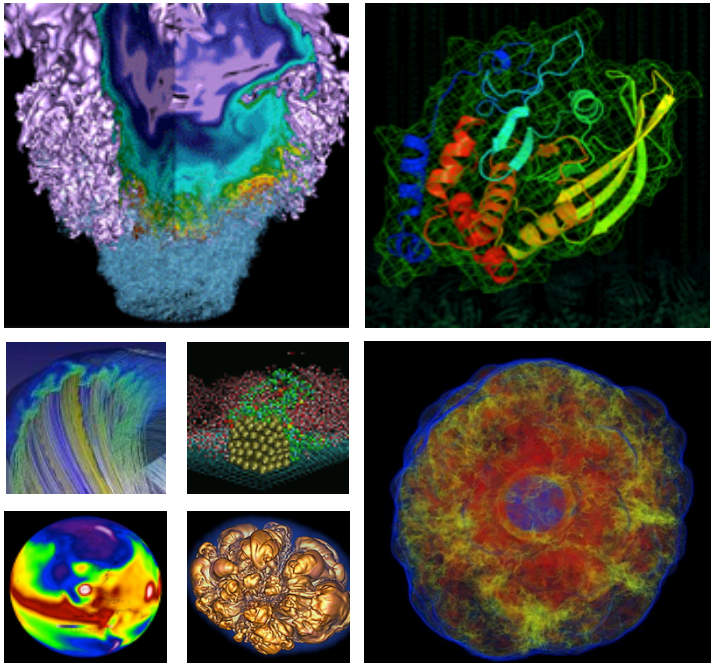
Many-Core for the Masses



Richard Gerber

NERSC Senior Science Advisor
High Performance Computing Department Head

Tsukuba University
University of Tokyo
March 6, 2018

NERSC

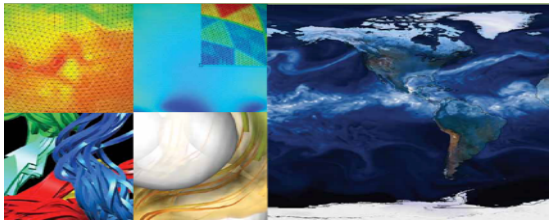
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

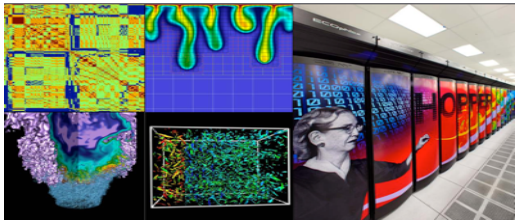
NERSC: Mission HPC for DOE Office of Science

U.S. DEPARTMENT OF ENERGY | Office of Science

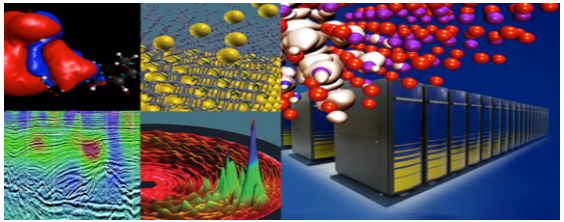
Largest funder of physical science research in U.S.



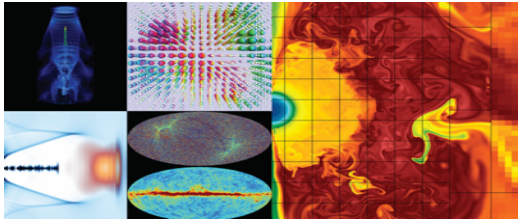
Bio Energy, Environment



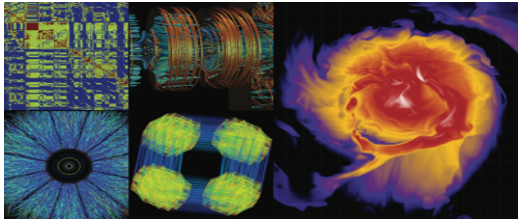
Computing



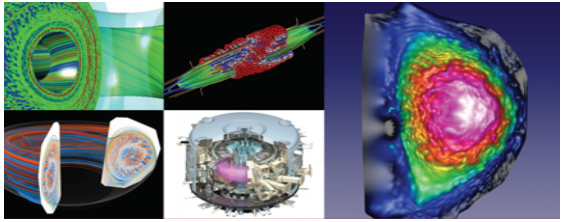
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

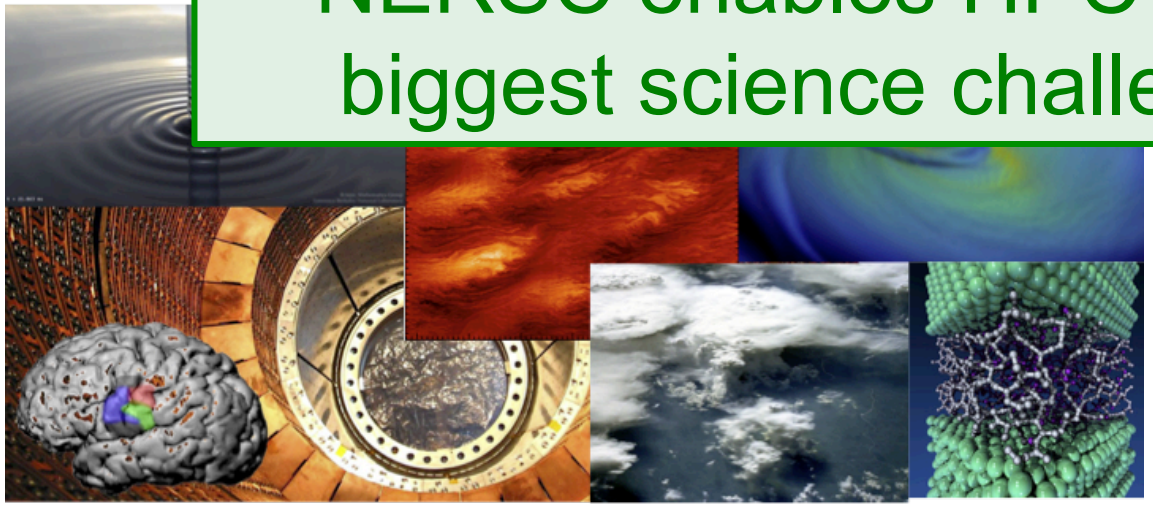
7,000 users, 800 projects, 700 codes, 48 states, 40 countries, universities & national labs

Focus on Science

NERSC users produce publish more than any other center in the world*; ~2,000/year

1,300 cit

NERSC enables HPC for the biggest science challenges.



2017

nature
International weekly journal of science

ature
lature Comm.
12 journals

Science
AAAS

9 in Science

15 in PNAS

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

6 Nobel-prize winning users

U.S. DEPARTMENT OF ENERGY | Office of Science

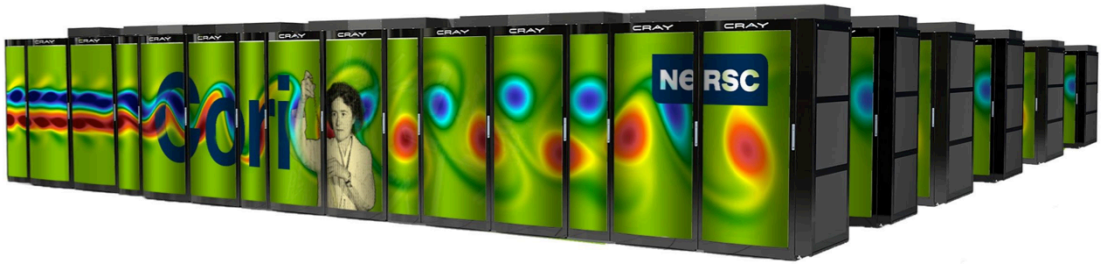
* as far as I can tell

BERKELEY LAB

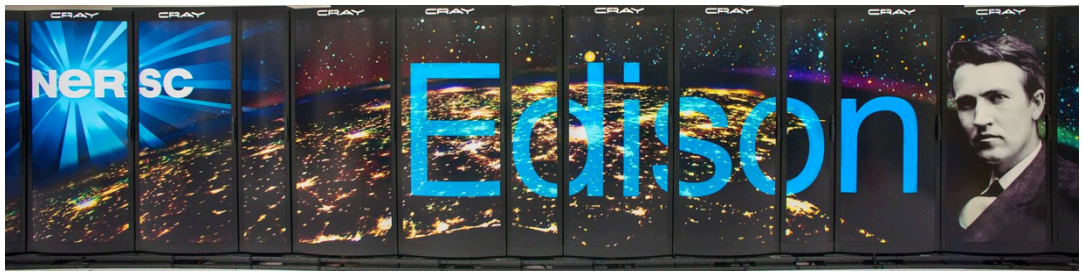
High Performance Computing Systems

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, 1.7 TB/sec
Haswell: ~1 B NHrs/yr; KNL: ~6 B NHrs/yr



Cori #8 on November 2017 Top 500 list
Oakforest-PACS #9



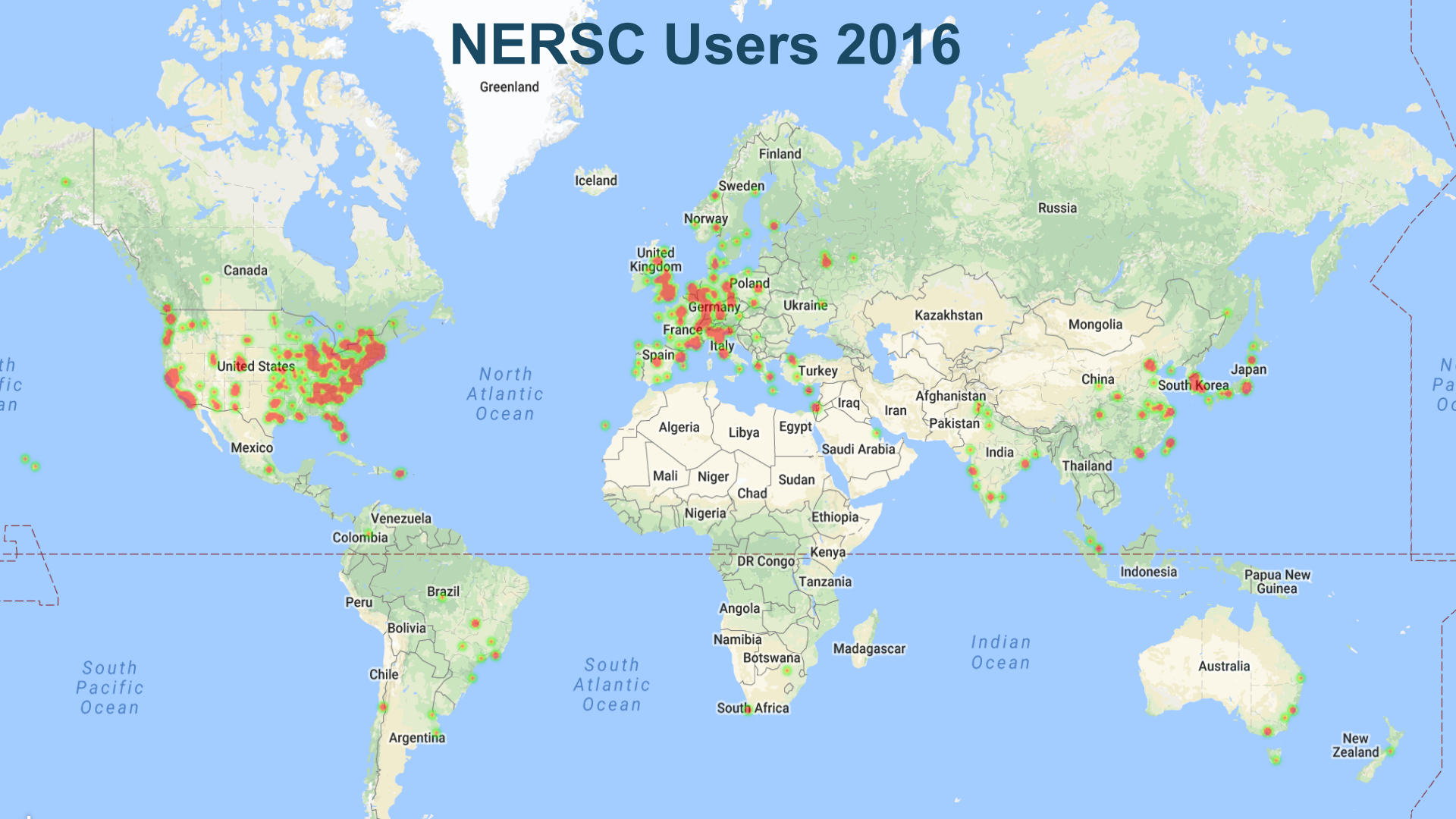
Edison

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
Edison: ~2 B NHrs/yr

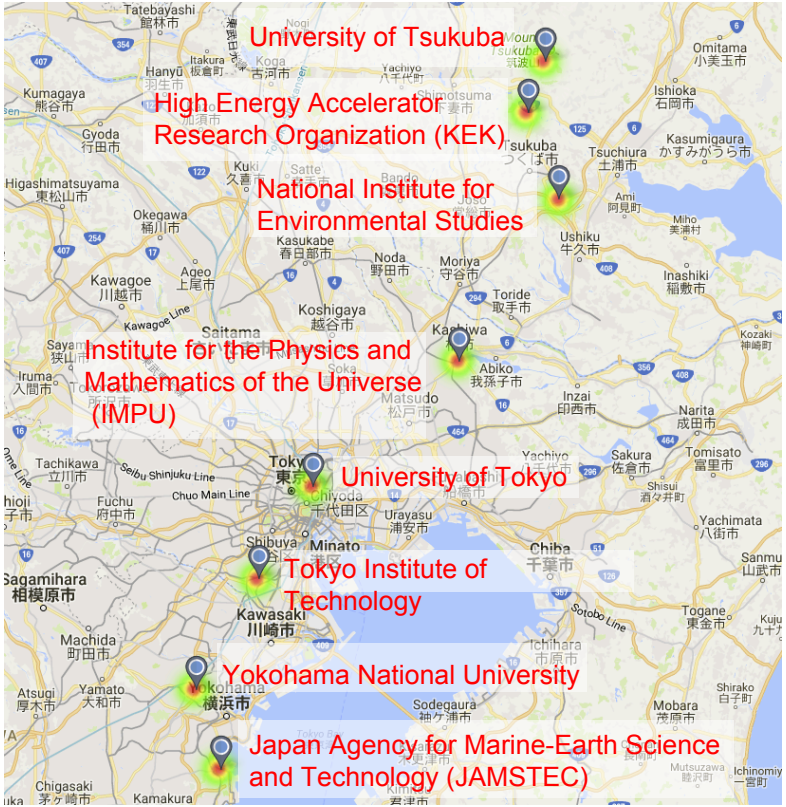
Allocation of Time at NERSC

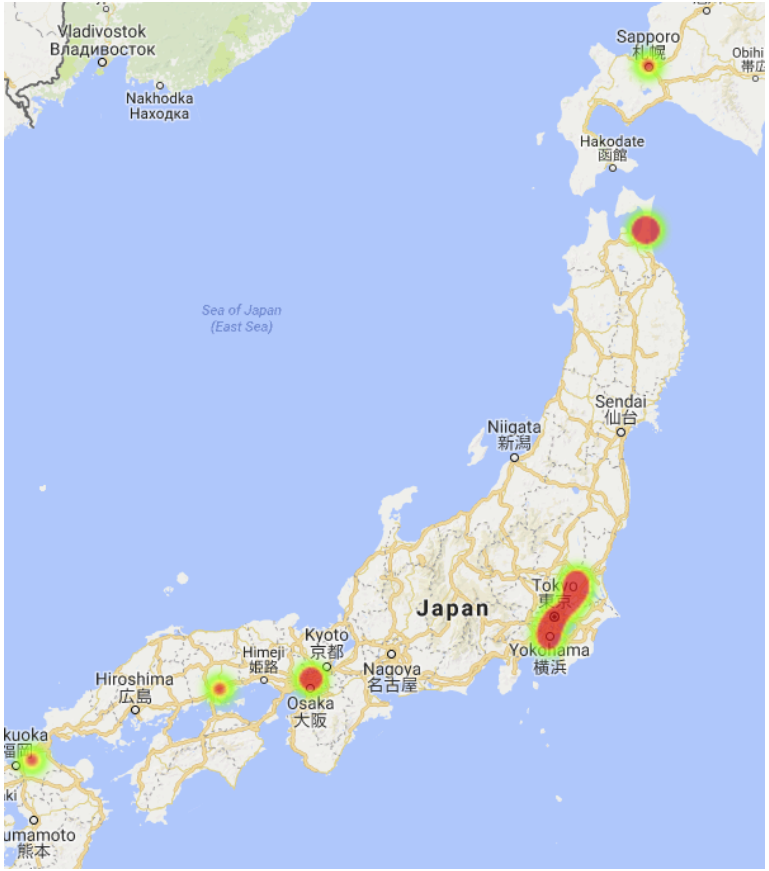
- **DOE Mission Science 80%**
  Distributed by DOE Office of Science program managers
- **ALCC 10%**
  Competitive awards run by DOE Advanced Scientific Computing Research Office
- **Directors Discretionary 10%**
  Strategic awards from NERSC

NERSC Users 2016



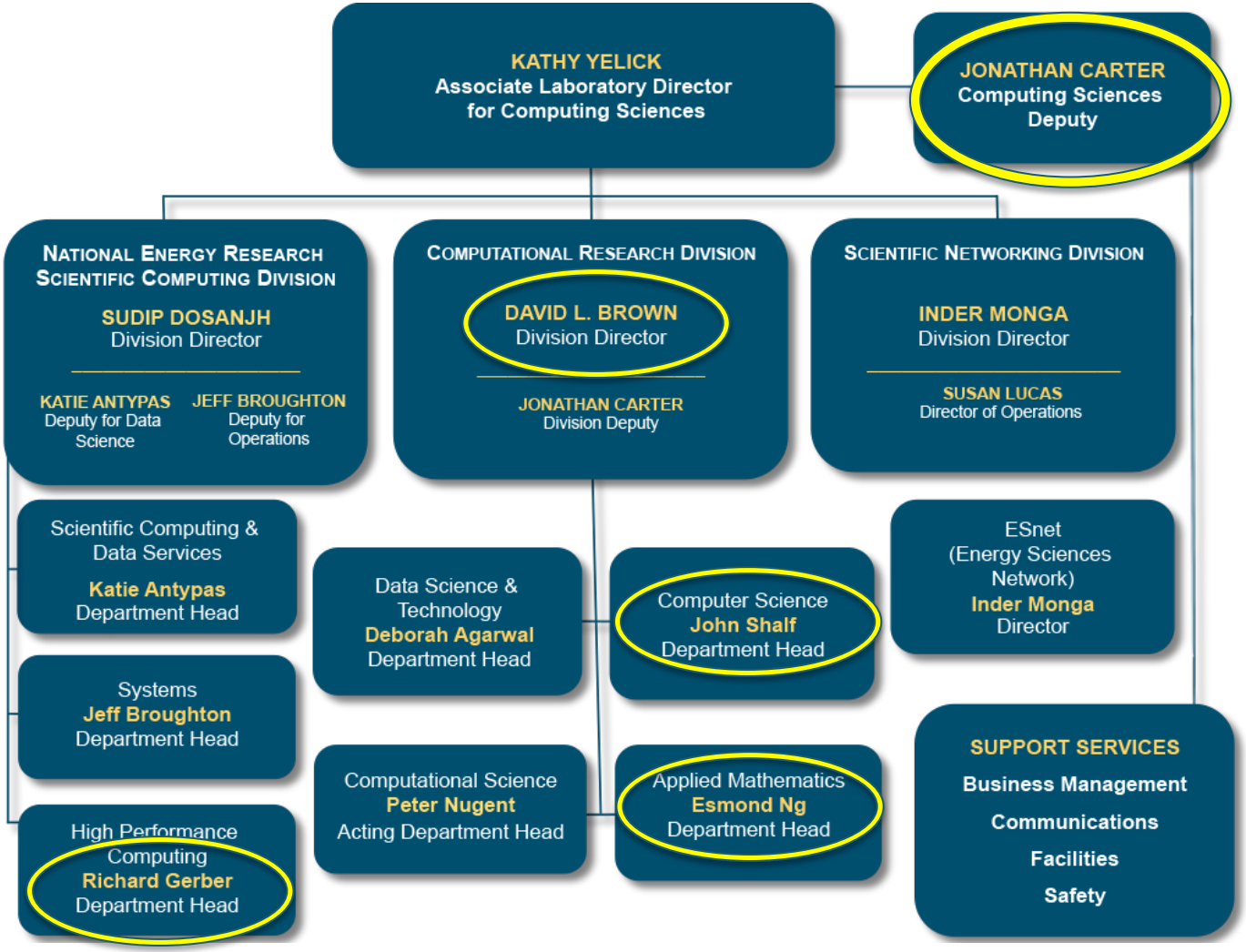
NERSC Users - Japan (34)





U.S. DEPARTMENT OF ENERGY | Office of Science

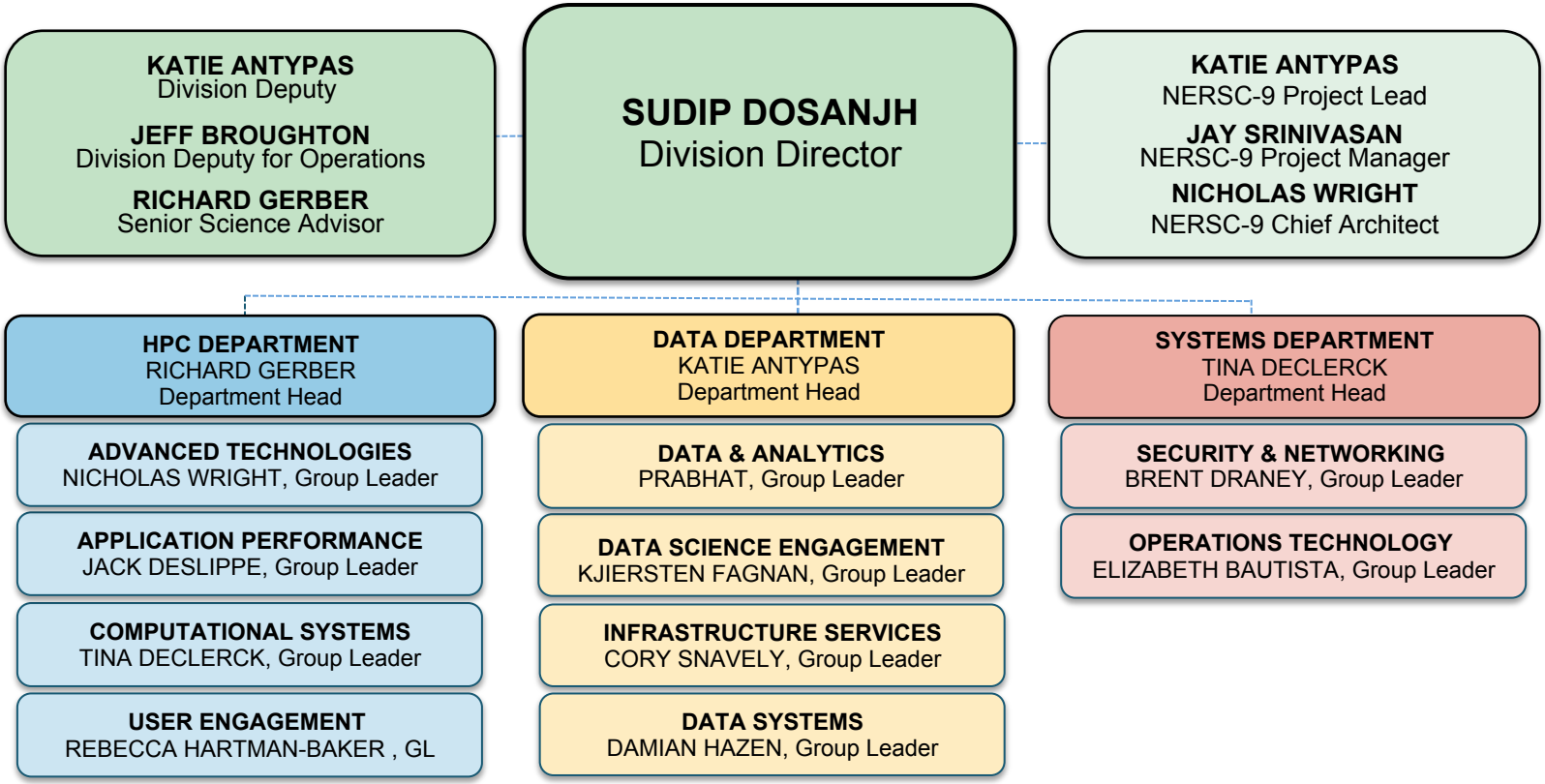
BERKELEY LAB



Berkeley Lab

Computing Sciences Area

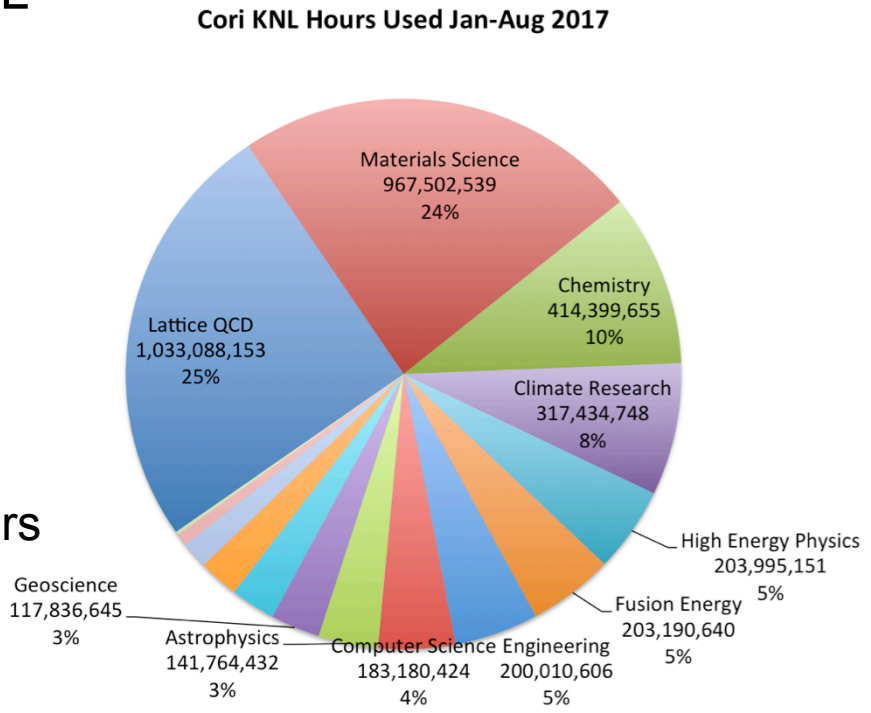
NERSC Organization 2018



Cori KNL Usage Year 1 (2017)

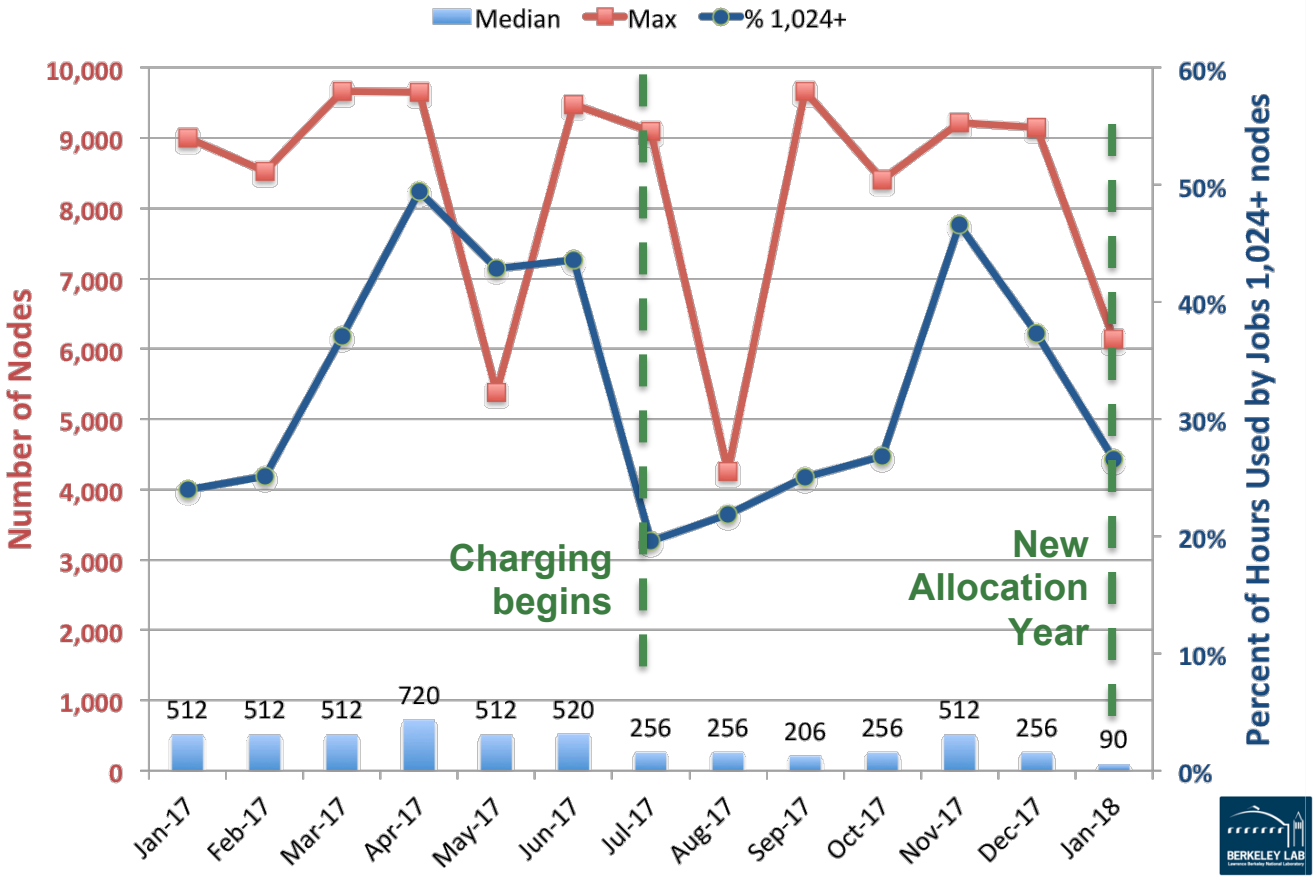
Adoption of KNL has been good; Cori KNL nodes are fully used by researchers

- Open to all users (free): March 2017
- Production (charging): July 2017
- 541 Projects Used KNL time in 2017
- 186 projects used > 1 M NERSC Hours (~10K node hours)
- 270 projects have used > 100 K NERSC Hrs
- 32% of hours used > 1,024 nodes (69K cores)
- 7 Gordon Bell submissions using Cori KNL



Job Scale on Cori KNL

- Jobs run at all scales on Cori
- Larger jobs during free time Jan.-June.
- Users constrained by allocation after July 1
- Large Director's Reserve projects helped big job usage in late 2017

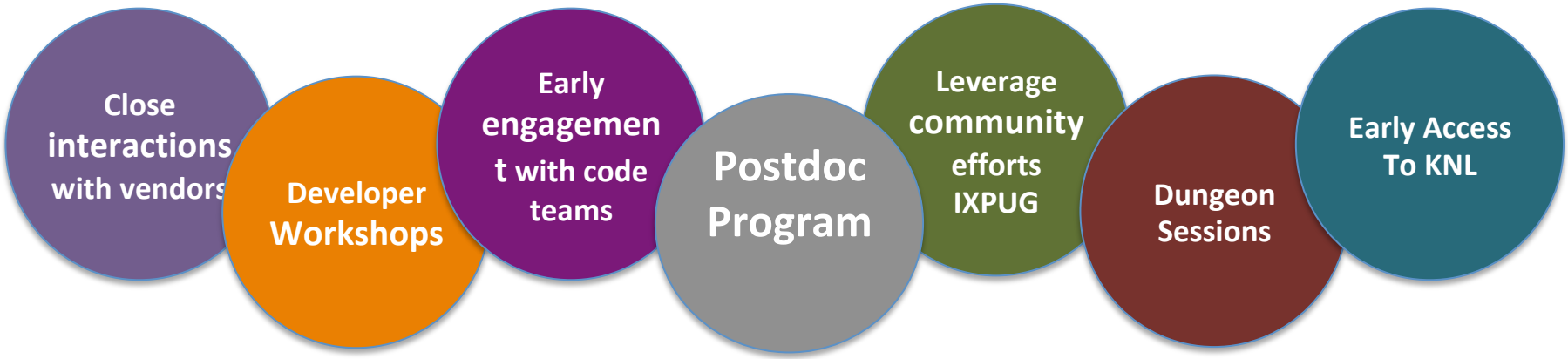


NERSC Exascale Scientific Application Program (NESAP)

Goal: Prepare DOE Office of Science users for many core

Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community.

20 applications cover (or serve as proxies for) > 50% of NERSC hours used



Opportunities & Challenges for Users

Edison ("Ivy Bridge):

- 5576 nodes
- 24 physical cores per node
- 48 virtual cores per node
- 2.4 - 3.2 GHz

- 8 double precision ops/cycle

- 64 GB of DDR3 memory
- 2.5 GB per physical core

- ~100 GB/s Memory Bandwidth

Cori ("Knights Landing"):

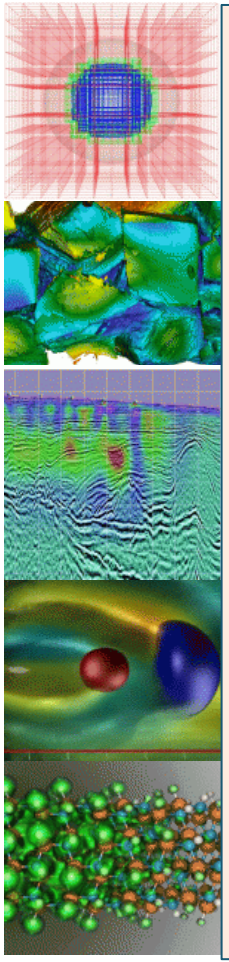
- 9600 nodes
- 68 physical cores per node
- 272 virtual cores per node
- 1.4 - 1.6 GHz

- 32 double precision ops/cycle

- 16 GB of fast memory (.25/core)
- 96GB of DDR4 memory (1.5/core)

- Fast memory has 400 - 500 GB/s
- No L3 Cache

NESAP Codes



Advanced Scientific Computing Research

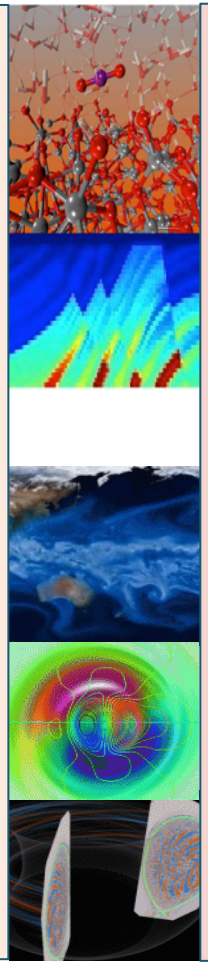
Almgren (LBNL) **BoxLib AMR**
Trebotich (LBNL) **Chombo-crunch**

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**

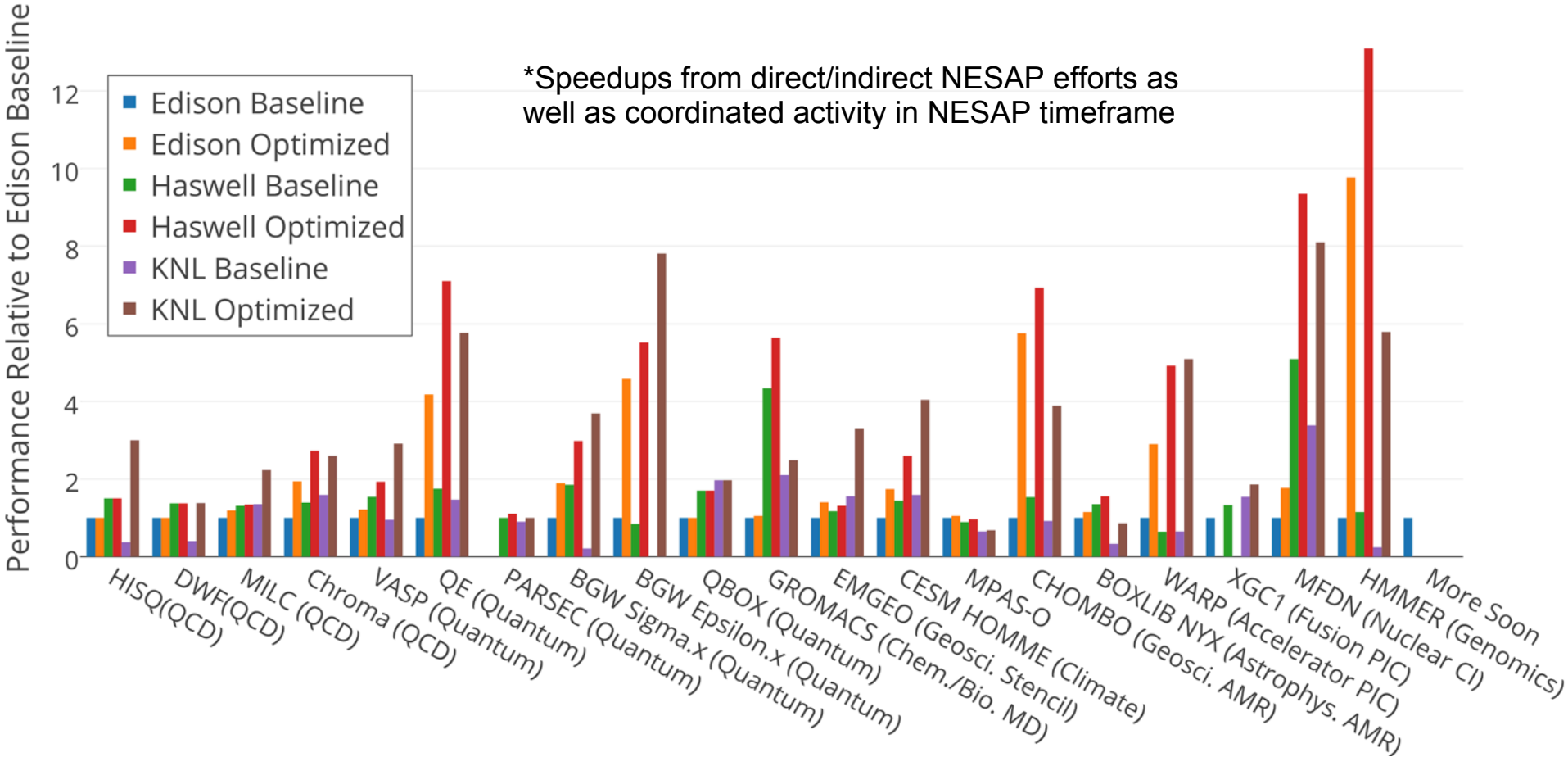
Biological and Environmental Research

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**

NESAP Code Performance on KNL



Preliminary N IL

Code Speedups Via NESAP (per node):

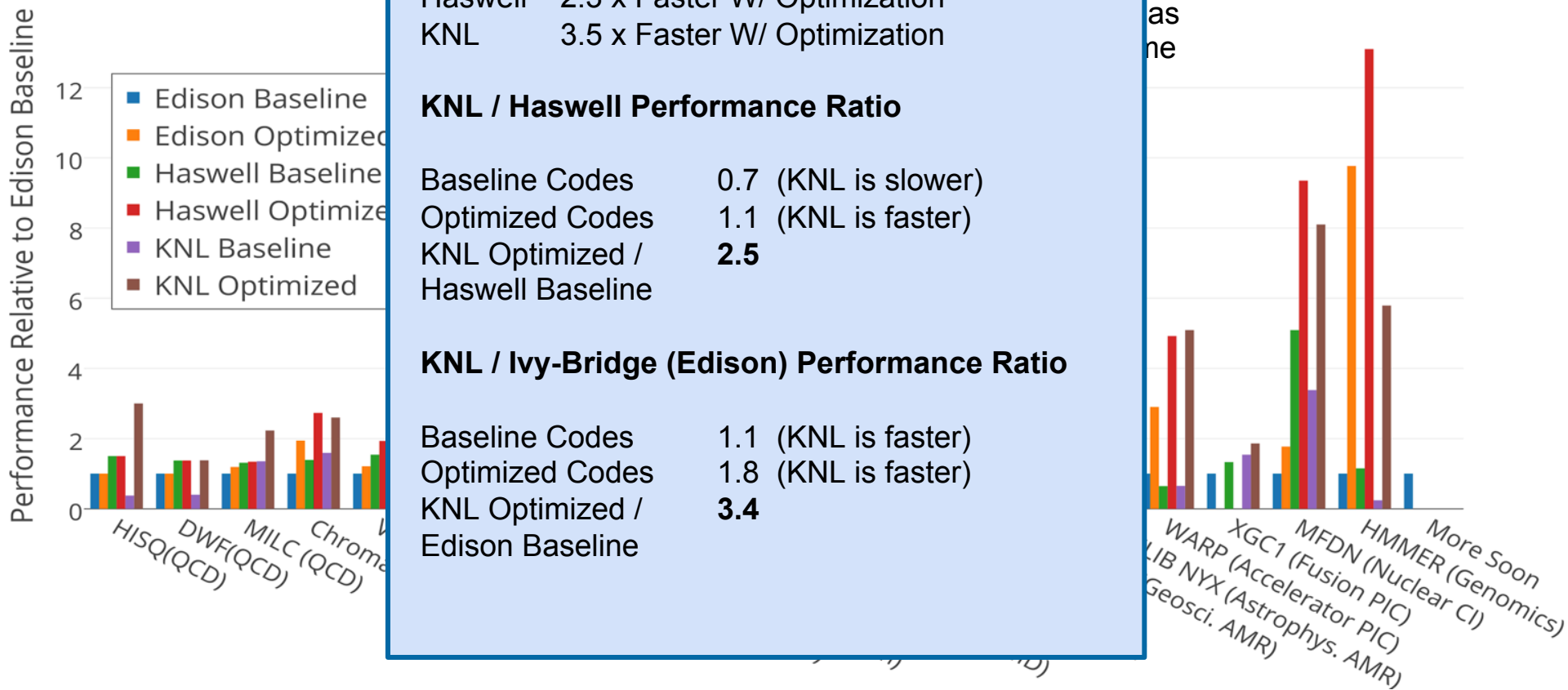
Haswell	2.3 x Faster W/ Optimization
KNL	3.5 x Faster W/ Optimization

KNL / Haswell Performance Ratio

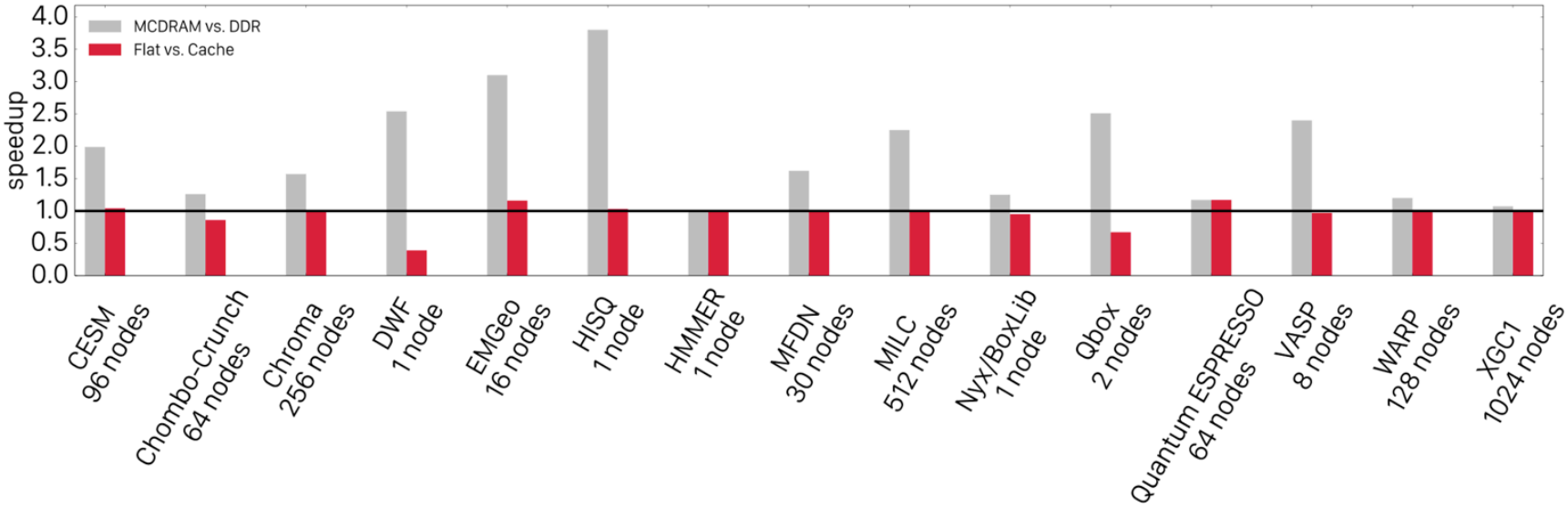
Baseline Codes	0.7	(KNL is slower)
Optimized Codes	1.1	(KNL is faster)
KNL Optimized / Haswell Baseline	**2.5**	

KNL / Ivy-Bridge (Edison) Performance Ratio

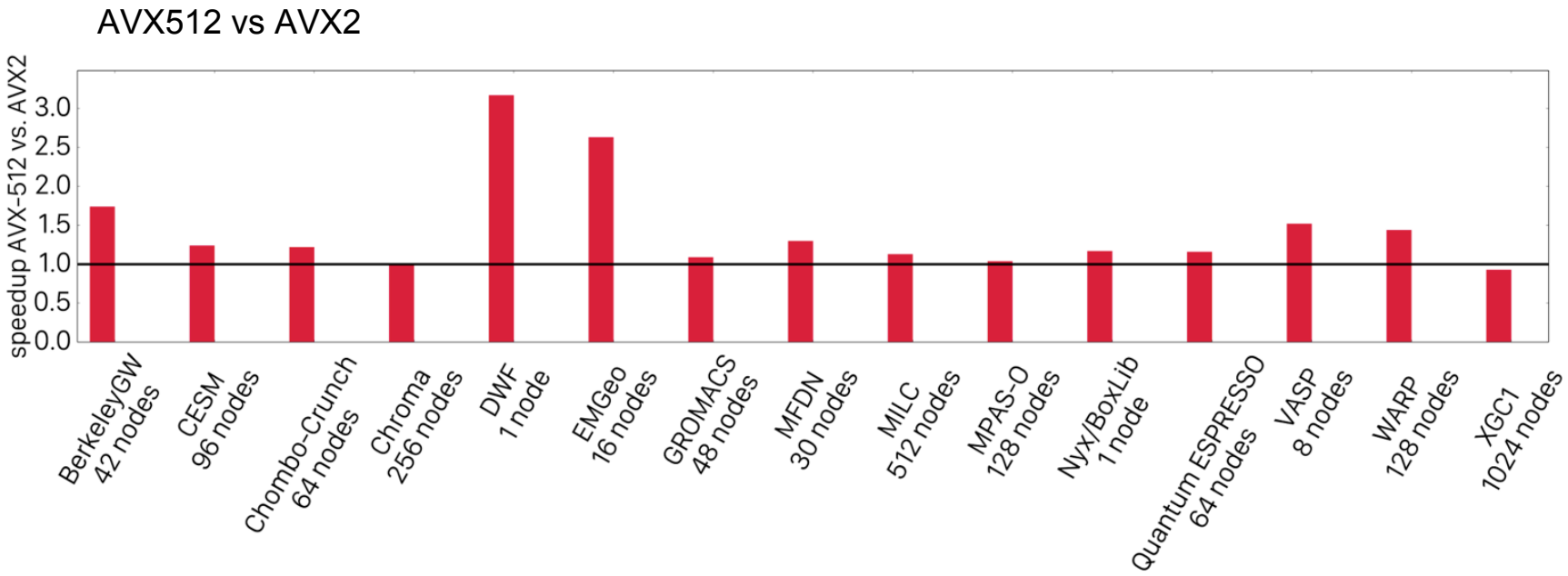
Baseline Codes	1.1	(KNL is faster)
Optimized Codes	1.8	(KNL is faster)
KNL Optimized / Edison Baseline	**3.4**	



NESAP MCDRAM Effects



NESAP VPU Effects



Some Things We Learned

- It is crucial to understand what limits performance. Tools like Advisor are necessary.
- To get good performance on KNL one needs good task/thread scaling and
  - a) efficient vectorization (Codes with high Arithmetic Intensity)
  - b) efficient use of MCDRAM (Codes with low AI)
  - c) both (Codes with AI near 1)
- The lack of an L3 cache can make cache blocking for L1/L2 important.
- Cache mode provides nearly the same performance as flat mode for most applications.
- MPI scaling is similar at the same number of ranks on Xeon and KNL. This translates to lower node counts on Xeon-Phi. Additional fine-grained parallelism needs to be exploited to take advantage of a large system like Cori.

Not All Projects Have Adopted KNL

Some legacy code perform poorly on KNL

Average performance of our "unoptimized well written" codes on KNL vs. Haswell dual socket: 70%

Some single-threaded codes see 500% slowdown

Projects that use >50% of their time on Cori KNL

2017

- 25% all users
- 37% of "big" users

2018 to date

- 20% of all users
- 30% of "big" users
- Goal: 50% of all users

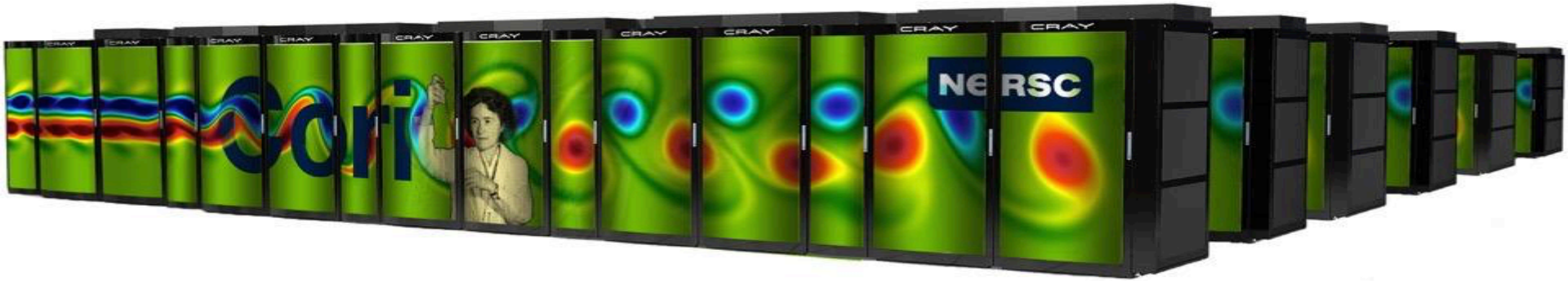
Moving the Community Toward Exascale

- Part of NERSC's mission is to help prepare the entire Office of Science workload for exascale
- Toward that end our goal is to have 50% of projects using KNL for >50% of their computing at NERSC.
- Training and workshops; web case studies; individual outreach
- Providing computer time for code development along with NERSC assistance
- We are also exploring languages, programming models, libraries, frameworks and working with standards committees, tool vendors and ECP to make advanced architectures useful to the broad community

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**		Selection underway	**~100-150 PF**
2024	**NERSC-10**			**1EF**

Edison is currently scheduled to retire in ~March 2019 (subject to change)

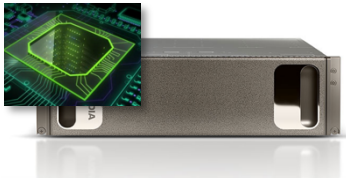


NERSC-9 System 2020

1. Provide 3-4x capability of Cori
2. Meet needs simulation and data analysis use cases including:
   a. Complex workflows
   b. Analytics and machine learning at scale
   c. Support for experimental facilities workflows
3. Prepare users for exascale and more specialization and heterogeneity

System will be announced in 2018

Proliferation of accelerators is altering the market.

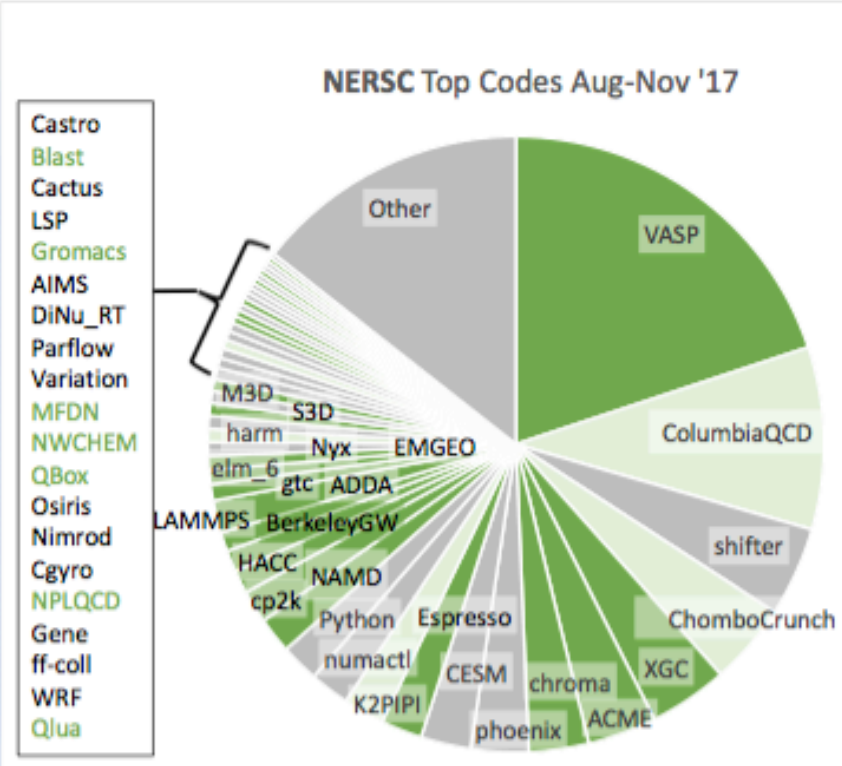


NVIDIA builds deep learning appliance

Microsoft and Baidu deploy FPGA's
Google designs own Tensor Processing Unit (TPU)

Intel buys
deep learning startup, Nervana
FPGA company, Altera

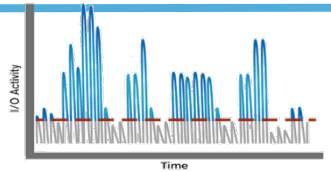
NERSC is assessing the GPU readiness of the workload



GPU Status	Description	Fraction
Enabled	At least some kernels GPU accelerated.	46%
Proxy	Port of algorithmically similar code exists.	19%
Unlikely	Code is not amenable to GPU acceleration.	25%
Unknown	Readiness cannot be assessed at this time.	11%

NERSC is addressing data 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

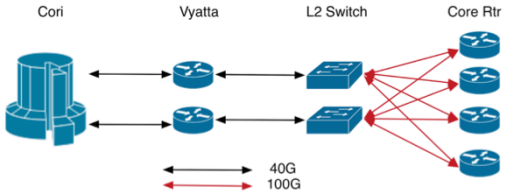
I need real-time feedback for my workflow

SchedMD

Real-time queues

Internal network limits how I can import data to supercomputer

SDN



There is limited software for analytics on HPC systems

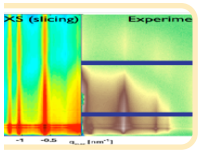
New analytics and ML libraries

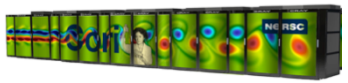
Caffe scikit learn Spark K TensorFlow

U.S. DEPARTMENT OF ENERGY | Office of Science

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

- Deploying large scale computing and storage resources
- Providing reusable building blocks for experimental scientists to build pipelines
- Providing scalable infrastructure to launch services
- Expertise on how to optimize pipelines





Summary

- Cori KNL is running well and is being used productively by a the DOE Office of Science workload
- But, there is still more work to do to get the majority NERSC users on Cori KNL and on the path to exascale
- NERSC 9 will be coming in 2020 and the architecture will be announced in 2018
- NESAP 2 will begin as soon as NERSC 9 is announced
- NERSC is targeting Cori and NERSC 9 for data-intensive workloads