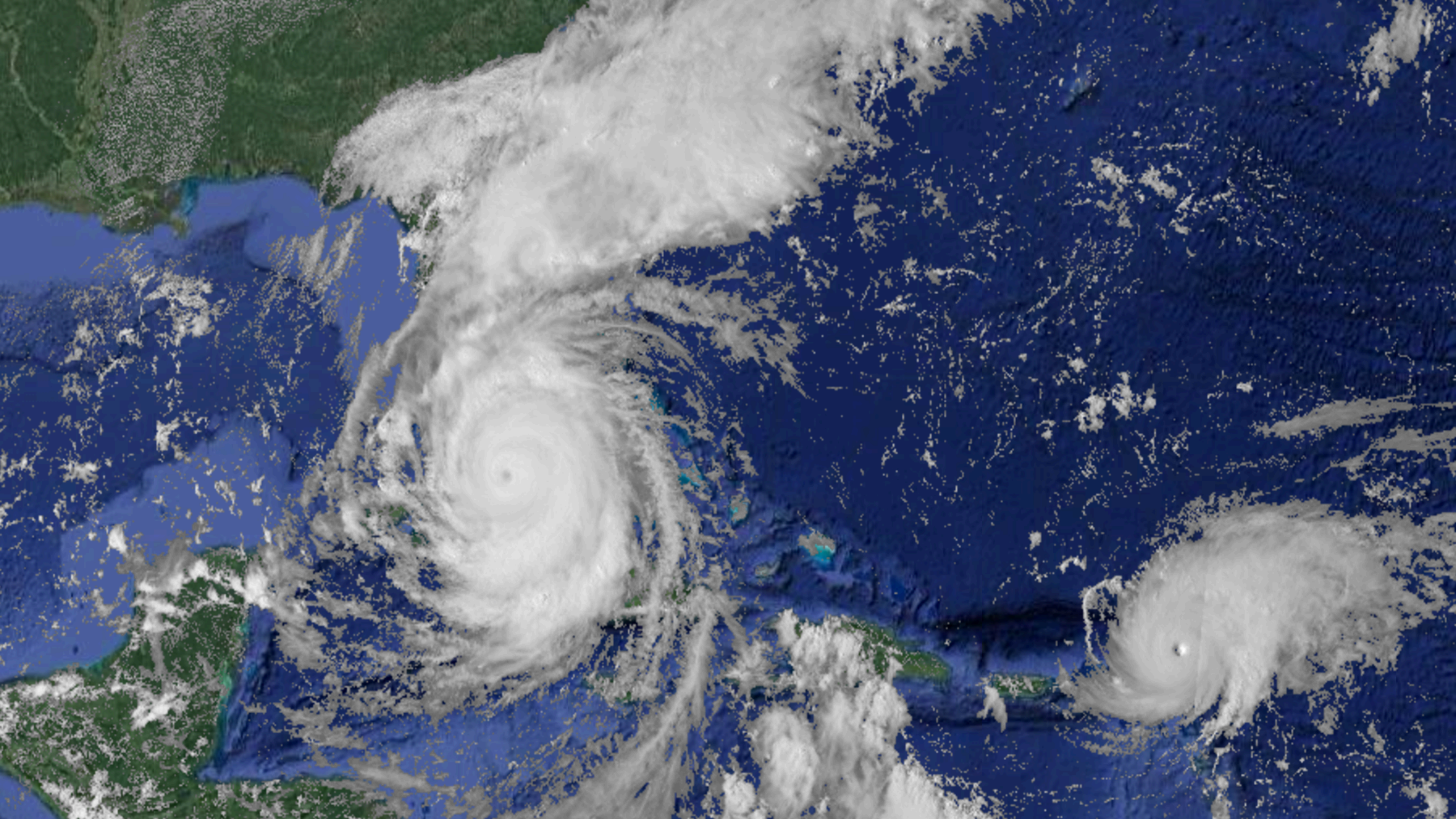
Preparing the NERSC Community for Next-Generation HPC Architectures

Richard Gerber

NERSC Senior Science Advisor
High Performance Computing Department Head

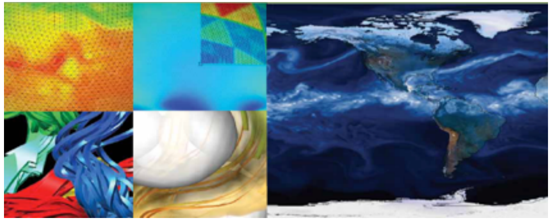


NERSC: the Mission HPC Facility for DOE Office of Science Research

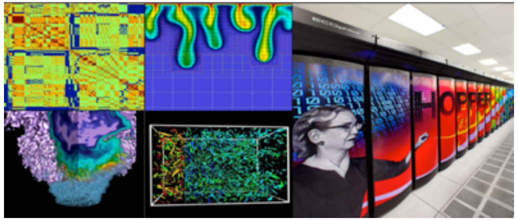
NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

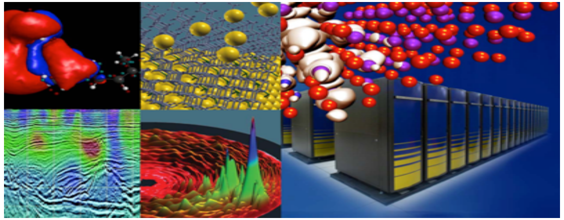
Largest funder of physical science research in the U.S.



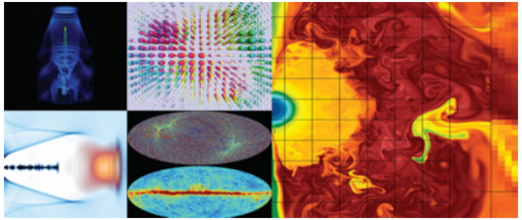
Bio Energy, Environment



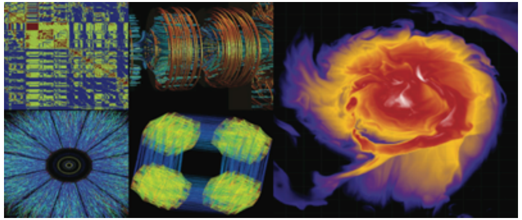
Computing



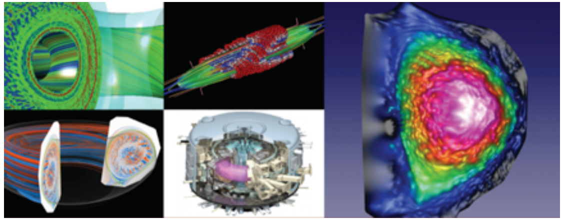
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



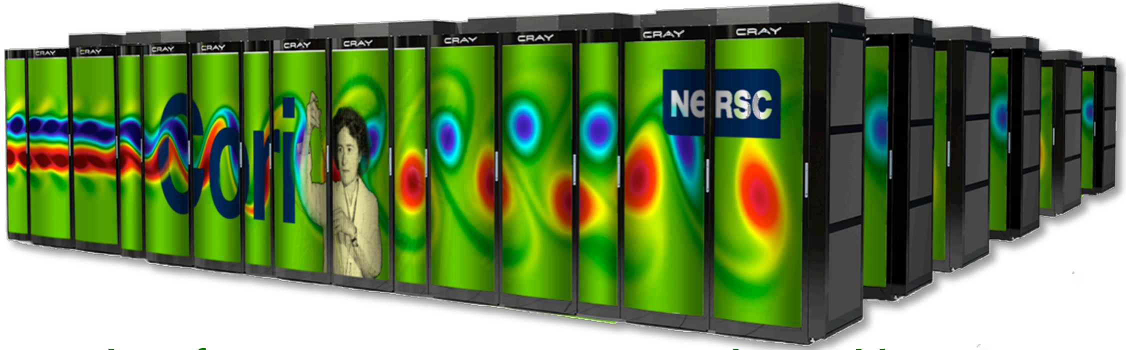
Fusion Energy, Plasma Physics

7,000 users, 750 projects, 750 codes, 48 states, 40 countries, universities & national labs

Production High Performance Computing Systems

Cori

9,600 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



#6 on list of Top 500 supercomputers in the world

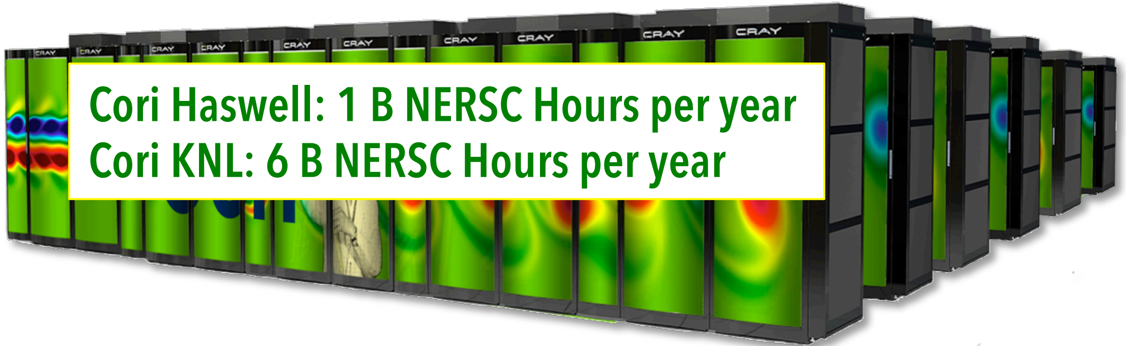
Edison

5,560 Intel Xeon "Ivy Bridge" Nodes
133 K cores, 357 TB memory
Cray XC30 / Aries Dragonfly interconnect
6 PB Lustre Cray Sonexion scratch FS

Production High Performance Computing Systems

Cori

9,600 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

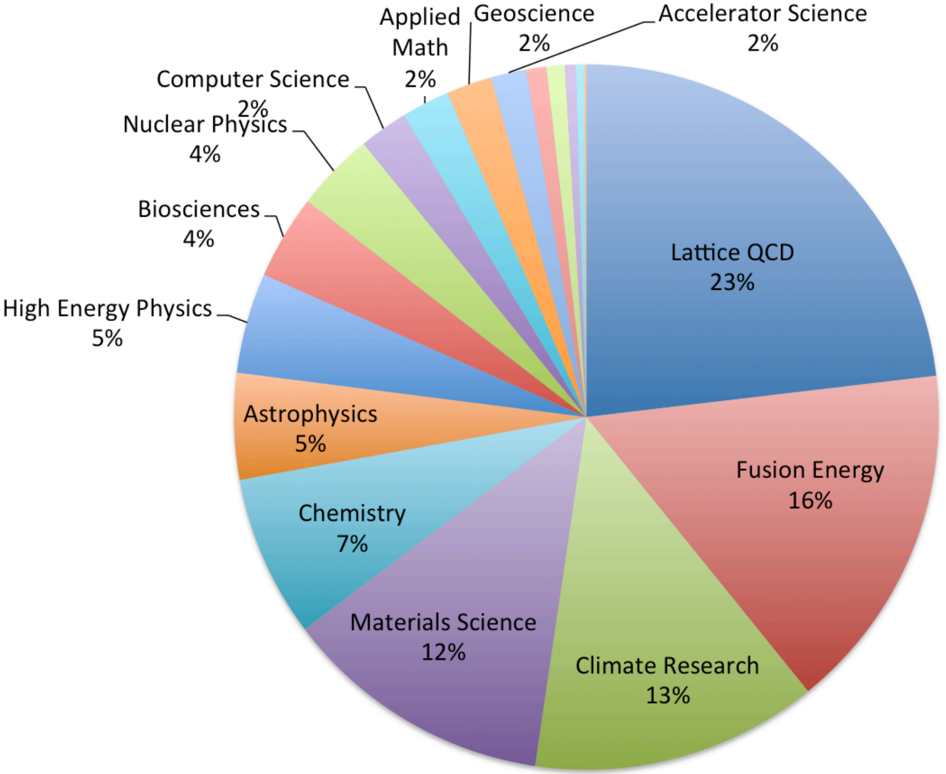


#6 on list of Top 500 supercomputers in the world

Edison

5,560 Intel Xeon "Ivy Bridge" Nodes
133 K cores, 357 TB memory
Cray XC30 / Aries Dragonfly interconnect
6 PB Lustre Cray Sonexion scratch FS

NERSC Usage Demographics 2016



84 Climate/Env Projects

With users from 127 organizations
715 different users

NCAR 52
Berkeley Lab 93
Livermore Lab 42
Los Alamos Lab 32
PNNL 95
UC Berkeley 34

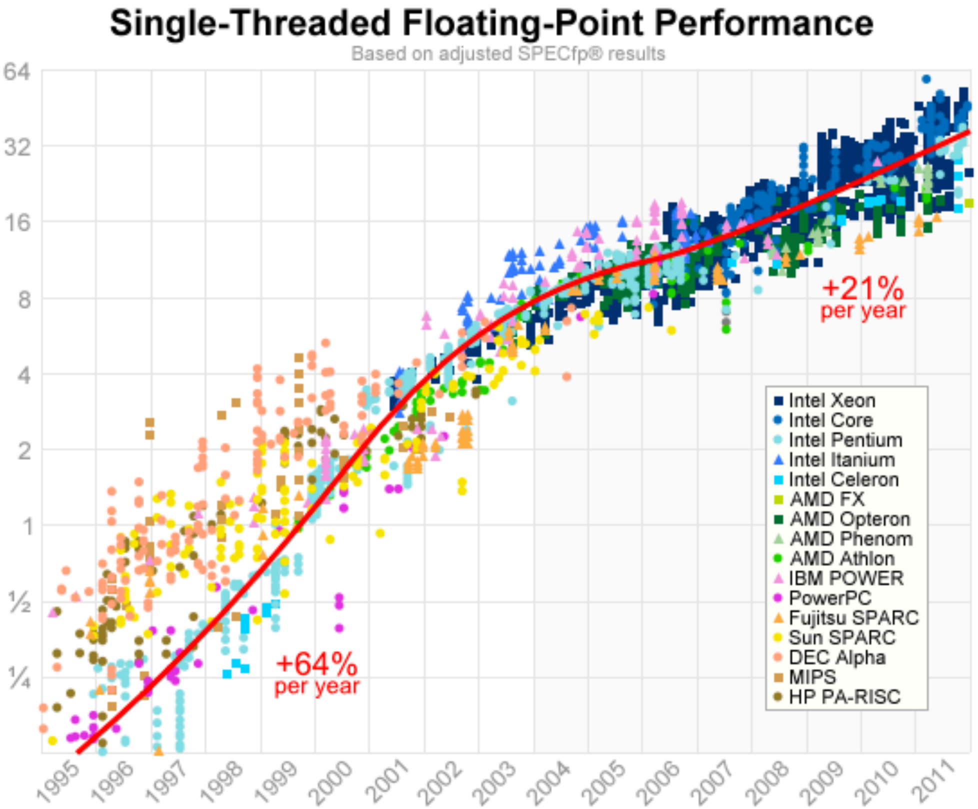
Code rank 2016
CESM #2
WRF #21

NERSC builds & installs CESM

“The only thing constant is change”

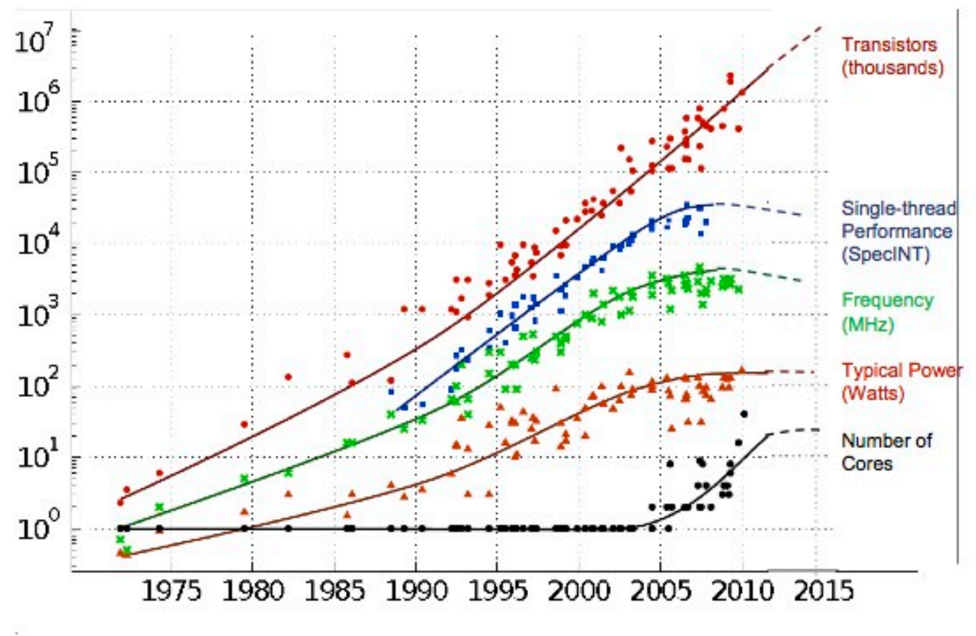
–Heraclitus of Ephesus

Single Processor Performance

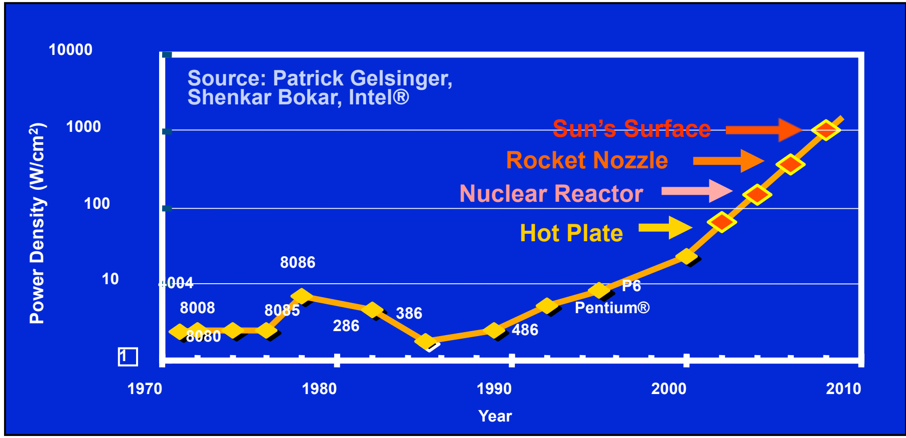


Every year there was a new CPU technology that enabled single-thread performance to increase

Change was coming and we kept telling our 7,000 users it was so …



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

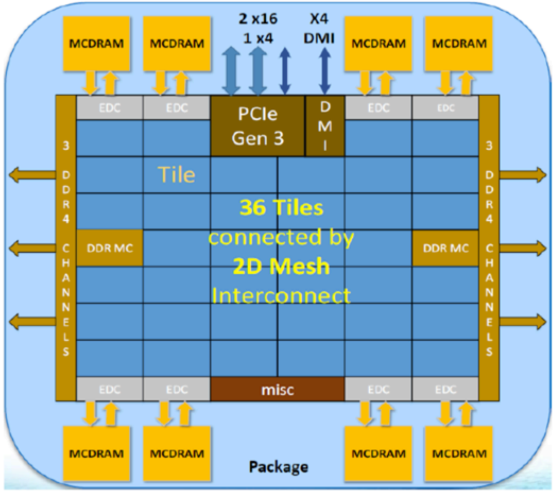


Driven by power consumption and dissipation toward lightweight cores

TECHENABLEMENT
Get known for the tech we teach

NERSC to Procure "Cori" a Knights Landing Based Cray XC Supercomputer

May 2, 2014 by Rob Farber — Leave a Comment



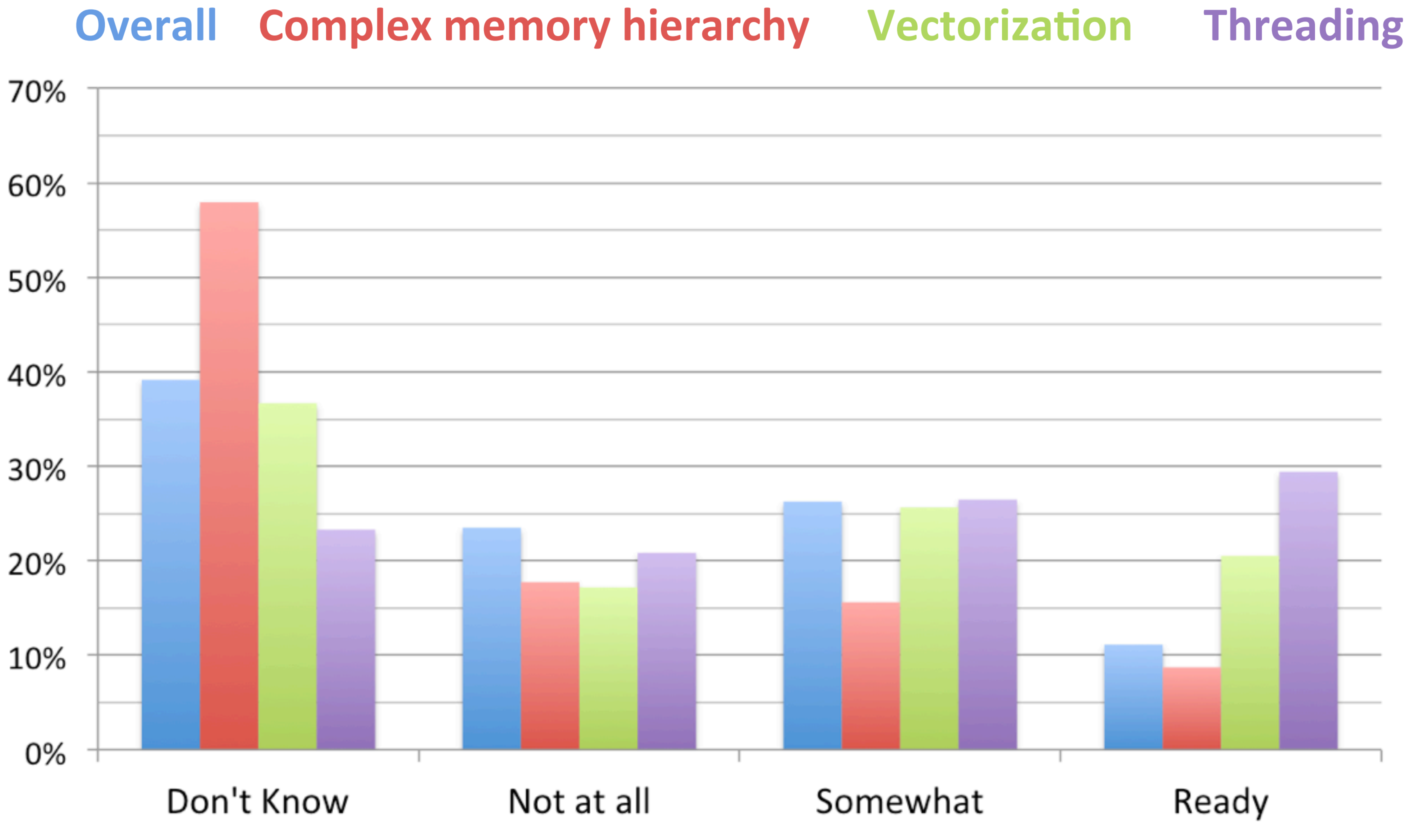
30 PFlop System will be a boon to science because of new capabilities, but the Intel Xeon Phi many-core architecture will require a code modernization effort to use efficiently.

KNL: 215-230 W
2-socket Haswell: 270 W

For the first time, NERSC's users will have lower single-thread performance in their next system.

68-272 threads
16 GB MCDRAM
2 x 512b vectors
2 x FMA / core

User Survey: Is Your Code Ready for Manycore?



We don't choose our users or codes. We support all DOE mission science.

Manycore is the future of HPC

Time to transition community

On the path to exascale

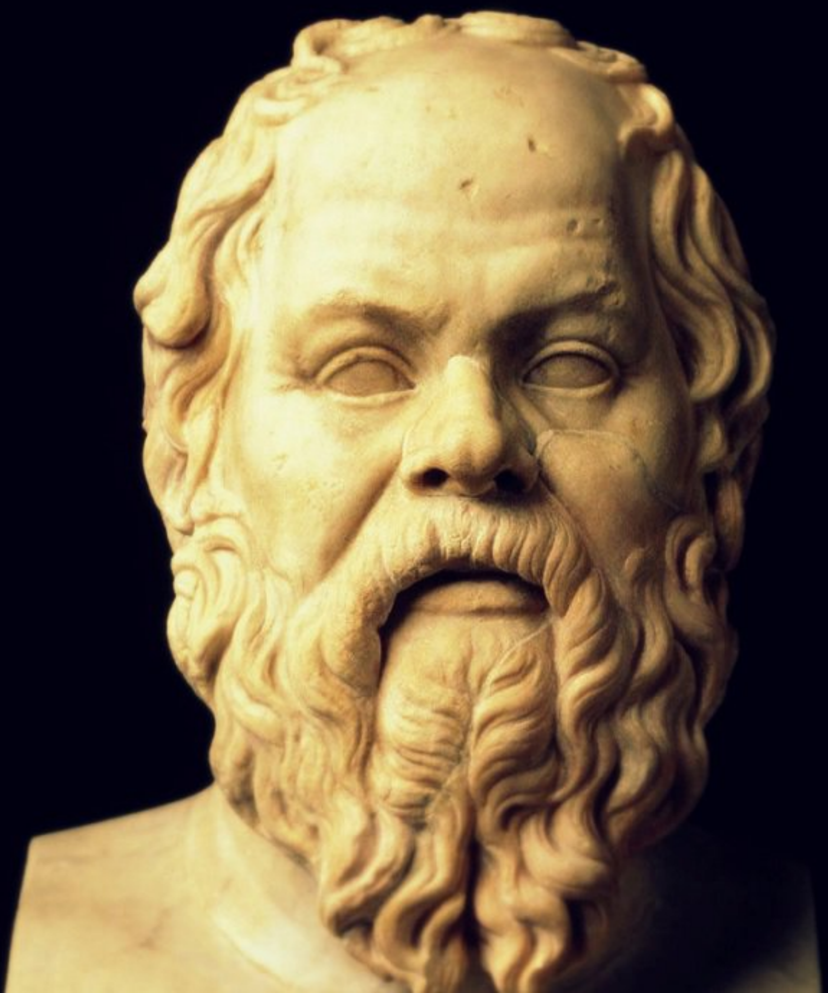
Homogeneous, x86-compatible CPU as a first step – not an accelerator

High bandwidth memory big win for many NERSC codes

NERSC's Challenge

How can NERSC's diverse community of 7,000 users, 750 projects, and 700 codes use Cori's Intel Xeon Phi Knights Landing processors at high performance

Business as usual was over



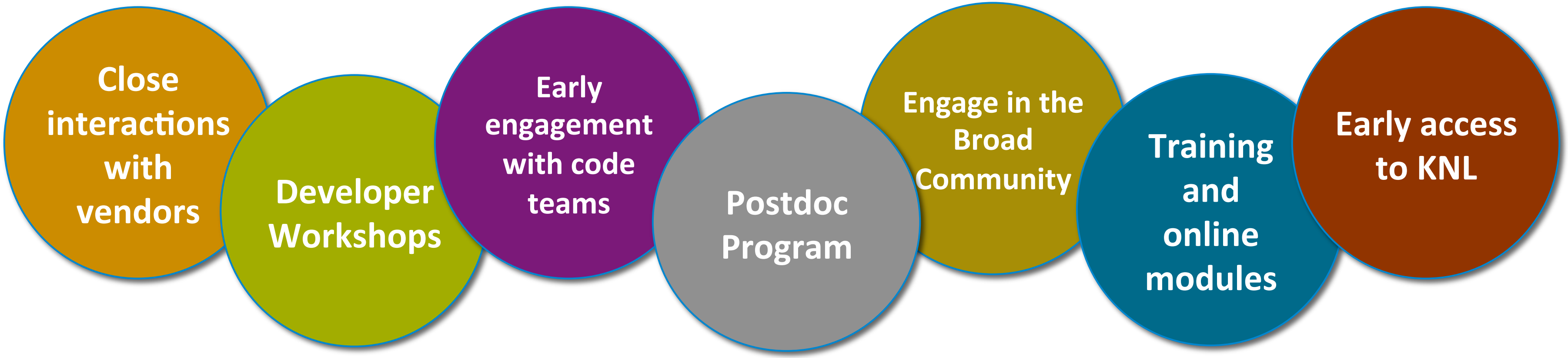
“The secret of change is to focus all of your energy, not on fighting the old, but on building the new.”

– Socrates

NERSC Exascale Scientific Application Program (NESAP)

Goal: Prepare Office of Science users for Cori's manycore CPUs

Partner with ~20 application teams and apply lessons learned to broad user community – accounts for ~ 50% of hours used



NESAP Call for Proposals – Early 2014

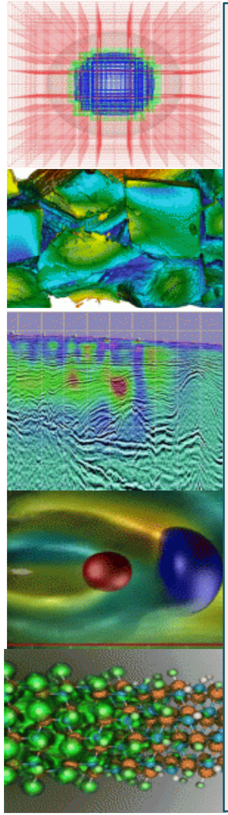
Selected projects must

- Work with NESAP liaison to produce profiling and scaling plots and vectorization and memory BW analyses.
- Commit 0.5-1.0 FTE to work on optimizing, refactoring, testing, and further profiling.
- Intermediate and final reports detailing the application's science and performance improvement as a result of the collaboration.

Evaluation criteria

- Importance to Office of Science research
- Representation all 6 OS programs
- Science potential
- Ability for code development and optimizations to be transferred to the broader community through libraries, algorithms, kernels or community codes
- Match NERSC/Vendor resources and expertise

NESAP Codes



Advanced Scientific Computing Research

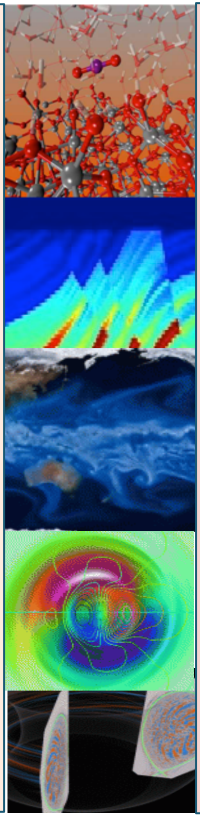
Almgren (LBNL) **BoxLib AMR**
Trebotich (LBNL) **Chombo-crunch**

High Energy Physics

Vay (LBNL) **WARP & IMPACT**
Toussaint(Arizona) **MILC**
Habib (ANL) **HACC**

Nuclear Physics

Maris (Iowa St.) **MFDn**
Joo (JLAB) **Chroma**
Christ/Karsch
(Columbia/BNL) **DWF/HISQ**



Basic Energy Sciences

Kent (ORNL) **Quantum Espresso**
Deslippe (NERSC) **BerkeleyGW**
Chelikowsky (UT) **PARSEC**
Bylaska (PNNL) **NWChem**
Newman (LBNL) **EMGeo**

Biological and Env Research

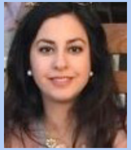
Smith (ORNL) **Gromacs**
Yelick (LBNL) **Meraculous**
Ringler (LANL) **MPAS-O**
Johansen (LBNL) **ACME**
Dennis (NCAR) **CESM**

Fusion Energy Sciences

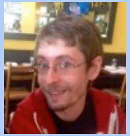
Jardin (PPPL) **M3D**
Chang (PPPL) **XGC1**

New Postdoc Program


Taylor Barnes
Quantum ESPRESSO


Zahra Ronaghi
Tomopy


Andrey Ovsyannikov
Chombo-Crunch


Bill Arndt
HIPMER/ HMMER/MPAS


Rahul Gayatri
SW4


Tuomas Koskela
XGC1


Kevin Gott
PARSEC

One Open Spot

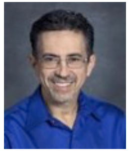
NERSC Application Performance Group formed

New hire: Charlene Yang (Pawsey)

NESAP Staff Contributors


Katie Antypas

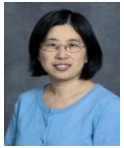

Jack Deslippe


Richard Gerber

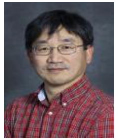

Nick Wright


Brandon Cook


Thorsten Kurth


Helen He


Stephen Leak


Woo-Sun Yang


Rebecca Hartman-Baker


Doug Doerfler


Zhengji Zhao


Brian Austin


Rollin Thomas


Brian Friesen
Former NESAP Postdoc

What is different about Cori for NERSC Users?

Edison (Cray XC w/ Intel Xeon Ivy-Bridge):

- 5000+ Nodes
- 12 Cores Per CPU
- 24 HW Threads Per CPU
- 2.4 GHz
- 8 DP Operations per Cycle
- 64 GB DDR Memory (2.6 GB/core)
- ~100 GB/s Memory BW
- 256b vector units
- 30 MB L3 cache per socket (12 cores)

Cori (Cray XC w/ Intel Xeon Phi KNL):

- 9600+ Nodes
- 68 Physical Cores Per CPU
- 272 HW Threads Per CPU
- 1.4 GHz
- 32 DP Operations per Cycle
- 16 GB of Fast Memory (0.24 GB/core)
  96GB of DDR Memory (1.4 GB/core)
  MCDRAM Has ~450 GB/s Memory BW
- No L3 cache
- 2 x 512b vector units

Optimization targets: OpenMP Threading, Vectors, Data management for MCDRAM

NESAP Optimization Strategy and Goals

NERSC

We're primarily working with existing codes to get them ready for Cori

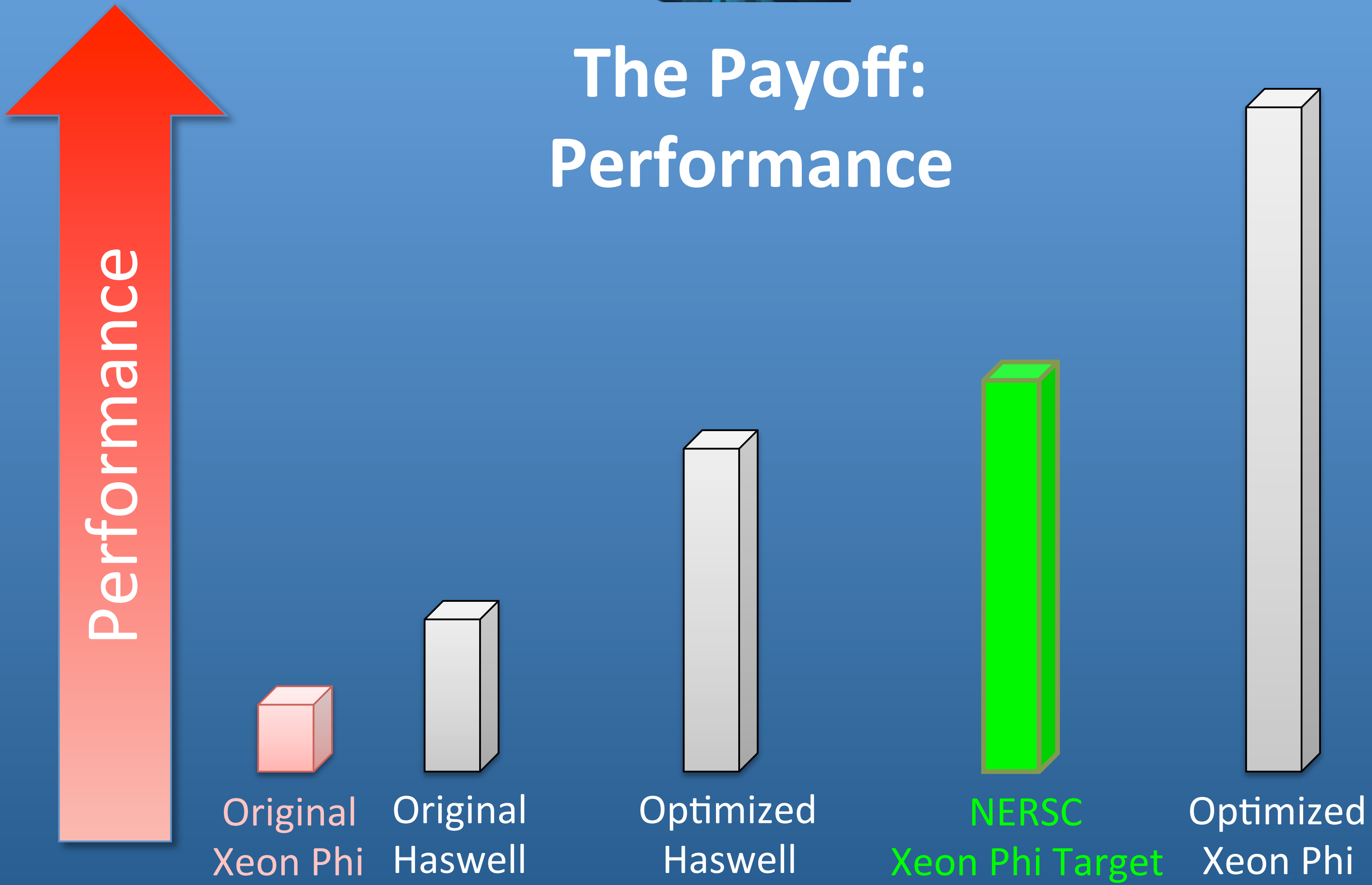
Goals

- Standard constructs for portability and maintainability
- Incorporate optimizations into code base by working directly with developers
- Collaborate closely with community to leverage expertise and expand NERSC influence and relevance

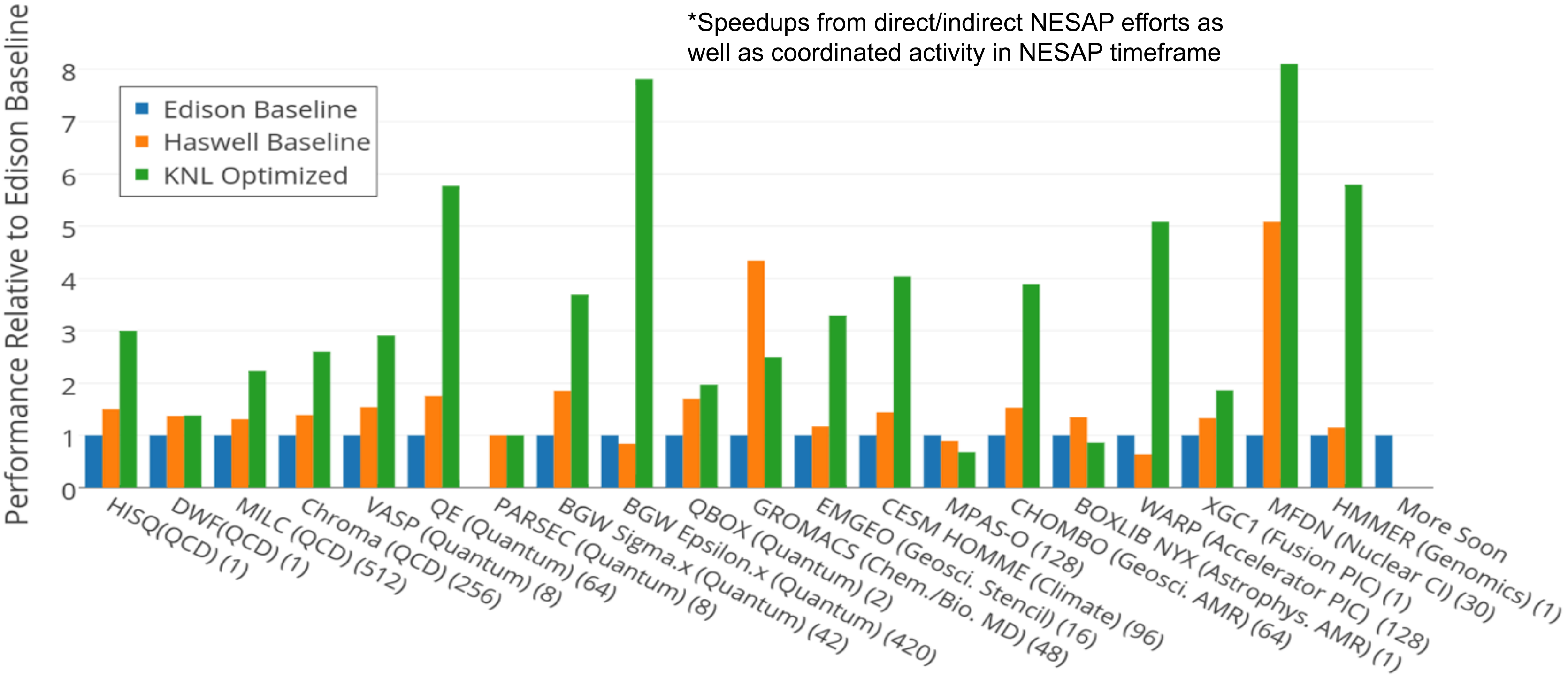
Strategy: Focus first on single-node optimization

- Enable fine-grained parallelism on light-weight cores via OpenMP
- Exploit dual 512b vector units
- Exploit 5X memory bandwidth due to MCDRAM by managing data access

The Payoff: Performance

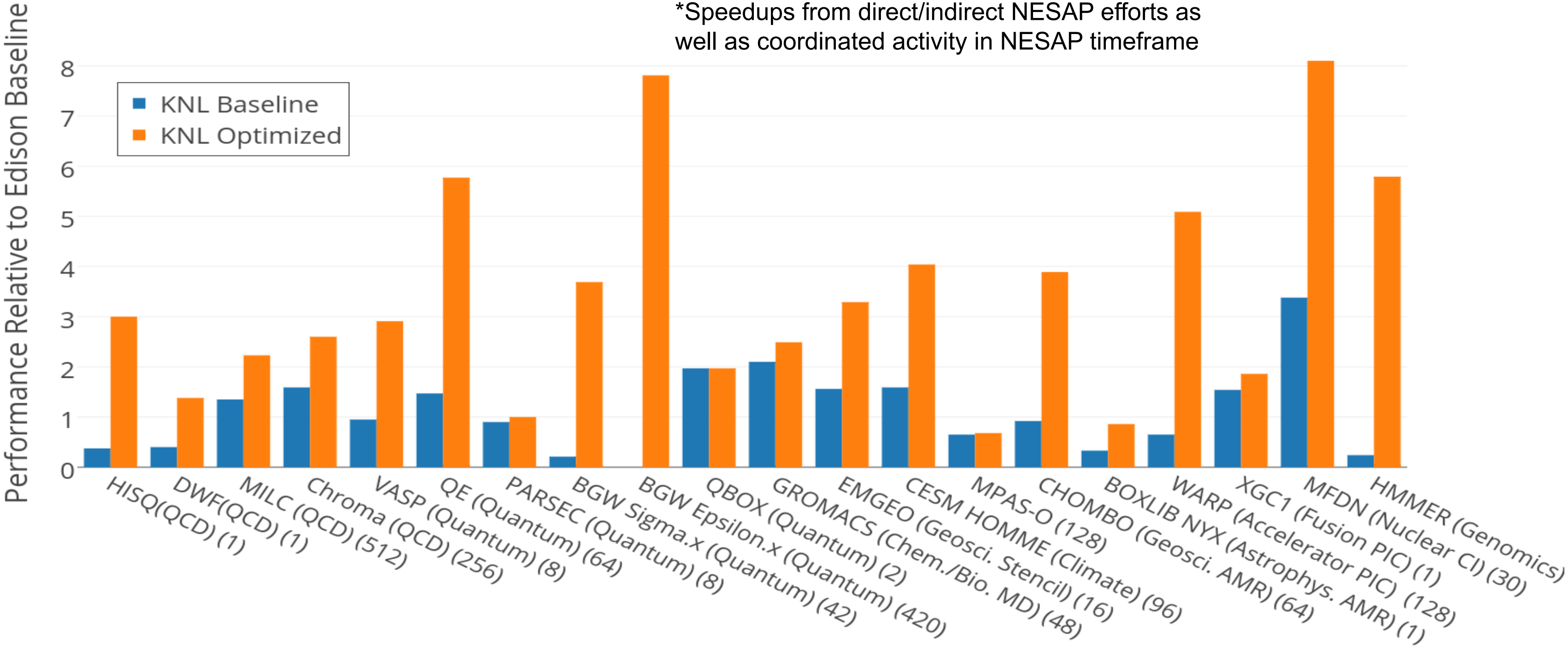


NESAP Code Performance on KNL



NESAP Code Performance on KNL

*Speedups from direct/indirect NESAP efforts as well as coordinated activity in NESAP timeframe



B: Baseline, original code
O: Optimized after NESAP work

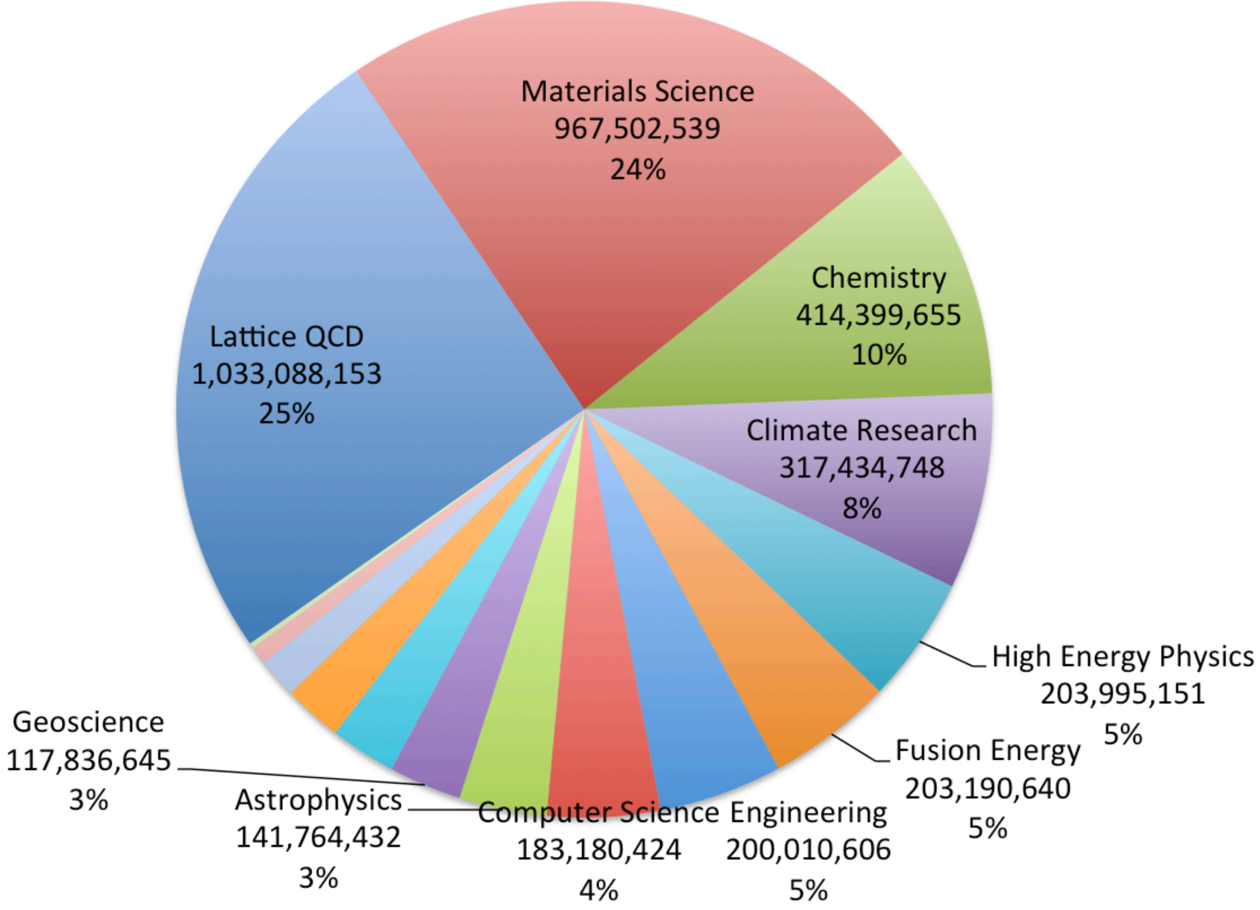
H: Haswell dual core
I: Ivy Bridge dual core
K: Xeon Phi KNL

Ratio	Performance per node	Comment
HB/IB	1.6 X	Business as usual; not on path to exascale
KO/HB	2.5 X	NESAP + KNL benefit over Haswell no opt
KO/IB	4.0 X	Cori KNL optimized benefit over Edison
HO/HB	2.3 X	NESAP code efforts only
KO/HO	1.2 X	Optimized KNL vs. optimized Haswell; on path to exascale
KB/HB	0.7 X	KNL vs. Haswell with no NESAP

KNL Usage by Science Category

Code Usage

CESM #7
ACME #20
WRF #29



Summary

Cori with light-weight Intel Xeon Phi processors provides unprecedented capability for DOE Office of Science research

NESAP has enabled large percentage of NERSC workload to run efficiently on new class of manycore system

Lessons learned and knowledge gained are being communicated to and applied by NERSC community

Postdoc program has been extremely valuable to NESAP and is helping to prepare next-generation workforce for HPC

Collaborations with application teams, vendors, and HPC community are necessary for success

There is no record in human history of a happy philosopher.

– H.L. Mencken

Climate Science at NERSC

Current Status

- HOMME-based atmosphere codes are running well on KNL
  - Performance is very good at small and medium scale
  - Scaling issues remain that are not understood completely
- Work is ongoing on large-scale coupled runs
  - OK at small scale, but not where would like to be
  - E.g. MPAS ice and ocean components
  - Postdoc is working on the issues
- Climate codes seem to be more sensitive to variability than most
  - Not understood, but seems to be system-related

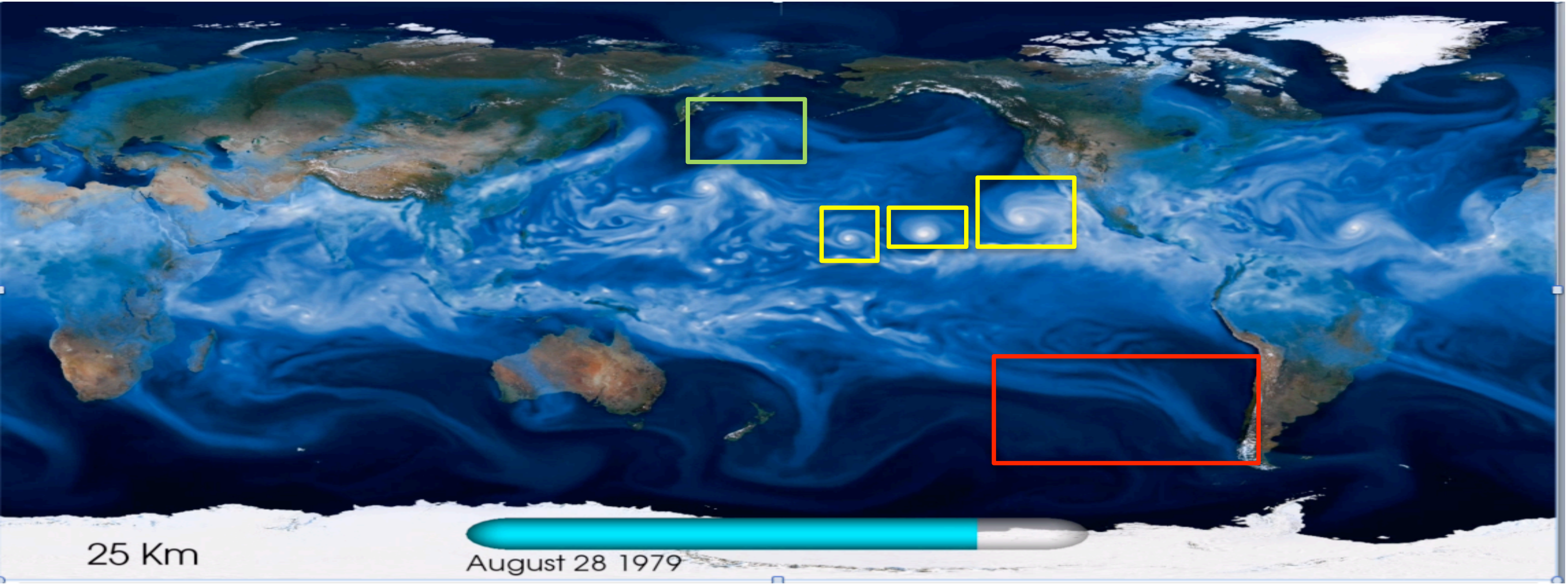
Deep Learning for Climate Datasets

How will extreme weather change in the future?

Need an objective tool for detecting extremes

- Pattern detection task
- Can Deep Learning come to the rescue?

Task: Find Extreme Weather Events



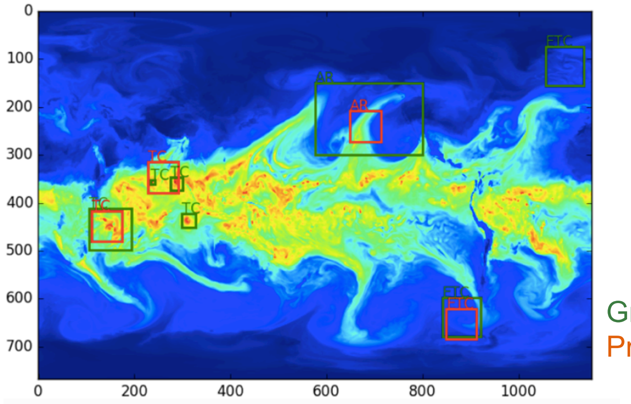
Deep Learning for Extreme Weather Detection

NERSC

First application of supervised and semi-supervised architectures for finding patterns in CAM5 data

DL methods are capable of extracting weather patterns with 85-99% accuracy (NIPS’17 paper)

Implementation scaled to 15PF on Cori Phase II (SC’17 paper)



Ground Truth
Prediction

Source and Contact: Prabhat (Prabhat@lbl.gov)

Looking Forward

NESAP for Data

Help experimental efforts transition to KNL and towards exascale

NESAP for NERSC 9 (2020) system when announced

Application portability recommendations (w/ ANL, ORNL)

- http://performanceportability.org

Explore ‘exascale’ programming models and languages

Influence standards committees (OpenMP, MPI)

Collaborate with CS researchers (algorithms & methods)

Transition broad community to manycore

84 Climate Projects at NERSC

PI Name	Org	Hrs (M)	Project Title
Leung, Ruby	PNNL	185.0	Accelerated Climate Modeling for Energy
O'Brien, Travis	LBNL	131.0	Calibrated and Systematic Characterization Attribution and Detection of Extremes
Meehl, Gerald	NCAR	40.6	Climate Change Simulations with CESM: Moderate and High Resolution Studies
Lin, Wuyin	BNL	32.0	Evaluation and improvement of Convective Parameterizations in ACME model
Collins, William D.	Berkeley Lab	23.6	Multiscale Methods for Accurate, Efficient, and Scale-Aware Models of the Earth System
Um, Junshik	U. Illinois U-C	22.2	Linear depolarization ratios of hexagonal ice crystals using an exact method: Applications to remote sensing and scattering database
Leung, Ruby	PNNL	10.6	Water Cycle and Climate Extremes Modeling (WACCEM)
Maxwell, Reed	Col Sch of Mines	9.3	High-resolution, integrated terrestrial and lower atmosphere simulation of the contiguous United States
Compo, Gil	U. Colo.	8.2	Ocean Atmosphere Reanalyses for Climate Applications (OARCA) 1815-2016

ACME v1 Coupled System

- Cori-KNL is fast and the most **efficient** system capable of running ACME v1 high-resolution (25 km atm) on much fewer nodes:
  - 1.1 SYPD, 825 nodes
  - 1.8M NERSC core-hours per simulated year
  - Atmosphere and atmosphere dycore run faster on KNL nodes vs conventional Xeon node, but the overall model is slower as of today
- ALCF MIRA:
  - 0.5 SYPD on 8192 nodes
  - 7.1M core-hours per simulated year
- OLCF Titan
  - 1.4 SYPD on 7448 nodes
  - 3.9M Titan core-hours per simulated year

Source: Mark Taylor, Sandia National Laboratory

CESM High Resolution

We are embarking on exciting set of runs using the highest possible fully-coupled resolution with the CESM, and this project is beginning on Cori with the start of a control simulation. We hope to create a full set of simulations with a control and 20th-21st century simulations. – Susan Bates, NCAR