Current and Next-Generation Supercomputing and Data Analysis at NERSC

Richard Gerber

NERSC Senior Science Advisor
High Performance Computing Department Head

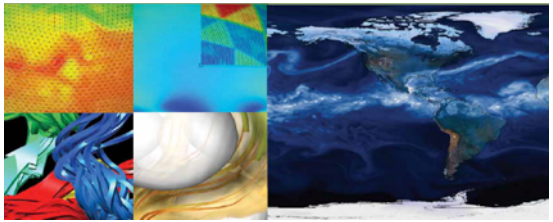
NERSC's Mission

NERSC's mission is to accelerate scientific discovery at the DOE Office of Science through high performance computing and data analysis.

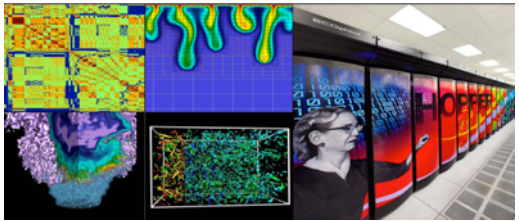
NERSC: the Mission HPC Facility for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

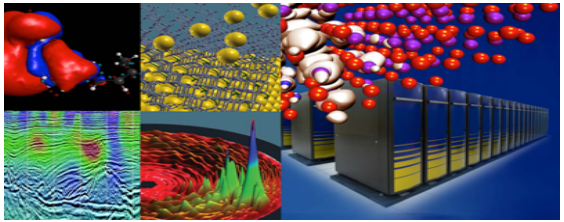
Largest funder of physical science research in the U.S.



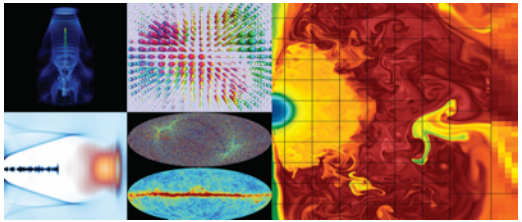
Bio Energy, Environment



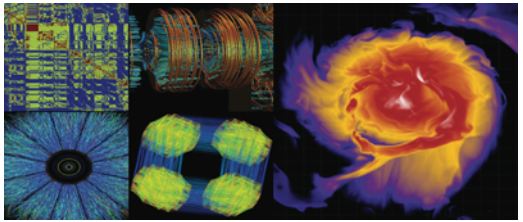
Computing



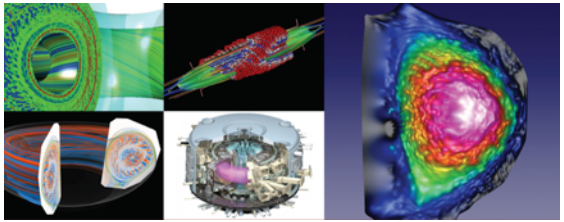
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

Focus on Science

- NERSC users produce more scientific publications than any other center in the world*; ~2K/year
- 1,036 citations via Web of Science in 2017 so far (not perfect!)

Journal	Articles
Physical Review B (condensed matter and materials)	55
Astrophysical Journal	36
Physical Review Letters	38
Journal of Chemical Physics	32
Nature Communications	30

nature
International weekly journal of science

5 in Nature
30 in Nature Comm.
70 in 12 journals

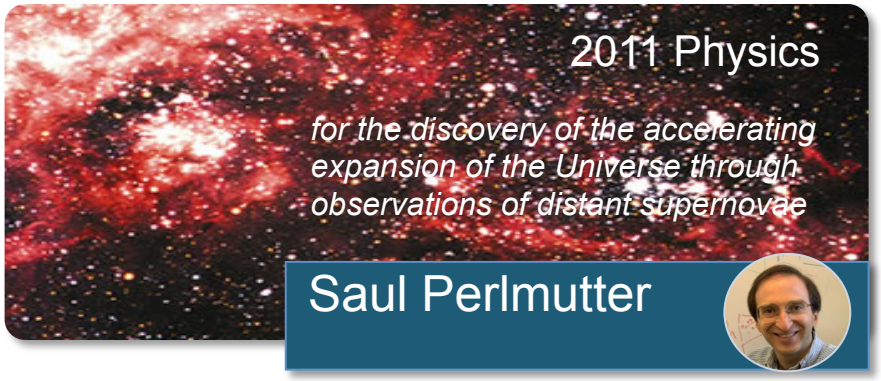
Science
AAAS

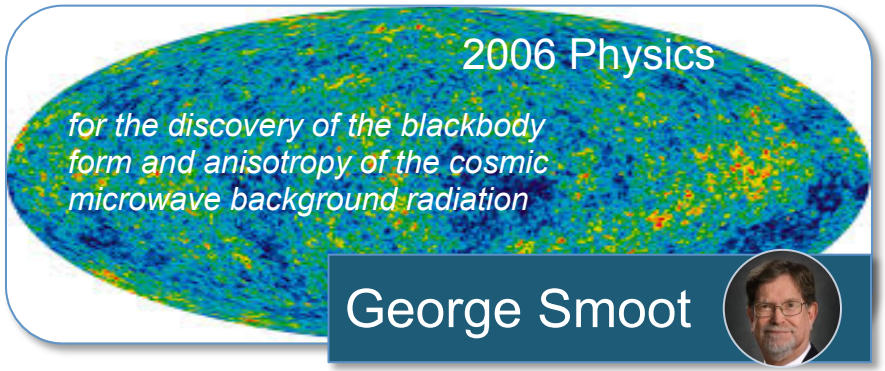
4 in Science

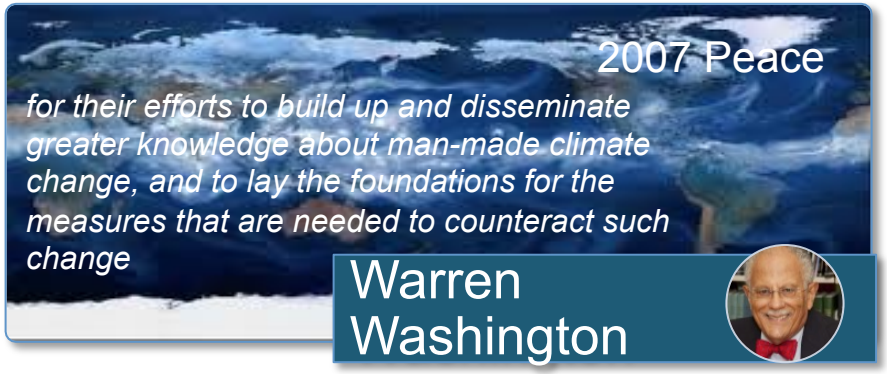
PNAS
Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

11 in PNAS

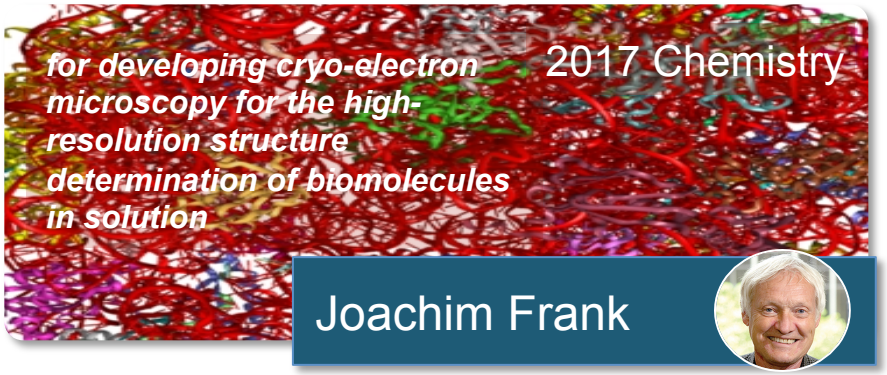
Nobel-Prize Winning Users



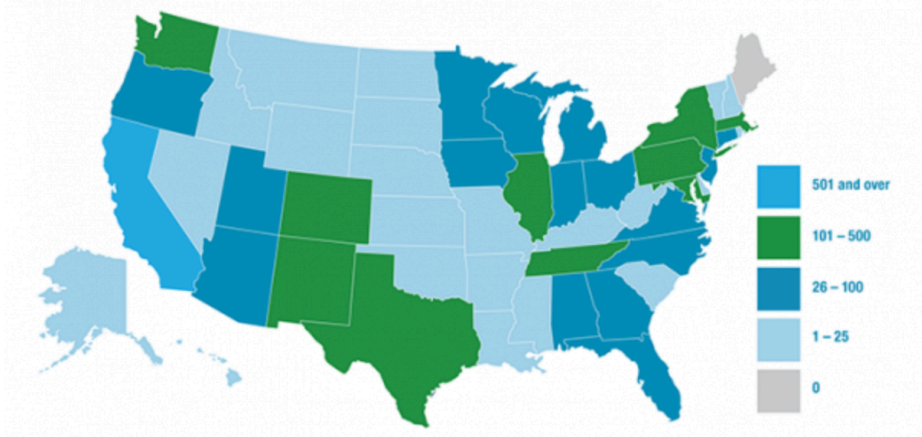




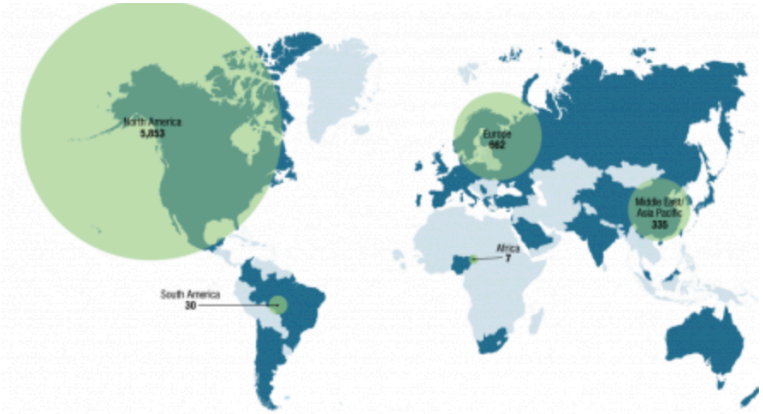
Nobel-Prize Winning Users



User Demographics

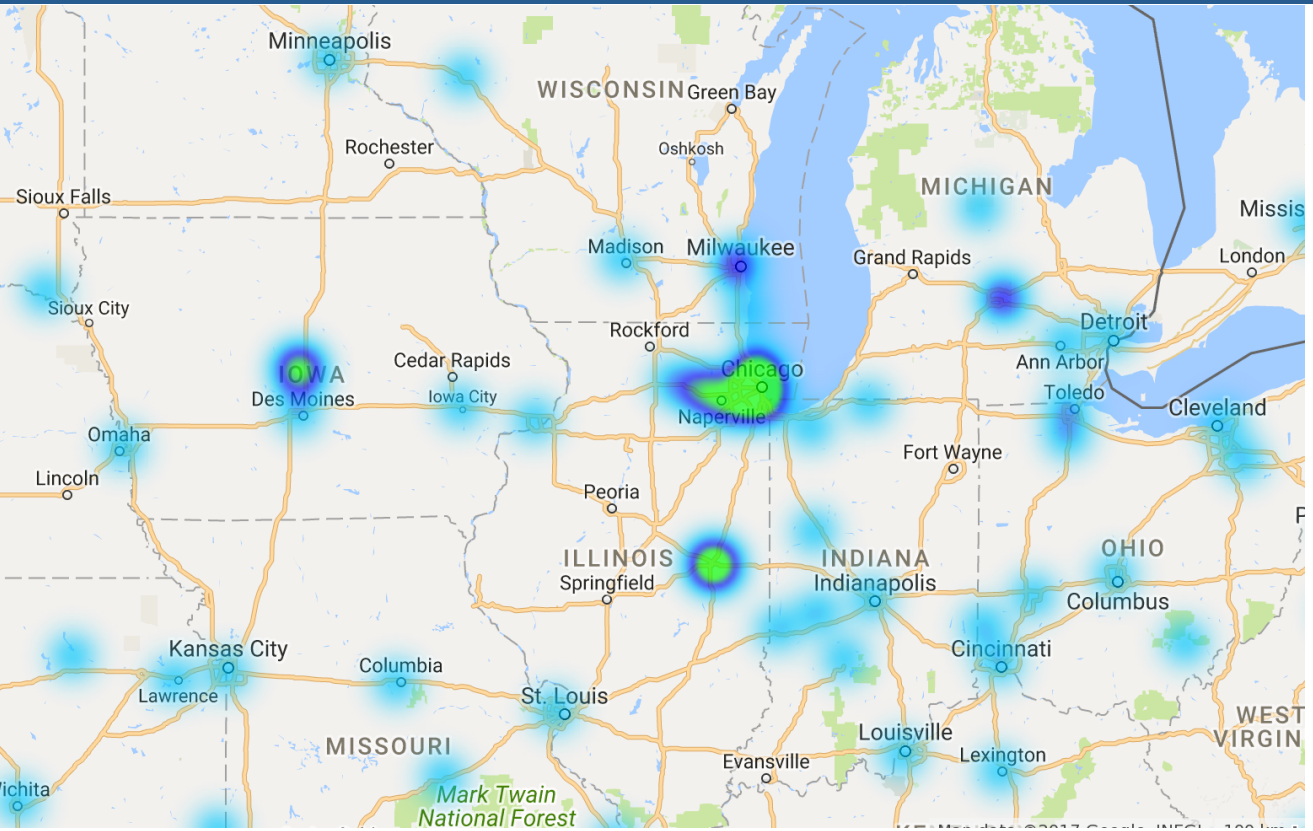


48 states
40 countries
Universities & national labs



7,000 users
800 projects
700 codes

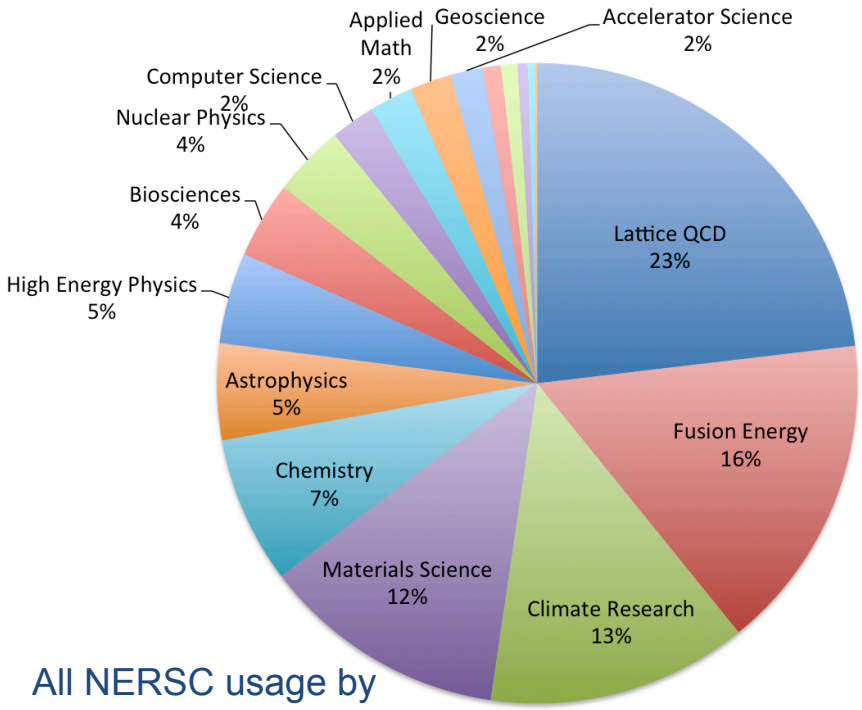
Large contingent of active users in Iowa



63 users in Iowa

Iowa State	29
Ames Lab	26
University of Iowa	5
Drake University	1
St. Ambrose	1
Krell Institute	1

Top Iowa Users



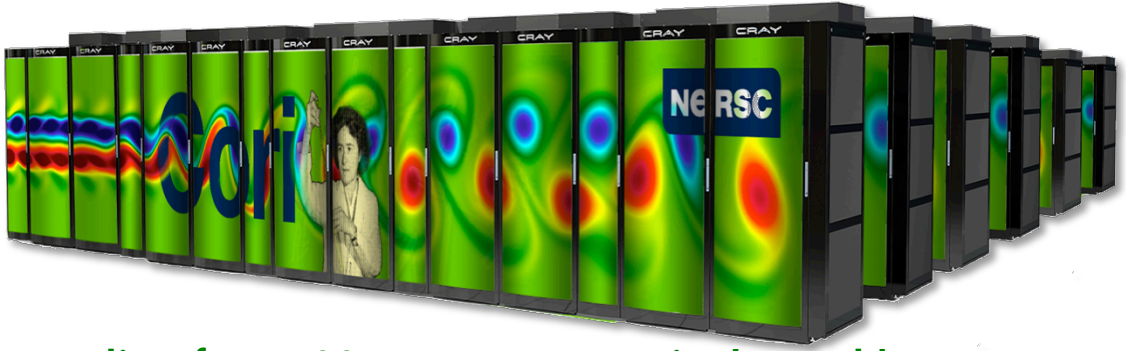
All NERSC usage by science category

Yong Han	Ames Lab	29 Mhr	Catalysis
Pieter Maris	Iowa St.	12 Mhr	Nuclear Physics
Xuewbu Ou	Ames Lab	1.5 Mhr	Materials
Minsung Kim	Ames Lab	1.4 Mhr	Materials
Xin Zhao	Ames Lab	1.1 Mhr	Materials
Lin Yang	Ames Lab	0.74 Mhr	Materials
Yang Li	Iowa St.	0.53 Mhr	Nuclear Physics
F. Zahariev	Ames Lab	0.53 Mhr	Metal Extractants
Diego Floor	U. Iowa	0.50 Mhr	LQCD
M. Dick-Perez	Ames Lab	0.36 Mhr	Metal Extractants
M. C. Nguyen	Ames Lab	0.34 Mhr	Materials
Nathan Weeks	Iowa St.	0.28 Mhr	Nuclear Physics

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

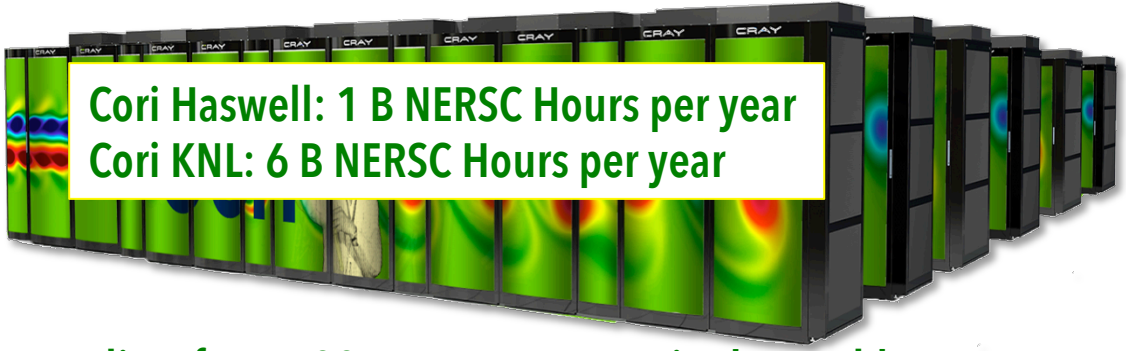
Edison

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

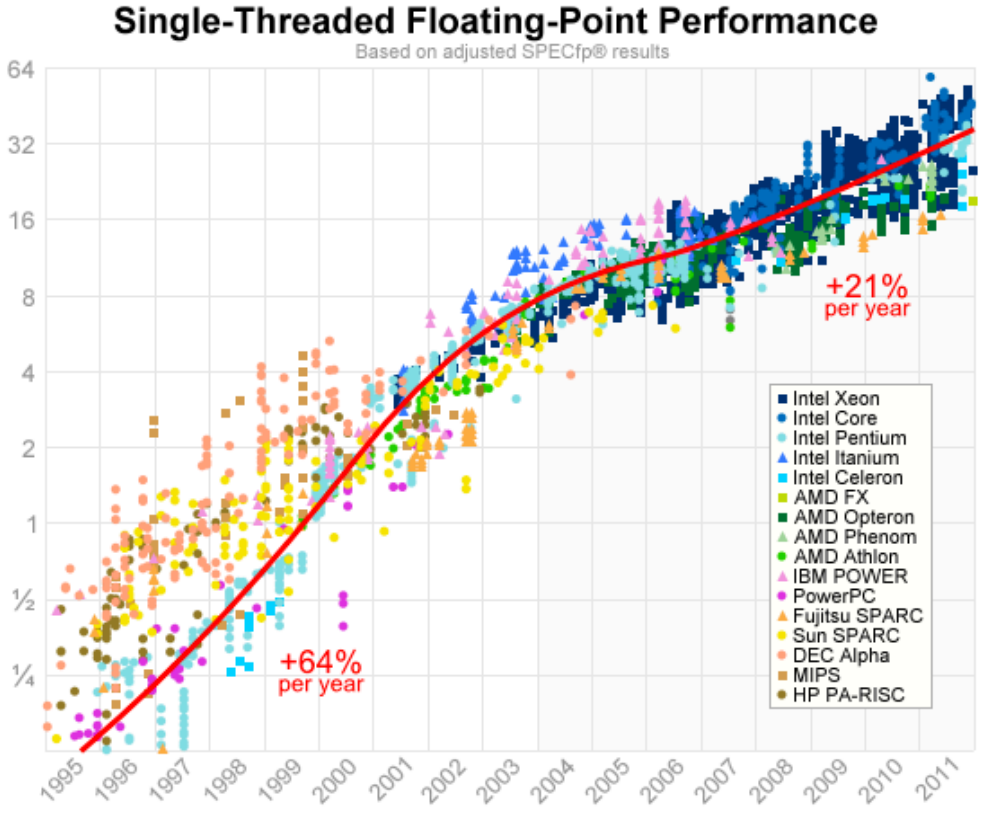
Edison

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

“The only thing constant is change”

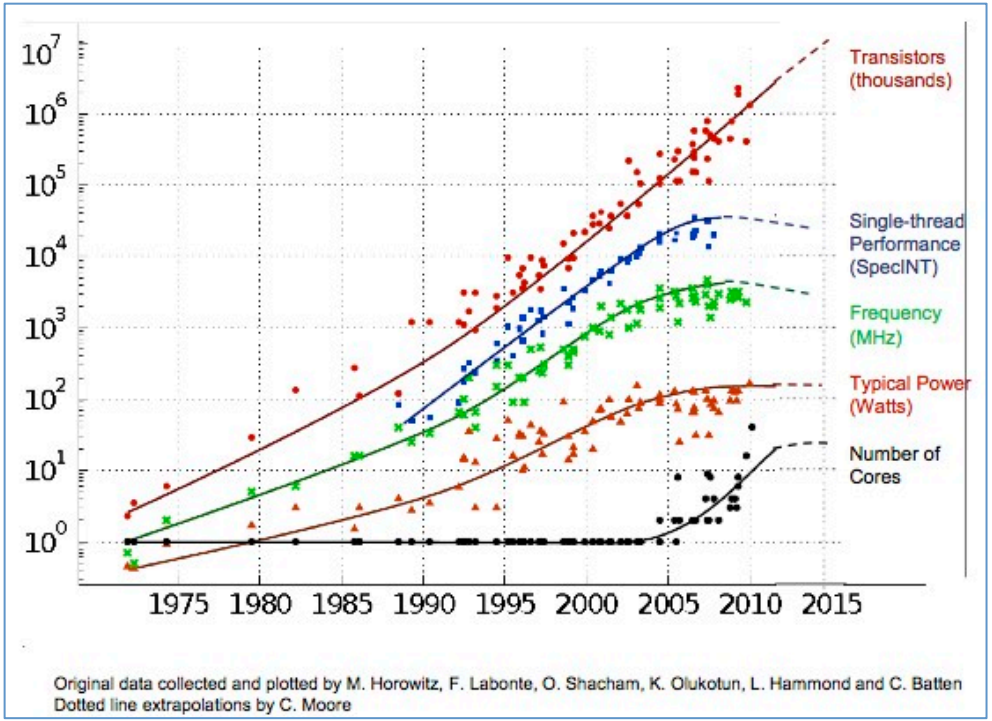
–Heraclitus of Ephesus

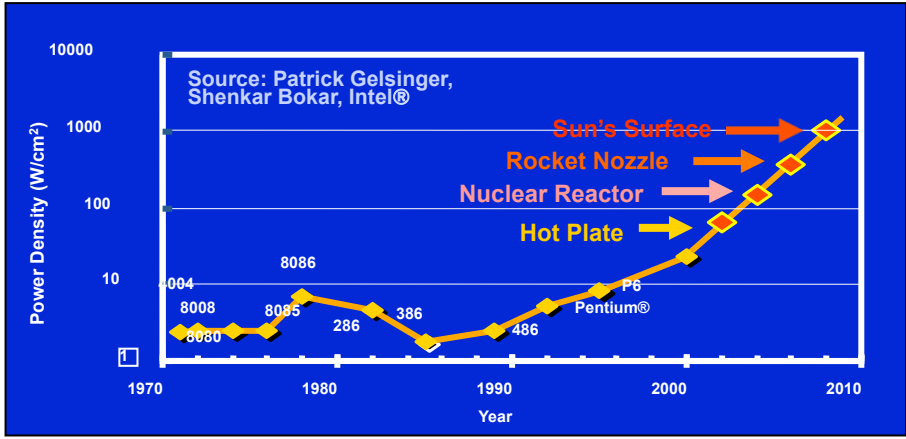
Single Processor Performance



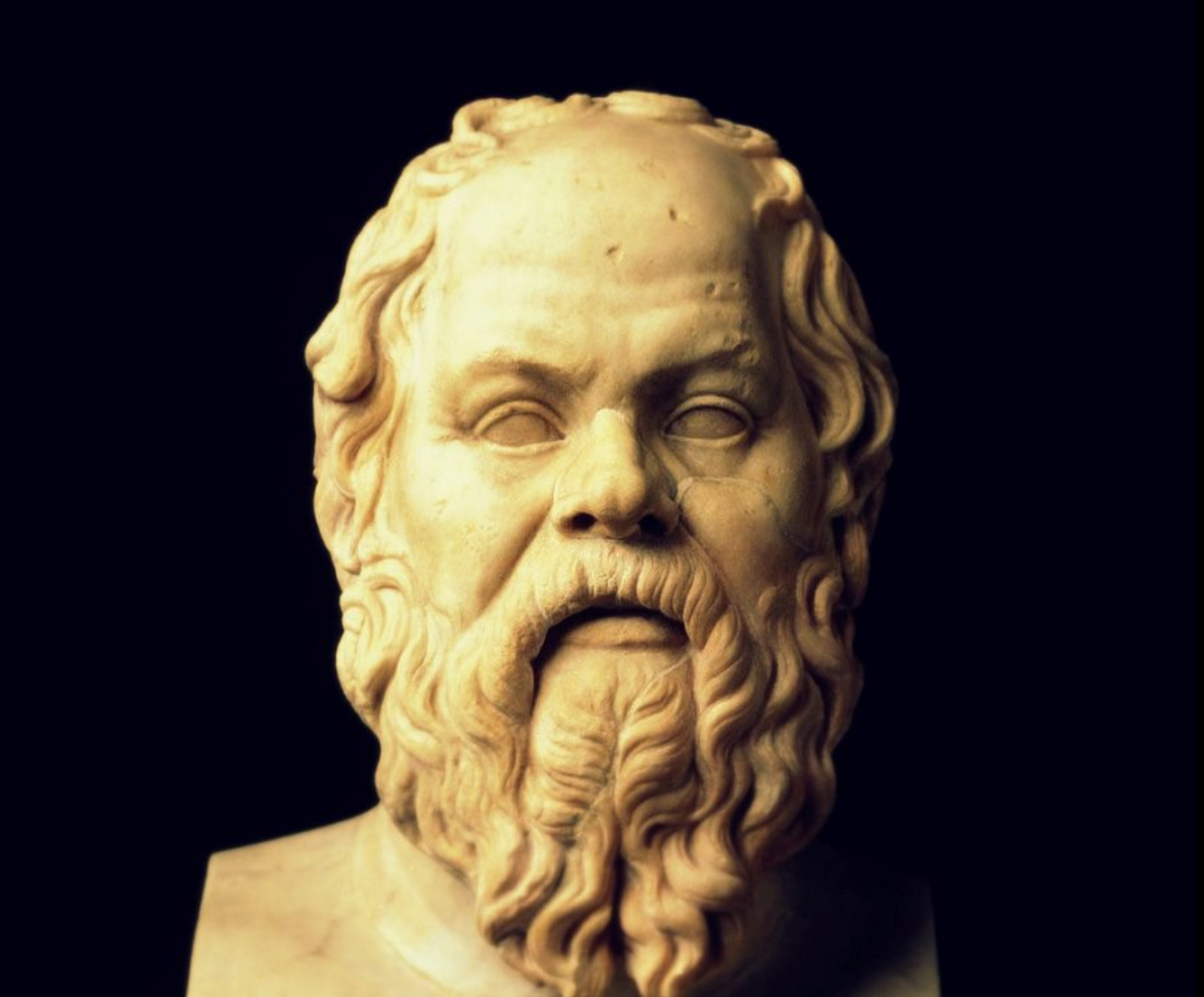
Every year there was a new CPU technology that enabled single-thread performance to increase

Change was coming ...





Driven by power consumption and dissipation toward lightweight cores



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

– Socrates

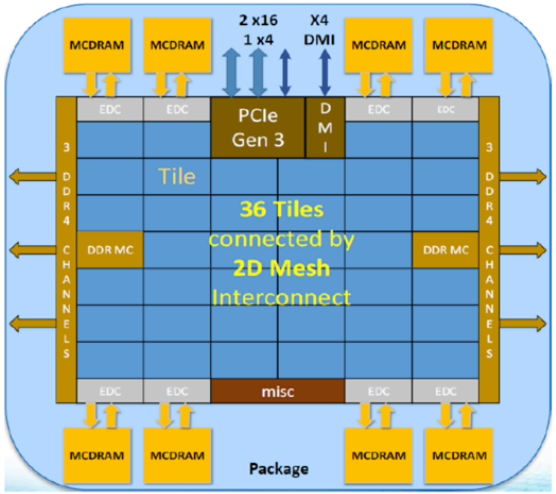
TECHENABLEMENT
Get known for the tech we teach

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

May 2, 2014 by Rob Farber — Leave a Comment

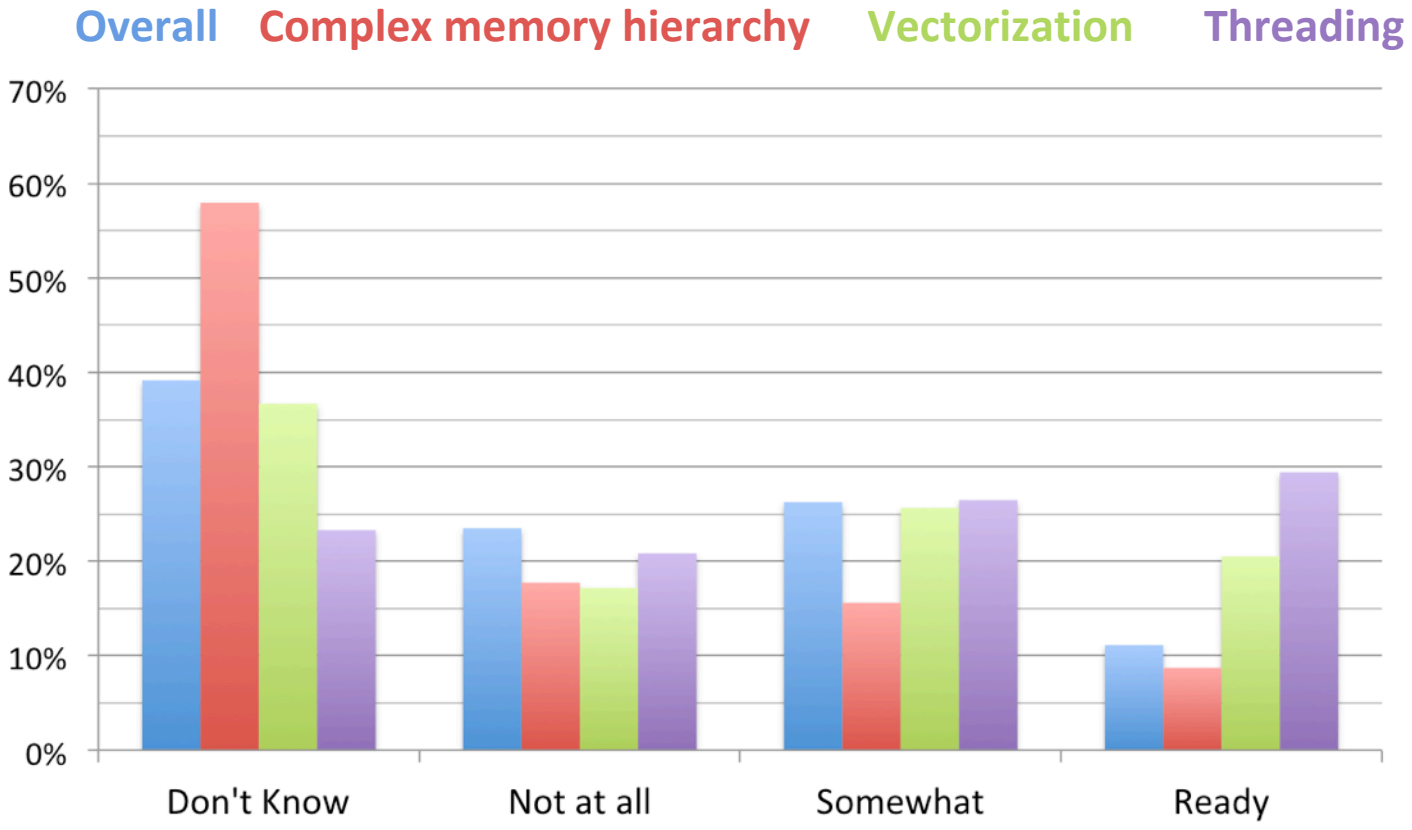
30 PFlop System will be a boon to science because of new capabilities, but the Intel Xeon Phi many-core architecture will require a code modernization effort to use efficiently.

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



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

2014 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

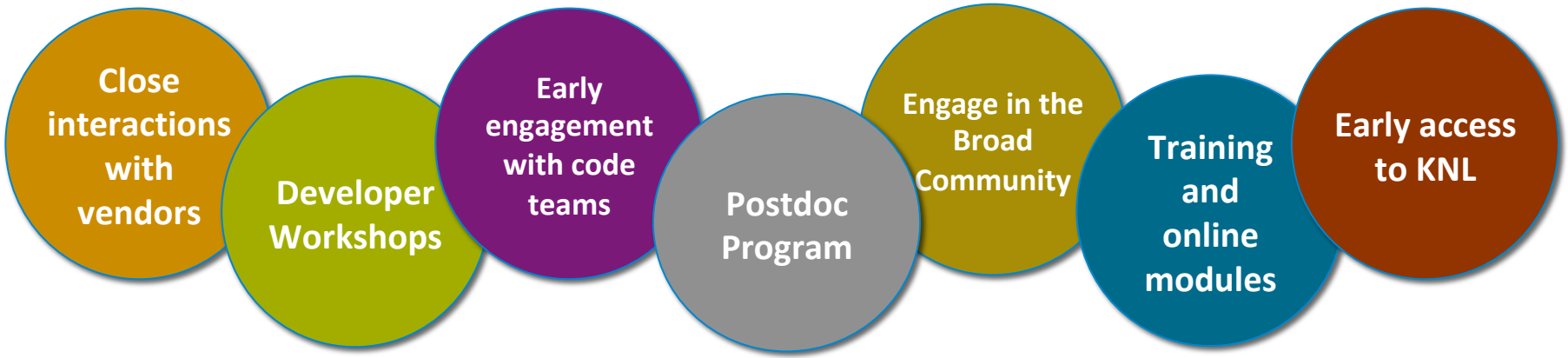
How can NERSC's diverse community of 7,000 users, 750 projects, and 700 codes use Cori's Intel Xeon Phi Knights Landing processors at high performance

Business as usual was over

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

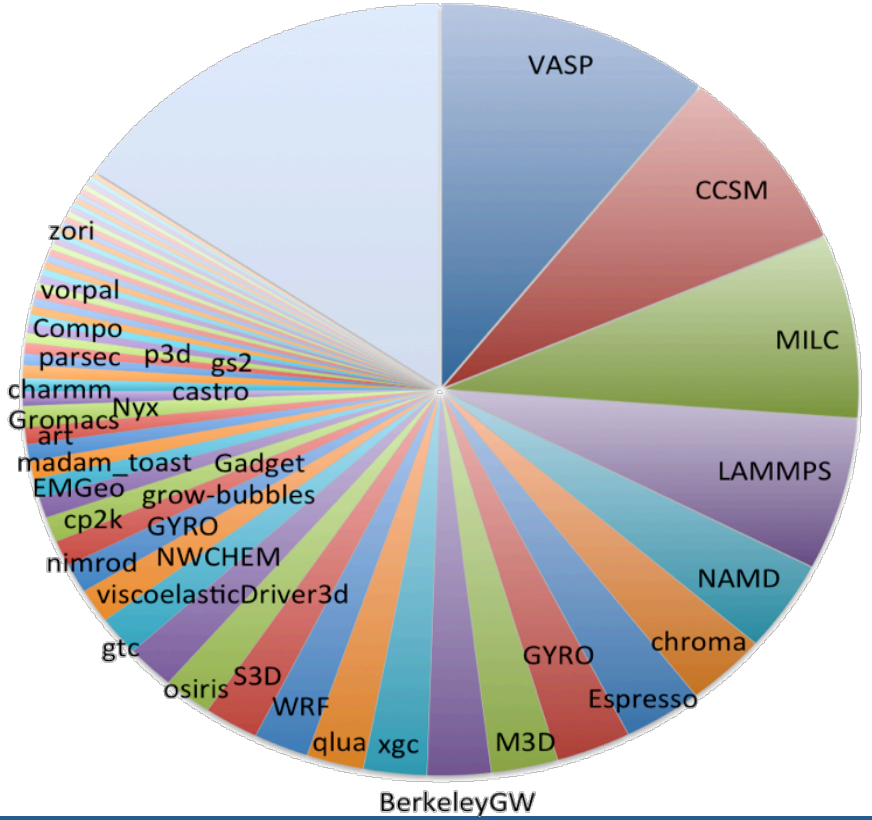
Selected projects must

- 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

Code Usage at Start of NESAP

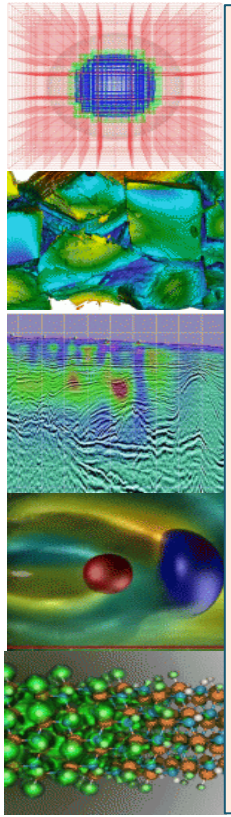


NERSC has 7,000 users, but a relatively small number of codes use a lot of hours

By working with ~20 codes, can cover ~50% of workload

A very long tail make up the last 25%

NESAP Codes



Advanced Scientific Computing Research

Almgren (LBNL) **BoxLib AMR**

Trebotich (LBNL) **Chombo-crunch**

High Energy Physics

Vay (LBNL) **WARP & IMPACT**

Toussaint(Arizona) **MILC**

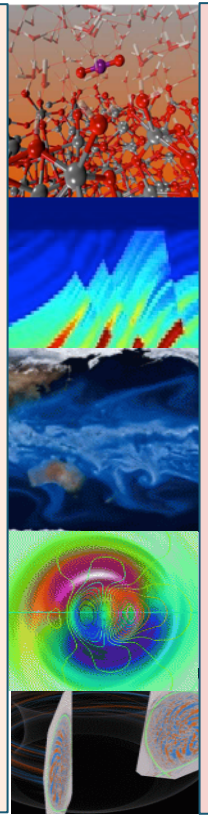
Habib (ANL) **HACC**

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**

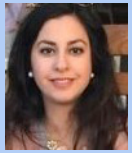
Fusion Energy Sciences

Jardin (PPPL) **M3D**

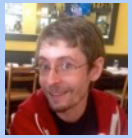
Chang (PPPL) **XGC1**

New Postdoc Program

Open


Zahra Ronaghi
Tomopy


Andrey Ovsyannikov
Chombo-Crunch


Bill Arndt
HIPMER/ HMMER/MPAS


Rahul Gayatri
SW4


Tuomas Koskela
XGC1


Kevin Gott
PARSEC

Open

NERSC Application Performance Group formed


Charlene Yang

NESAP Staff Contributors


Katie Antypas

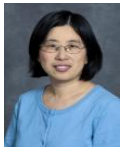

Jack Deslippe


Richard Gerber

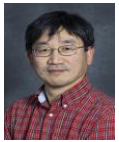

Nick Wright


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
- 256b vector units
- 64 GB DDR Memory (2.6 GB/core)
- ~100 GB/s Memory BW
- 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
- 2 x 512b vector units
- 16 GB of Fast Memory (0.24 GB/core)
  96GB of DDR Memory (1.4 GB/core)
  MCDRAM Has ~450 GB/s Memory BW

NESAP Optimization Strategy

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

Goals

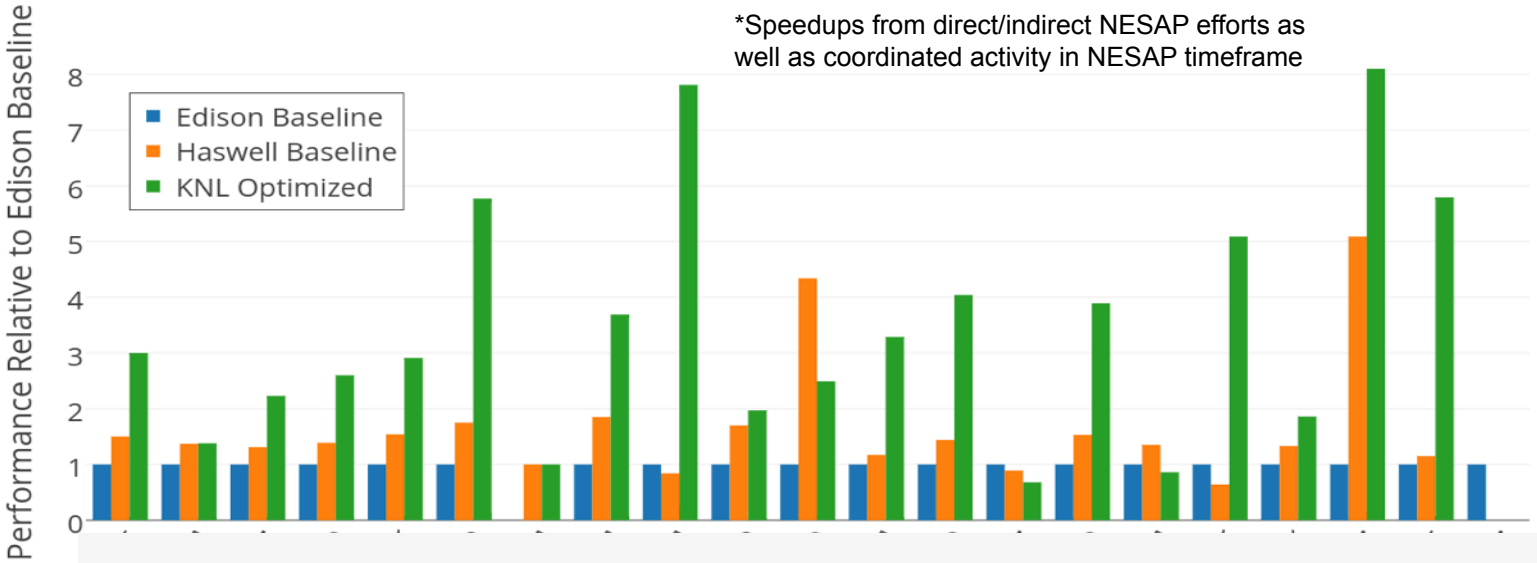
- Standard constructs for portability and maintainability
- Incorporate optimizations into code base by working directly with developers
- Collaborate closely with community to leverage expertise, communicate lessons learned, and expand NERSC influence and relevance

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

Two Years Later ...

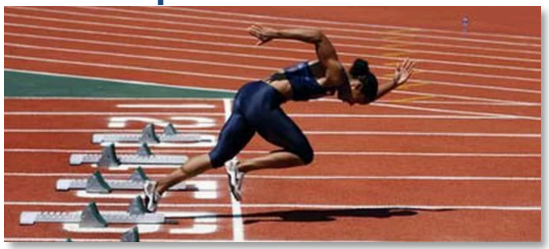
NESAP Code Performance on KNL



NESAP + KNL: Upgrade toward exascale

Old Code

Optimized for KNL



Exascale

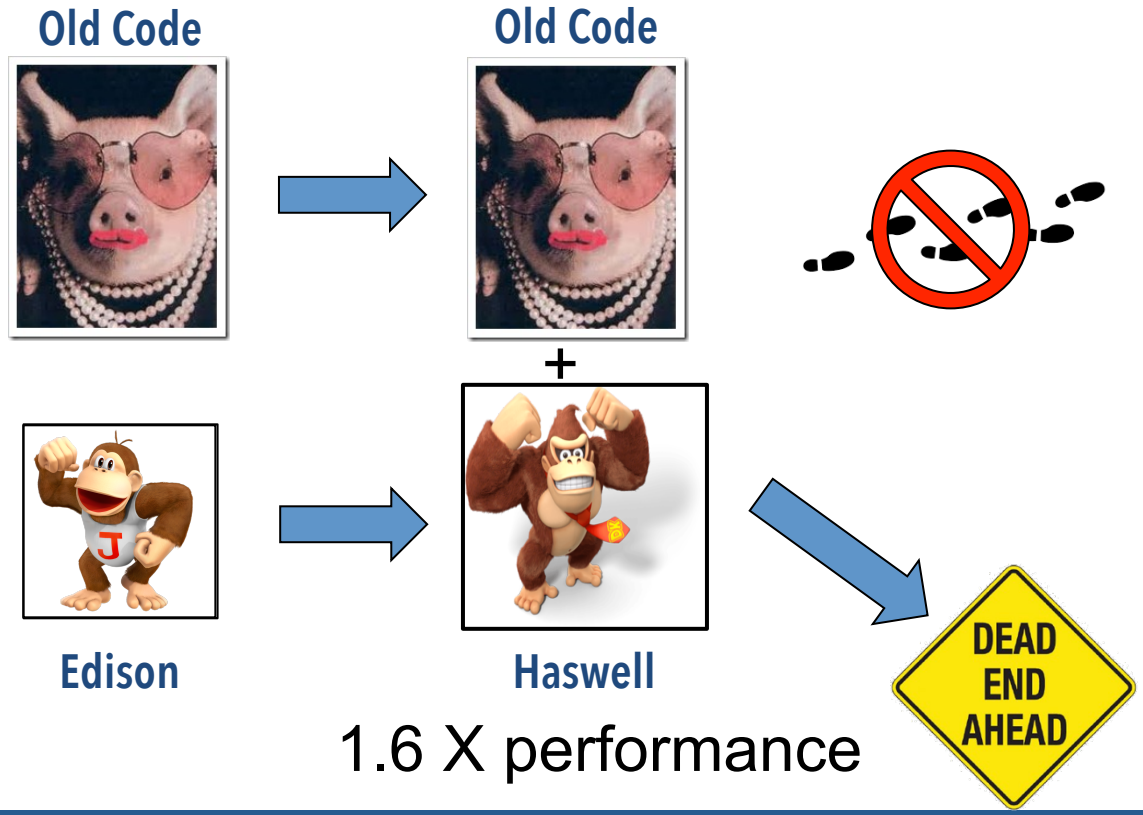
+

Haswell

KNL

= 2.5 X performance increase per node

Business as Usual: Dead End



NESAP Only

Old Code

Optimized for KNL

+

Haswell

Haswell

= 2.1 X performance

NESAP optimization efforts by themselves improved code performance by 2.1 X on x86 (Intel Xeon) processors

No motivation to optimize codes while next-gen processors gave ~60% improvements by themselves

NESAP + KNL vs. NESAP + Haswell

Optimized for KNL

Optimized for KNL

Haswell

vs.

KNL

Exascale

1.2 X performance
Less energy

Remaining Challenge for the Masses

Old Code

Haswell

vs.

Old Code

KNL

= 0.7 X performance

NERSC and DOE Office of Science remaining challenge is to get broad community to run efficiently on manycore

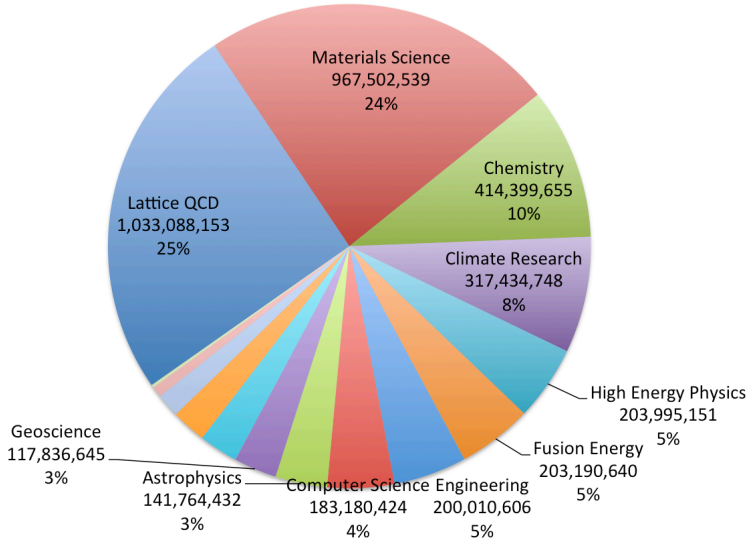
Summary: Good Adoption of KNL

NERSC Exascale Science Application Program (NESAP) codes are running on KNL at about 3.5X-4X their pre-NESAP performance on Edison per node.

150 projects have used > 1 M NERSC Hours on KNL
233 projects have used > 100 K NERSC Hours on KNL
Still leaves ~500 to move over

32% of hours used by jobs using > 1,024 nodes (69K cores)

NERSC supported 6 Gordon Bell submissions using Cori KNL



Cori provides a large increase in NERSC Hours available to Office of Science researchers at NERSC (3X+ in 2018 over 2016)

Example Science Highlights

A Record Quantum Circuit Simulation

Scientific Achievement

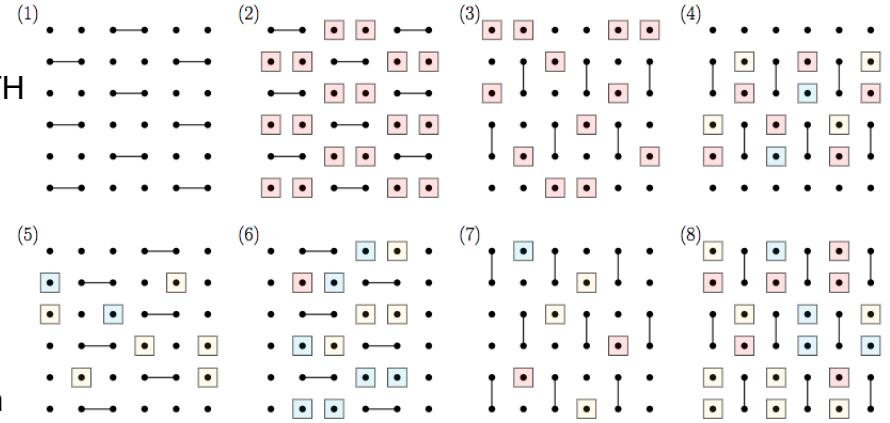
Researchers from the Swiss Federal Institute of Technology (ETH Zurich) used NERSC's 30-petaflop supercomputer, Cori, to successfully simulate a 45-qubit (quantum bit) quantum circuit, the largest simulation of a quantum computer achieved to date.

Significance and Impact

The current consensus is that a quantum computer capable of handling 49 qubits will offer the computing power of the most powerful supercomputers in the world. This new simulation is an important step in achieving "quantum supremacy"— the point at which quantum computers finally become more powerful than ordinary computers.

Research Details

- In addition to the 45-qubit simulation, the researchers also simulated 30-, 36- and 42-qubit quantum circuits.
- For the 45-bit simulation, they used 8,192 of 9,688 Intel Xeon Phi processors and 0.5 petabytes of memory.



Thomas Häner, Damian S. Steiger, 0.5 Petabyte Simulation of a 45-Qubit Quantum Circuit, arXiv:1704.01127 [quant-ph]

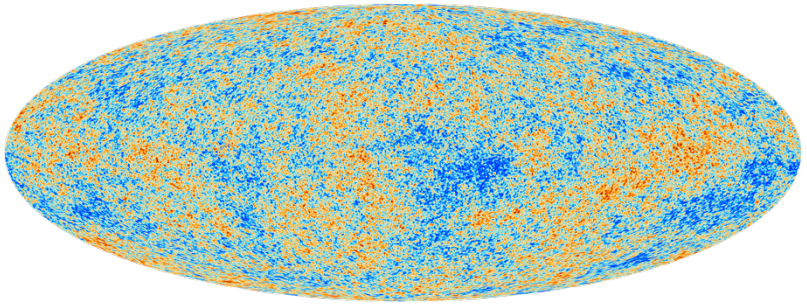
NERSC PI: T. Haner, ETH Zurich

Berkeley Lab Software Scales to 658,784 Cori Cores for Cosmic Microwave Background Analysis

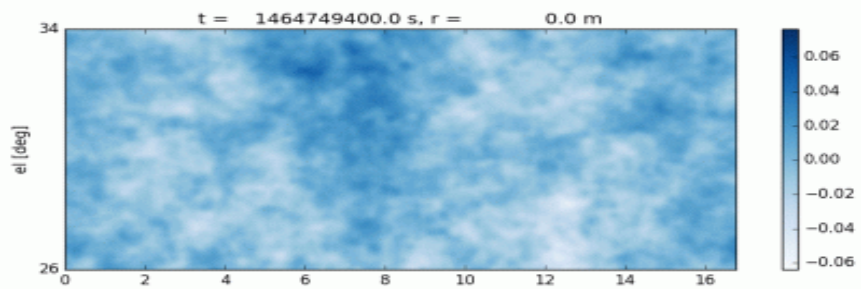
Berkeley Lab CS Computational Cosmology Center, NERSC, Intel, Cray collaborate via NESAP to scale key CMB software to full Cori system.

The TOAST (Time Ordered Astrophysics Scalable Tools) data simulation and reduction framework achieved a critical project milestone for upcoming experiments.

The ground-based CMB-S4 project will gather 35X more data than the Planck satellite did and will require TOAST's enhanced capabilities on Cori.



NERSC was used to reduce and interpret Planck results to create this map of the CMB



TOAST has incorporated modules to account for atmospheric effects

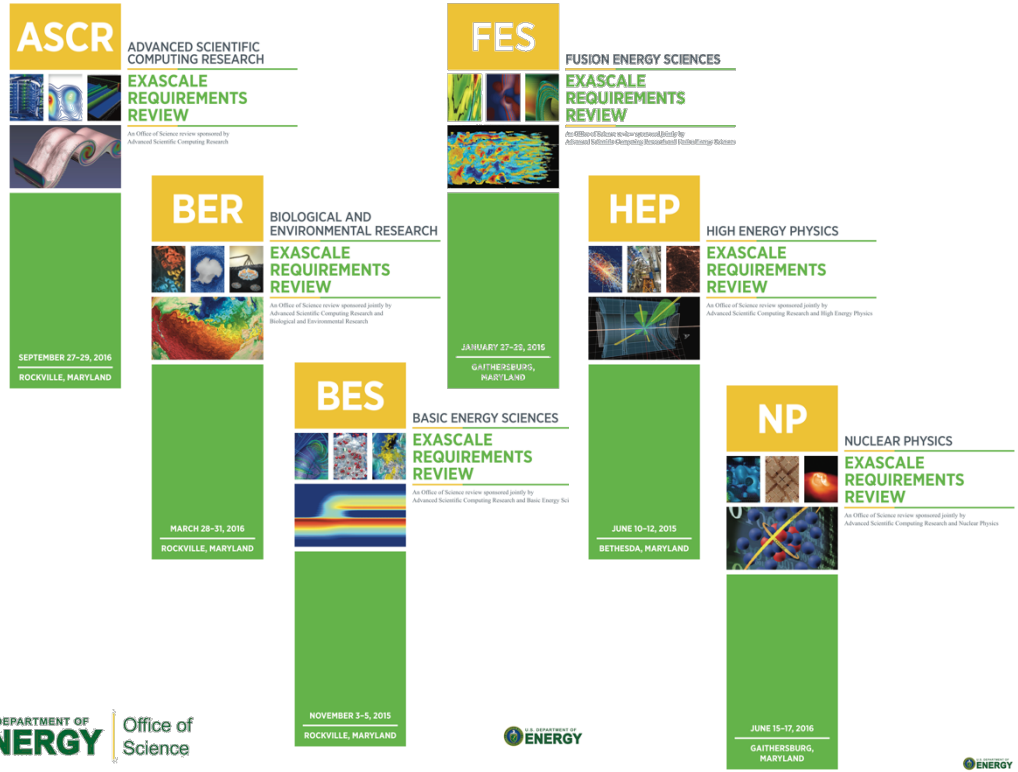
There is no record in human history of a happy philosopher.

– H.L. Mencken

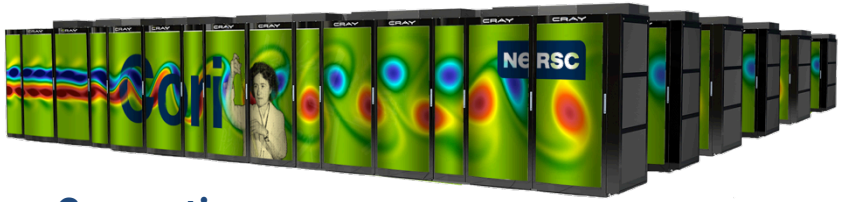
Data

DOE Exascale Requirements Reviews

- Reviews with 6 Office of Science programs
- Scientists, CS researchers, facility staff
- Requirements and productivity needs for an exascale ecosystem
- Identify opportunities for collaborations among SC programs and facilities

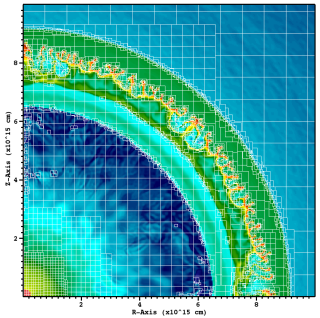


Cross Cutting Findings



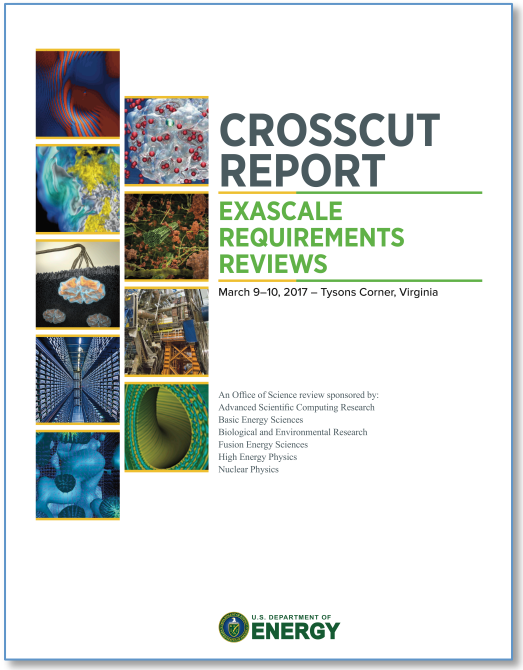
Computing

Data



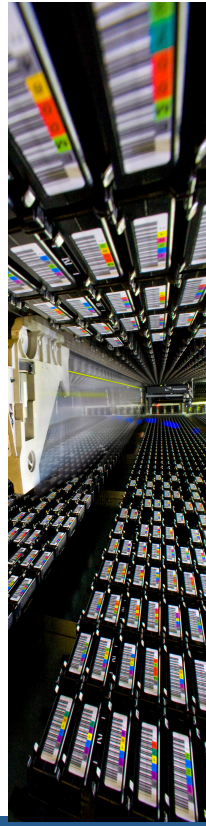
Software and Applications

Workforce and Training



Data Findings

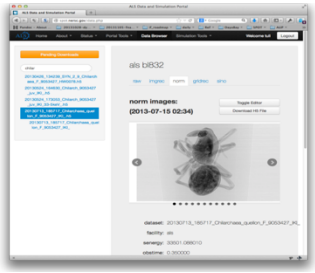
- Support for analysis tools, many of which differ significantly from traditional simulation software.
- Complex workflows that include data movement, co-scheduling.
- Data management, archiving, and curation well beyond what is in common practice today.
- Ability to share, transfer, and access data at remote sites with ease.
- Input/output capability of large HPC systems that scale with their computational capability.
- Scheduling and allocation policies to support workflow needs, especially data analysis requirements at experimental facilities: real-time, pseudo-real time, co-scheduling, variable job requirements, and allocations based on other resources like disk and memory.



NERSC's data vision

Transform science by enabling seamless large-scale data pipelines and analysis on leading-edge HPC systems and platforms

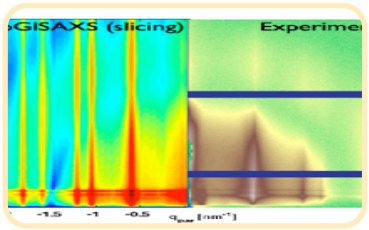
Superfacility: A network of connected facilities, software and expertise to enable new modes of discovery



Experimental Facilities

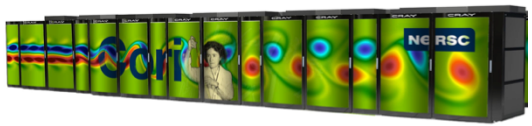
Real-time analysis and Data management

New Mathematical Analyses



Fast Implementations on latest computers

Computing Facilities



ESnet
ENERGY SCIENCES NETWORK

Network for Big Data Science

To support this vision we have created 4 initiatives, each with measurable goals

Superfacility Initiative: **We will implement the superfacility model, collaborating with Office of Science User Facilities to enable seamless and high performing end-to-end workflows**

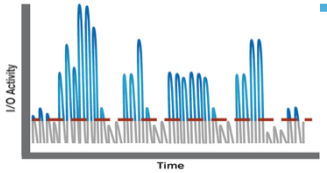
Systems Initiative: **We will evaluate emerging technologies and design NERSC-9 & -10 to support large-scale data analysis and simulations with tight integration to storage systems and edge services**

Software Initiative: **We will partner with the community to develop and enable data analytics and management software to run at scale on HPC systems**

User Engagement Initiative: **We will engage with the user community to optimize and support data pipelines for large scale HPC systems**

NERSC has begun to address some issues raised by users

I/O is too slow



Burst Buffer more than doubles I/O bandwidth

It's difficult to bring complex software stacks to HPC systems

User defined images with Shifter

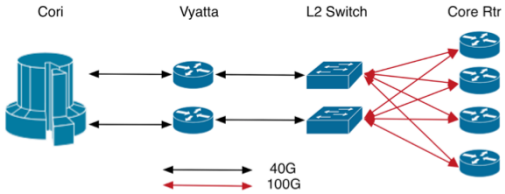
I need real-time feedback for my workflow

SchedMD

Real-time queues

Internal network limits how I can import data to supercomputer

SDN



There is limited software for analytics on HPC systems

New analytics and ML libraries

Caffe scikit learn Spark K TensorFlow

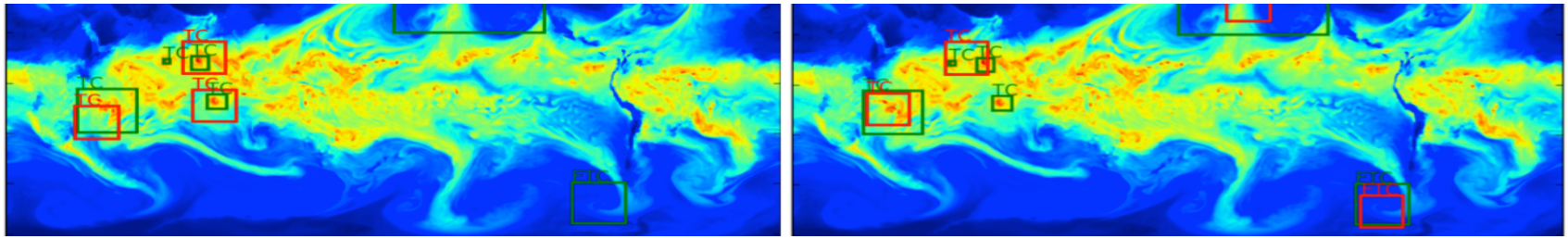
Deep Learning on Cori KNL

NERSC is actively exploring Deep Learning for Science

- Collaborating with leading vendors to optimize and deploy stack
- Collaborating with leading research institutions to develop methods
- Drive real science use cases

Deep Learning at 15 PF on NERSC Cori (Cray + Intel KNL)

- Trained in 10s of minutes on 10 terabyte datasets, millions of Images
- 9600 nodes, optimized on KNL with IntelCaffe and MKL (NERSC / Intel collaboration)
- Synch + Asynch parameter update strategy for multi-node scaling (NERSC / Stanford)



Identified extreme climate events using supervised (left) and semisupervised (right) deep learning. Green = ground truth, Red = predictions (confidence > 0.8). [NIPS 2017]

Looking Forward

- Data Initiative & NESAP for Data
  - Help experimental efforts transition to KNL and towards exascale
- Continue NESAP work on Cori KNL
  - Transition broad community to manycore through training and web
  - Application portability w/ ANL, ORNL (http://performanceportability.org)
  - Explore ‘exascale’ programming models and languages
  - Influence standards committees (OpenMP, MPI)
- NESAP for NERSC 9 (2020) system when announced

NERSC Systems Timeline

2007/2009	NERSC-5	Franklin	Cray XT4	102/352 TF
2010	NERSC-6	Hopper	Cray XE6	1.28 PF
2014	NERSC-7	Edison	Cray XC30	2.57 PF
2016	NERSC-8	Cori	Cray XC	30 PF
2020	NERSC-9			100PF-300PF
2024	NERSC-10			1EF