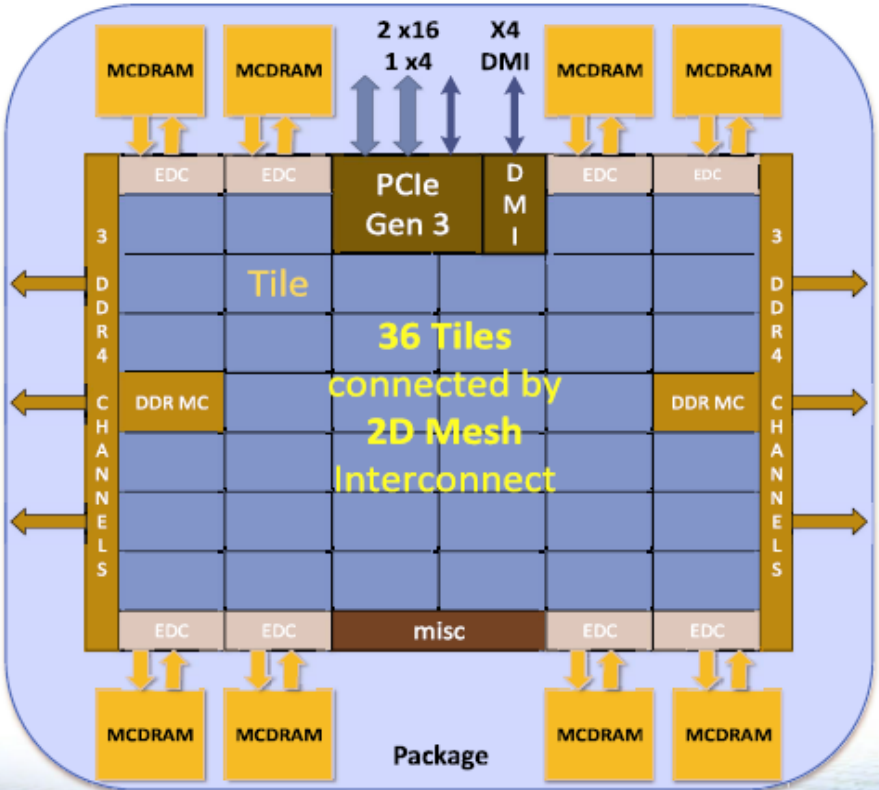
128 GB DDR4 @ 2133 MHz

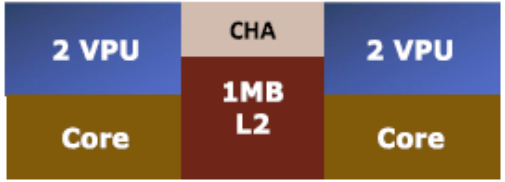
Cray Aries 3-level dragonfly network connects KNL and Haswell nodes

NVRAM Burst Buffer 1.8 PB, 1.5 TB/sec

30 PB Lustre scratch, >700 GB/sec I/O

Knights Landing Overview





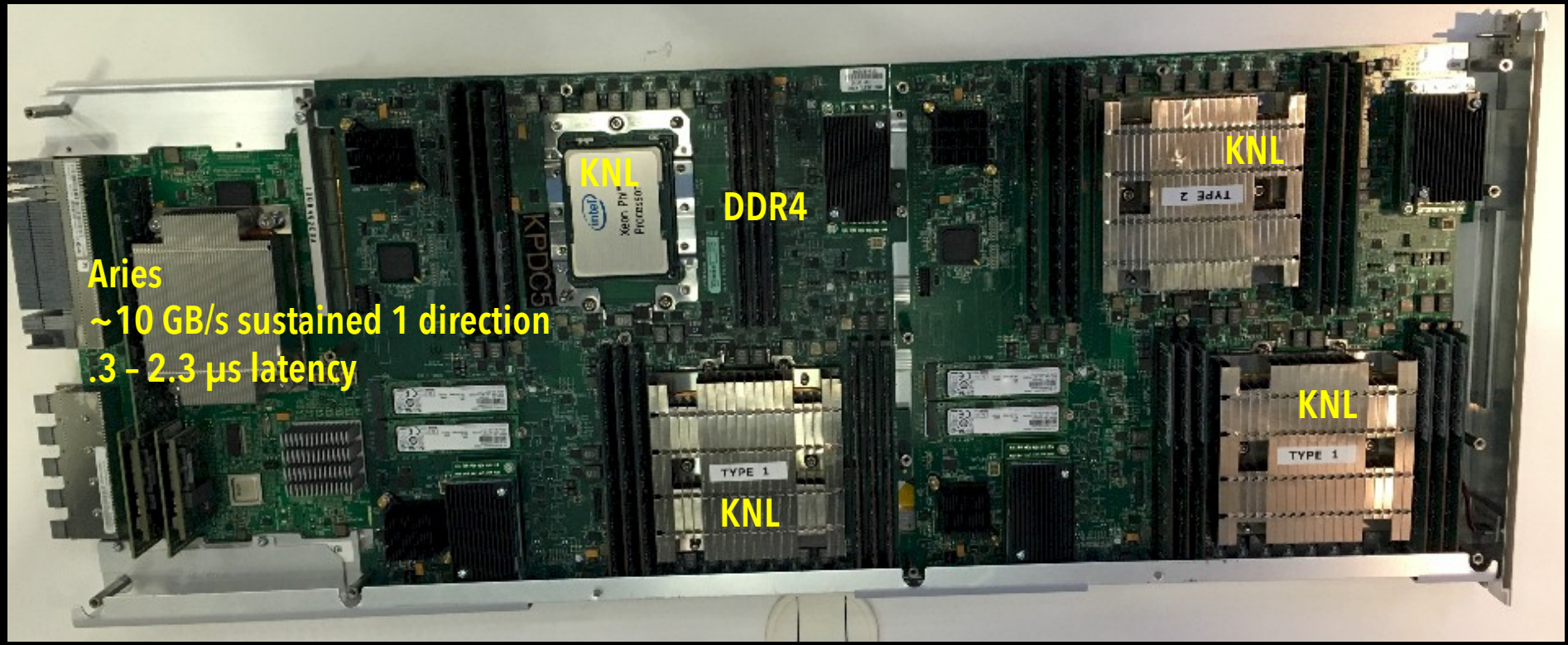
Chip: 36 Tiles interconnected by **2D Mesh**
Tile: 2 Cores + 2 VPU/core + 1 MB L2

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. †Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). ††Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Cray XC40 KNL Blade

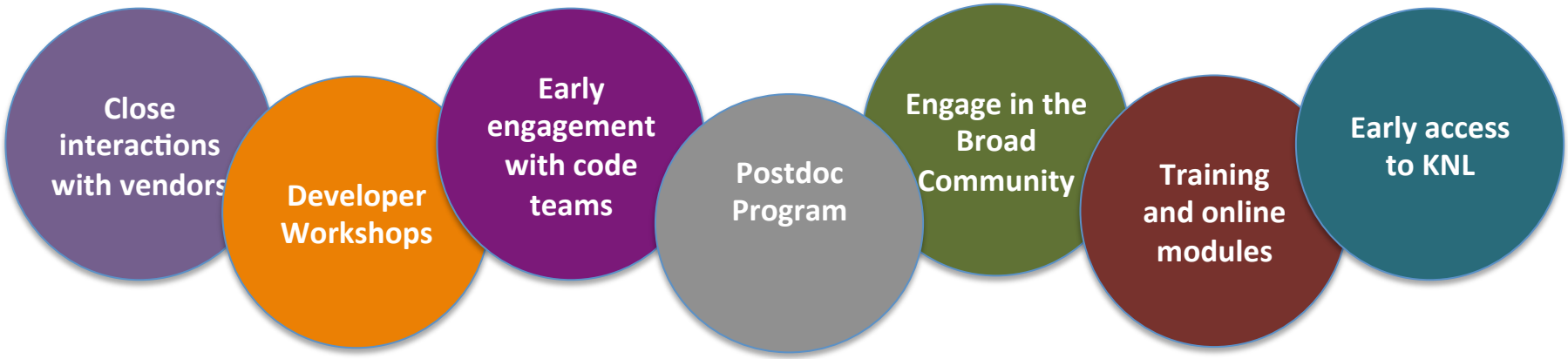


NERSC Exascale Scientific Application Program (NESAP)

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:



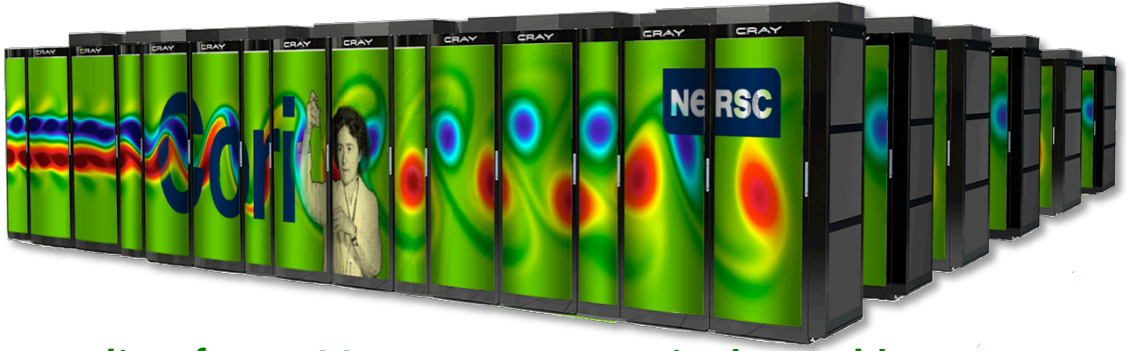
NERSC at a Glance

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

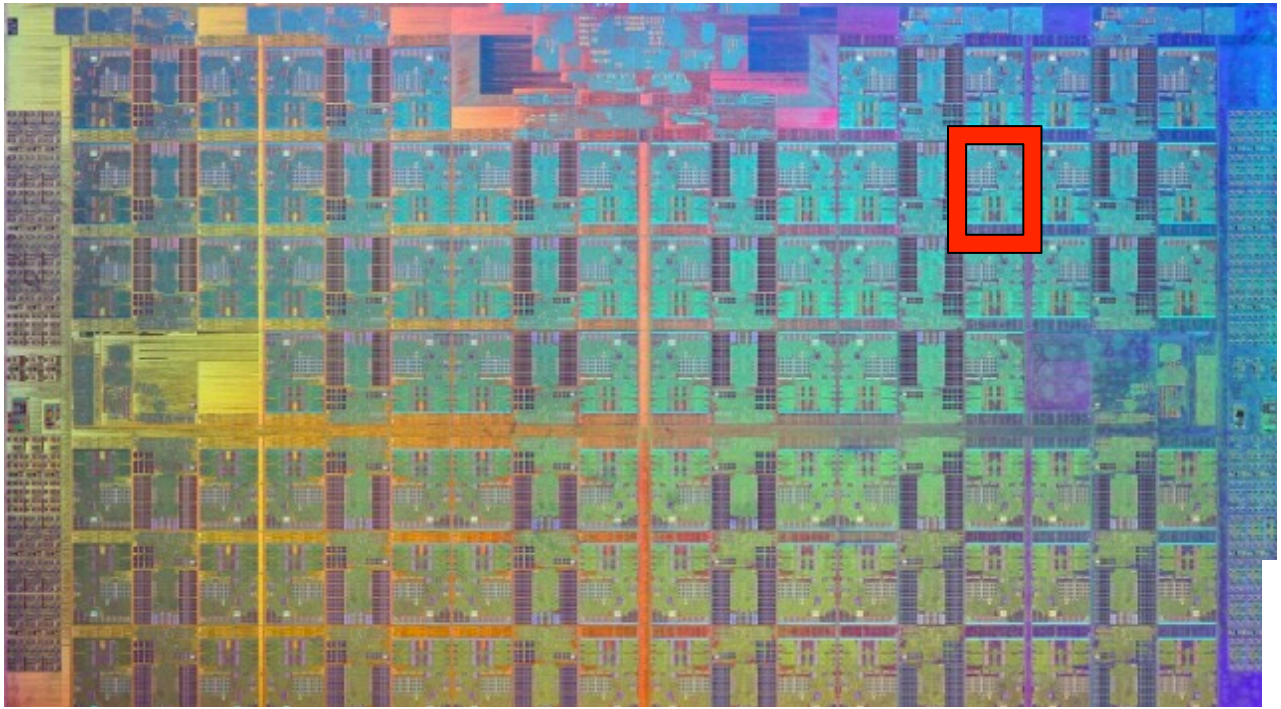


#5 on list of Top 500 supercomputers in the world

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

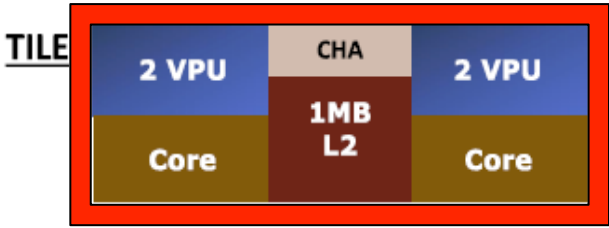
Thread-Level Parallelism for Xeon Phi Manycore



Xeon Phi "Knights Landing"

68 Cores with 1-4 threads

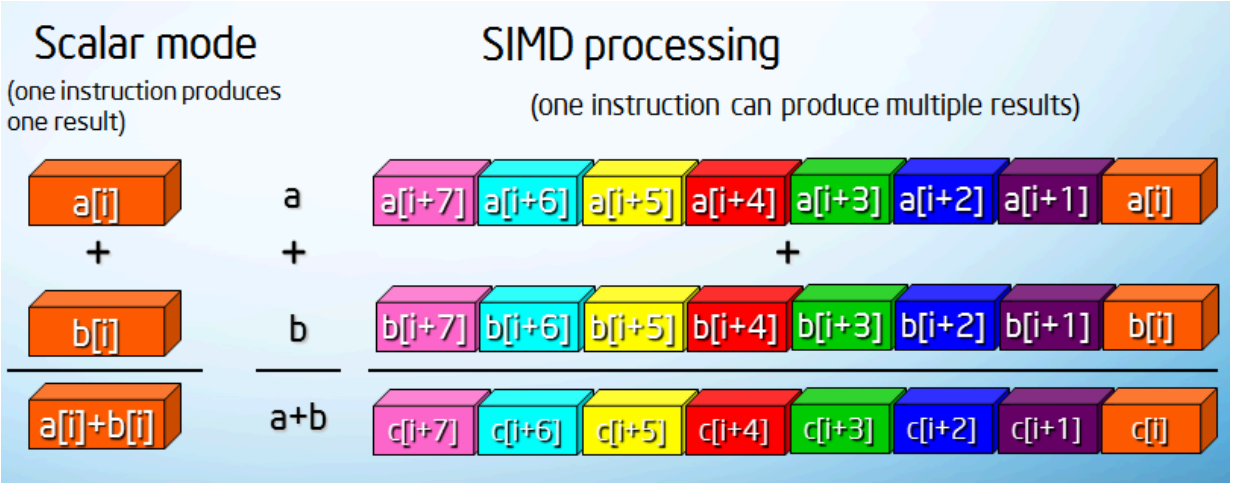
Commonly using OpenMP to express threaded parallelism



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

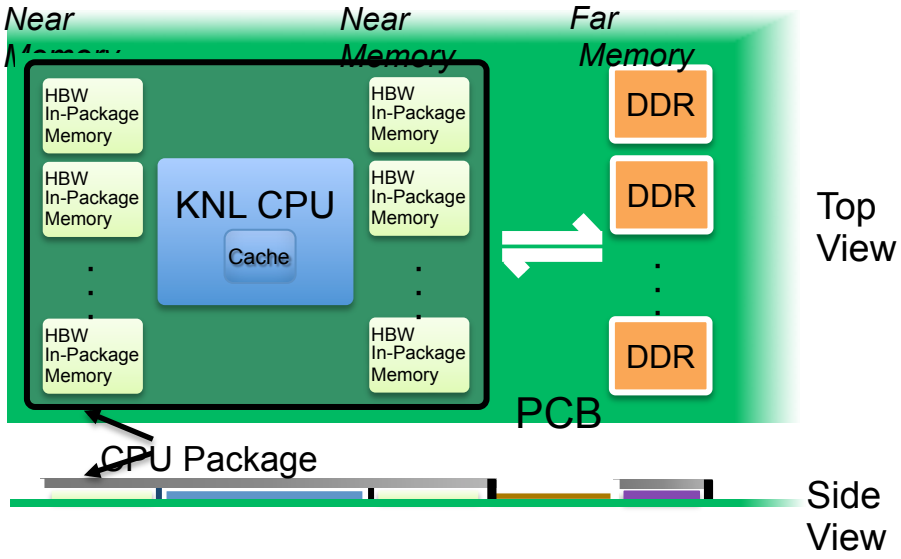


Knights Landing Integrated On-Package Memory

Cache Model — Let the hardware automatically manage the integrated on-package memory as an "L3" cache between KNL CPU and external DDR

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

Hybrid Model — 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