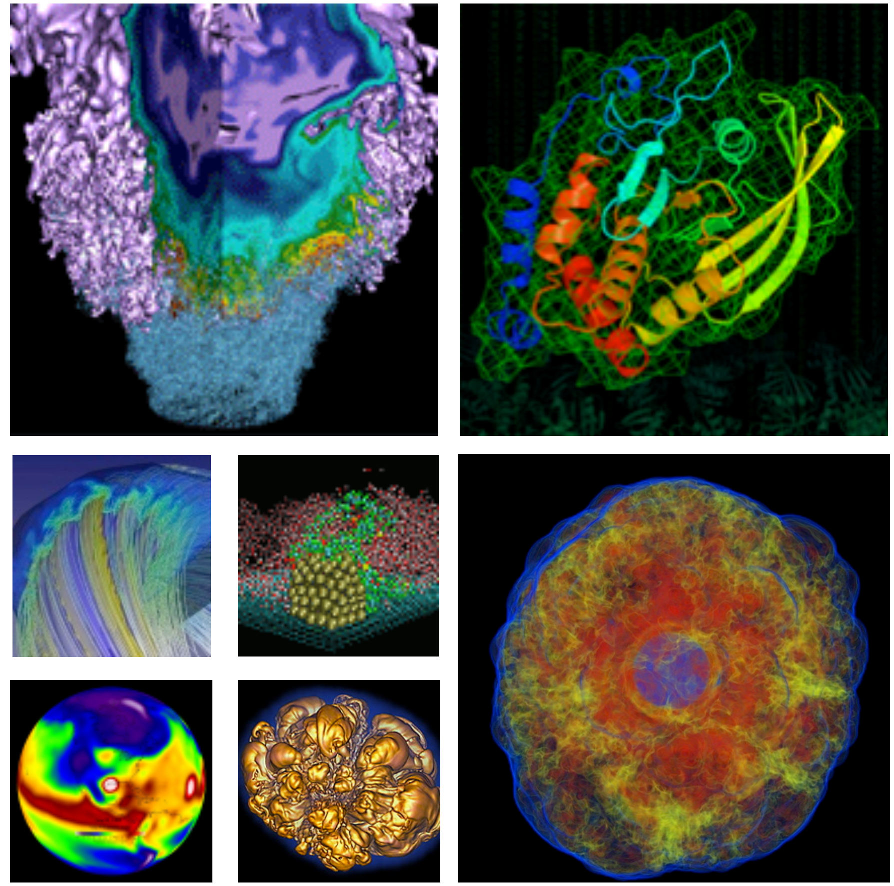
Application Preparedness for Next Generation Computational Systems and Integration with Data-Intensive Workflows



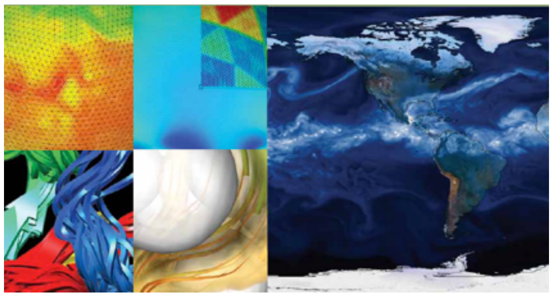
Richard Gerber
NERSC Senior Science Advisor
HPC Department Head

February 27, 2017

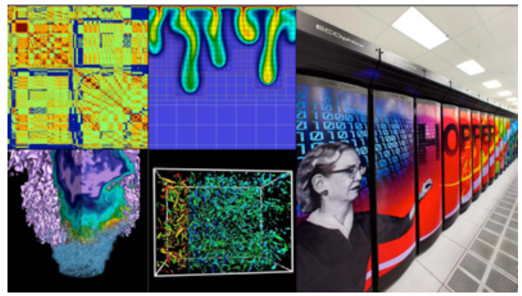
NERSC Provides Mission HPC and Data Resources for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

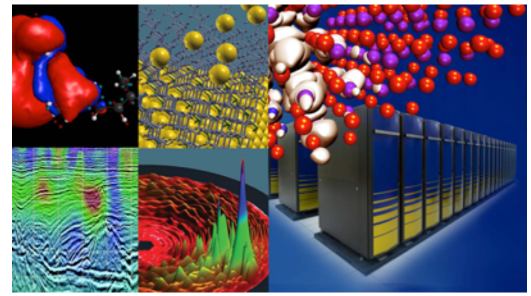
Largest funder of physical science research in U.S.



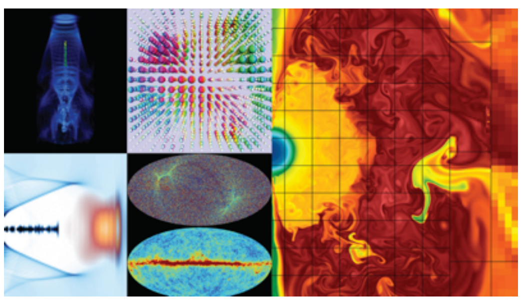
Bio Energy, Environment



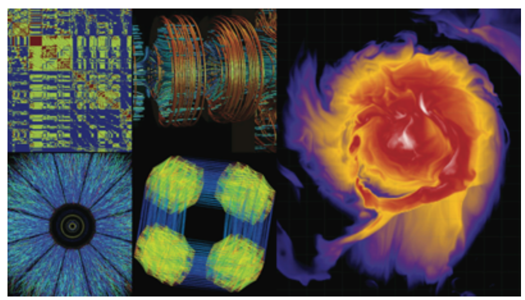
Computing



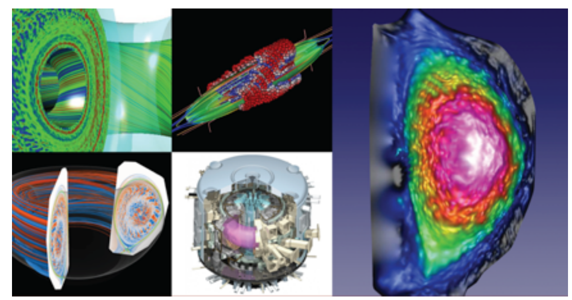
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

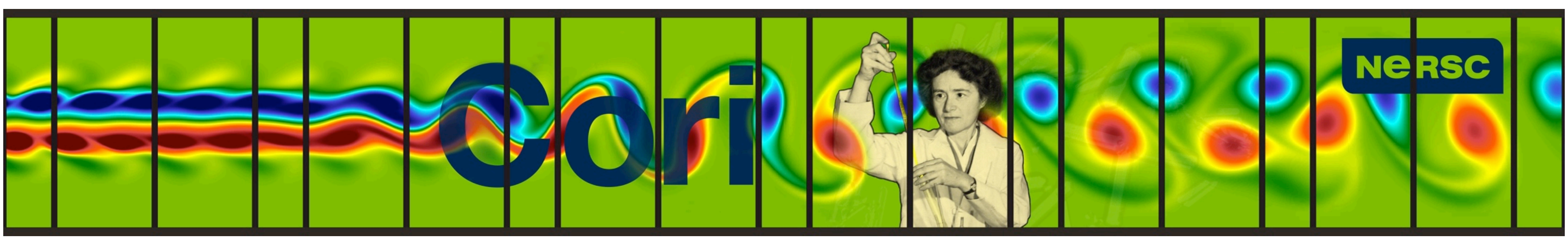
Focus on Science

- **NERSC supports the broad mission needs of the six DOE Office of Science program offices**
- **6,000 users and 750 projects**
- **Supercomputing and data users**
- **NERSC science engagement team provides outreach and POCs**

2,078 refereed publications in 2015

The NERSC-8 System: Cori

- **Cori will support the broad Office of Science research community and begin to transition the workload to more energy efficient architectures**
- **Cray XC system with over 9,300 Intel Knights Landing compute nodes – mid 2016**
  – Self-hosted, (not an accelerator) manycore processor with up to 72 cores per node
  – On-package high-bandwidth memory
- **Data Intensive Science Support**
  – 10 Haswell processor cabinets (Phase 1) to support data intensive applications – Summer 2015
  – NVRAM Burst Buffer to accelerate data intensive applications, 1.5 PB, 1.5 TB/sec
  – 28 PB of disk, >700 GB/sec I/O bandwidth
- **Robust Application Readiness Plan**
  – Outreach and training for user community
  – Application deep dives with Intel and Cray
  – 8 post-docs integrated with key application teams

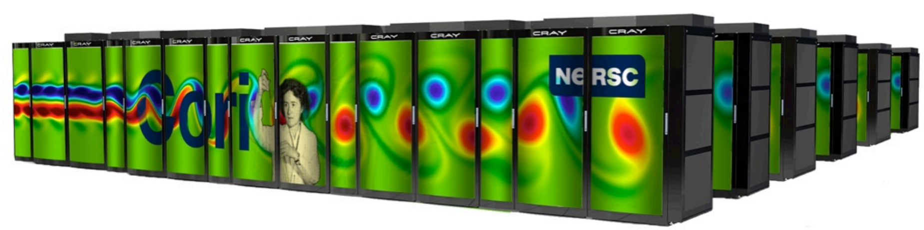


Intel "Knights Landing" Processor

- **Next generation Xeon-Phi, >3TF peak (3X/thread over KNC)**
  - Single socket processor - Self-hosted, not a co-processor, not an accelerator
  - Up to 72 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
  - 512b vector units (32 flops/clock – AVX 512)
  - High bandwidth on-package memory (16GB) 5X bandwidth of off-package DDR4 DRAM
- **Presents an application porting challenge to efficiently exploit KNL performance features**

Cori Phase 1

- **Running with all NERSC users in production mode**
- **1,630 Compute Nodes (52,160 cores)**
  - Two Haswell processors/node
  - 16 cores/processor at 2.3 GHz
  - 128 GB DDR4 2133 MHz memory/ node
- **Cray Aries high-speed "dragonfly" topology interconnect**
- **22 login nodes for advanced workflows and analytics**
- **SLURM batch system**
- **Lustre File system**
  - 28 PB capacity, >700 GB/sec peak performance



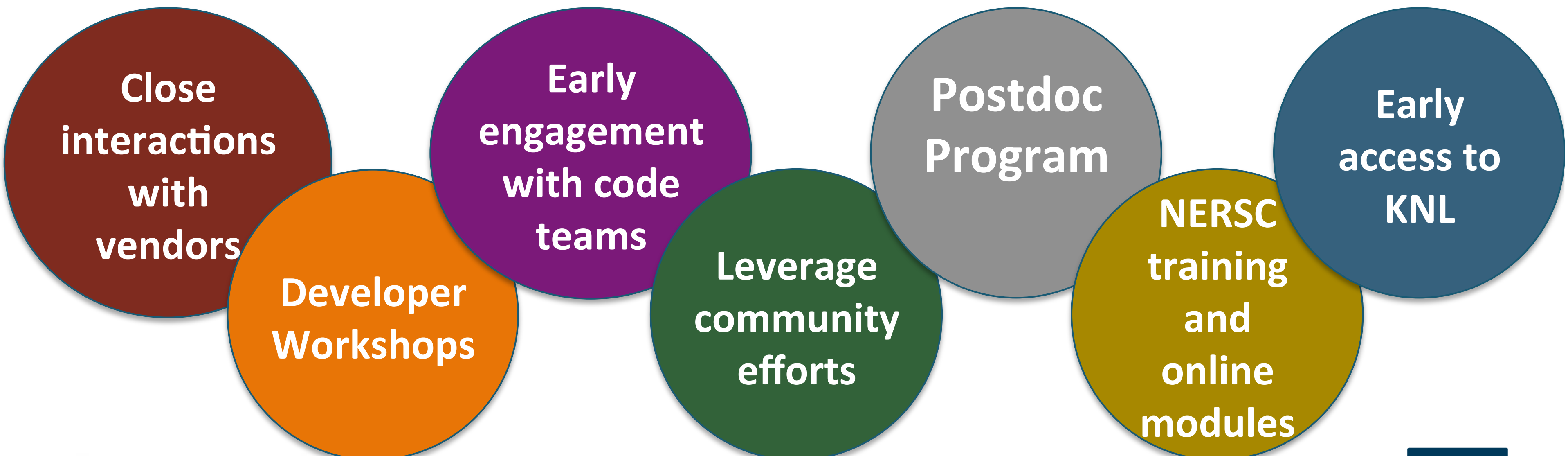
NERSC's Current Big System is Edison

- Edison is the HPCS* demo system (serial #1)
- First Cray Petascale system with Intel processors (Ivy Bridge), Aries interconnect and Dragonfly topology
- Very high memory bandwidth (100 GB/s per node), interconnect bandwidth and bisection bandwidth
- 5,576 nodes, 133K cores, 64 GB/node
- Exceptional application performance

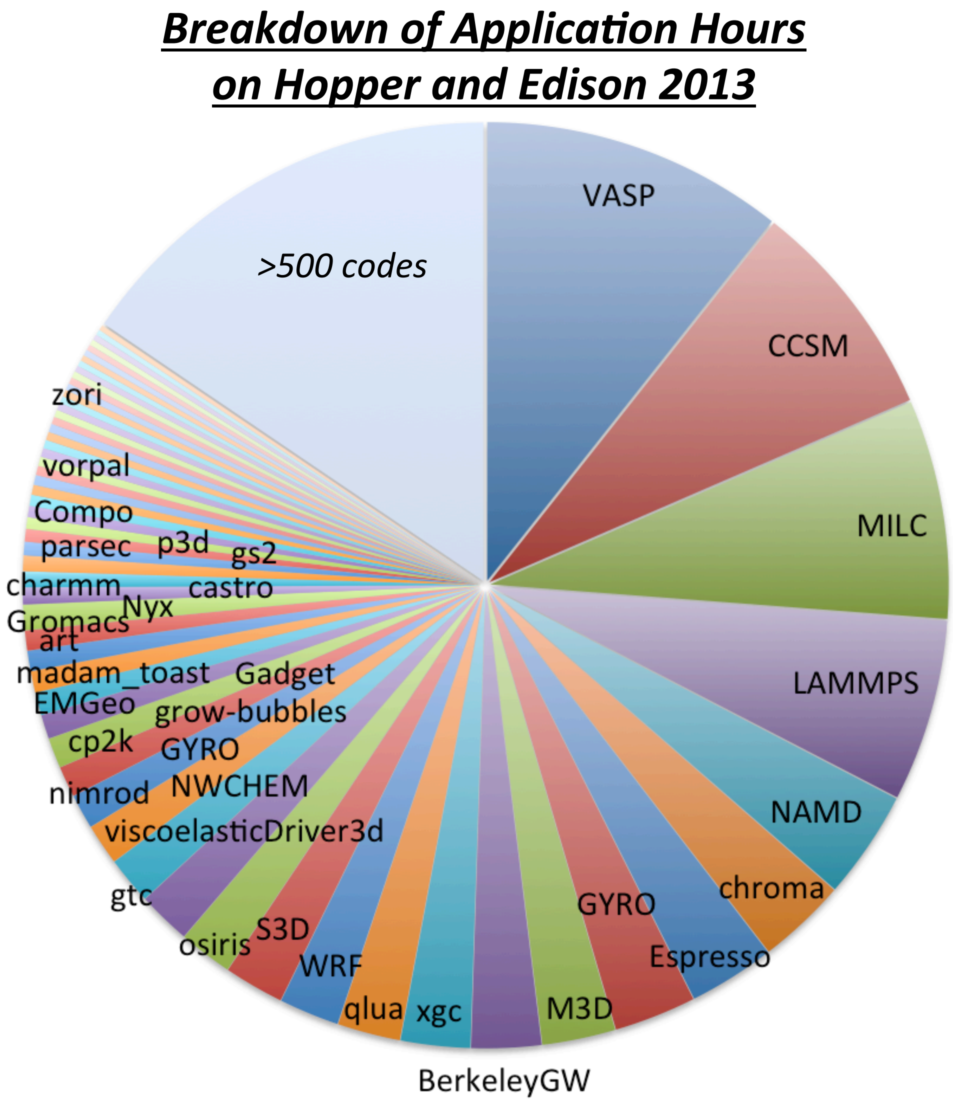
*DARPA High Productivity Computing System program

NERSC Exascale Science Application Program

- Goal: Prepare DOE Office of Science user community for Cori manycore architecture
- Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community
- NESAP activities include:



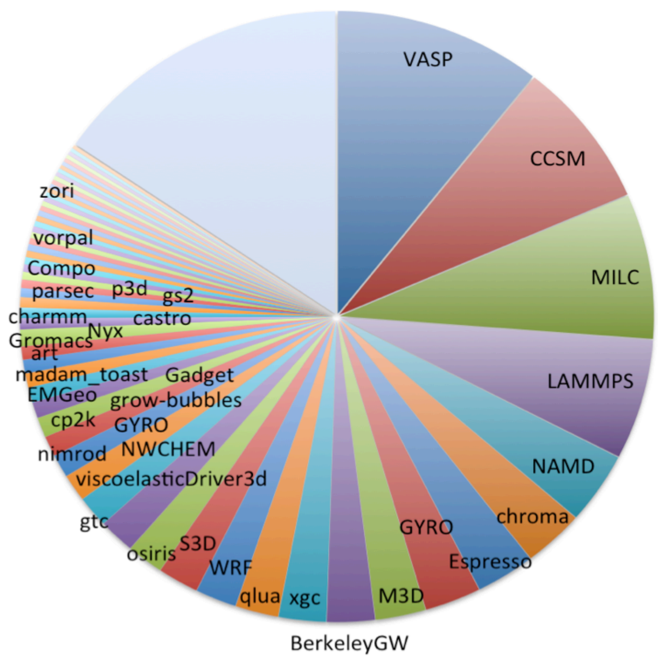
We are initially focusing on 20 codes

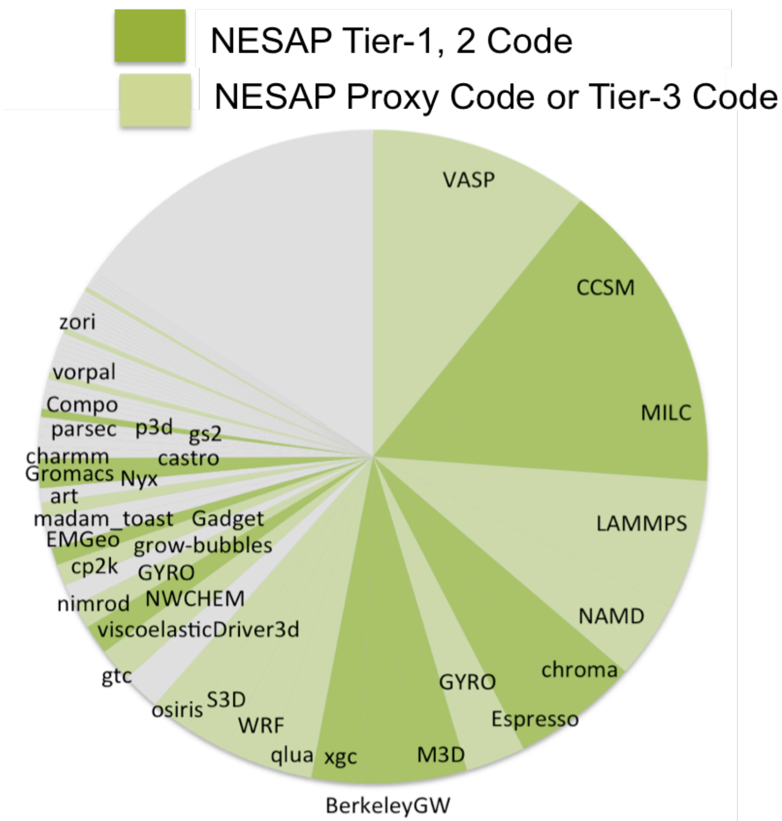


- 10 codes make up 50% of the workload
- 25 codes make up 66% of the workload
- Edison will be available until 2019
- Training and lessons learned will be made available to all application teams

Code Coverage

Breakdown of Application Hours on Hopper and Edison 2013

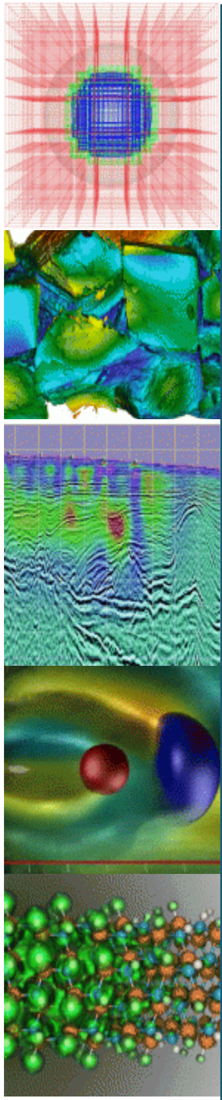




U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NESAP Codes



Advanced Scientific Computing Research

Almgren (LBNL) **BoxLib AMR Framework**

Trebotich (LBNL) **Chombo-crunch**

High Energy Physics

Vay (LBNL) **WARP & IMPACT**

Toussaint(Arizona) **MILC**

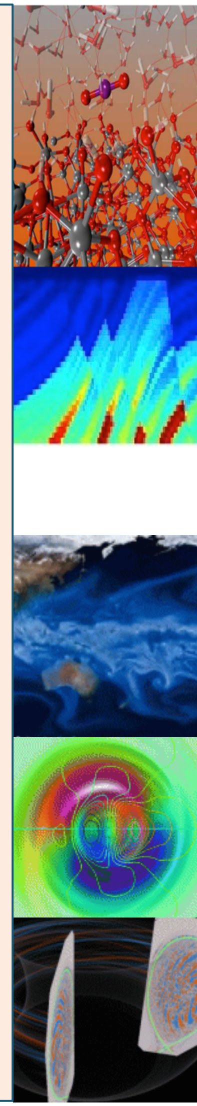
Habib (ANL) **HACC**

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**

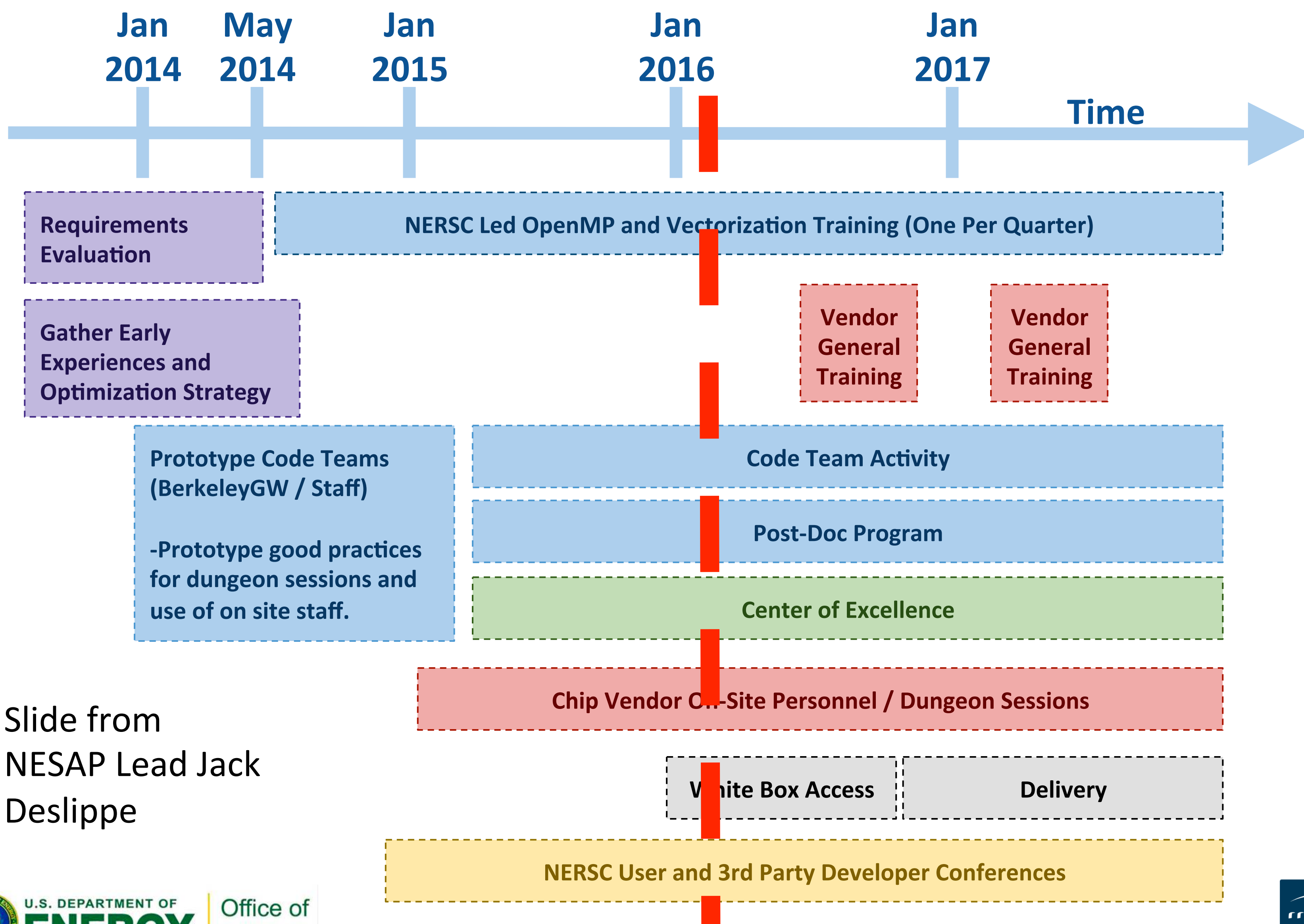
Resources for Code Teams

- **Early access to hardware**
  - Access to Babbage (KNC cluster) and early “white box” test systems expected in early 2016
  - Early access and significant time on the full Cori system
- **Technical deep dives**
  - Access to Cray and Intel staff on-site staff for application optimization and performance analysis
  - Multi-day deep dive (‘dungeon’ session) with Intel staff at Oregon Campus to examine specific optimization issues
- **User Training Sessions**
  - From NERSC, Cray and Intel staff on OpenMP, vectorization, application profiling
  - Knights Landing architectural briefings from Intel
- **NERSC Staff as Code Team Laisons (Hands on assistance)**
  - New Application Performance Group
- **Postdocs (6 of 8 hired) embedded with Tier 1 projects**

Intel Xeon Phi User Group (IXPUG)

- **NERSC hosted IXPUG 2015 in Sept. at the CRT facility**
- **Over 100 attendees**
- **Week long community event with training sessions, hackathons and technical briefings and community meetings**
- **DFT for Exascale community workshop on last day**

NESAP Timeline



Slide from NESAP Lead Jack Deslippe

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NESAP Code Status

Advanced (waiting for hardware)

- Chroma
- DWF
- Gromacs
- BerkeleyGW
- MILC
- HACC

Lots of Progress

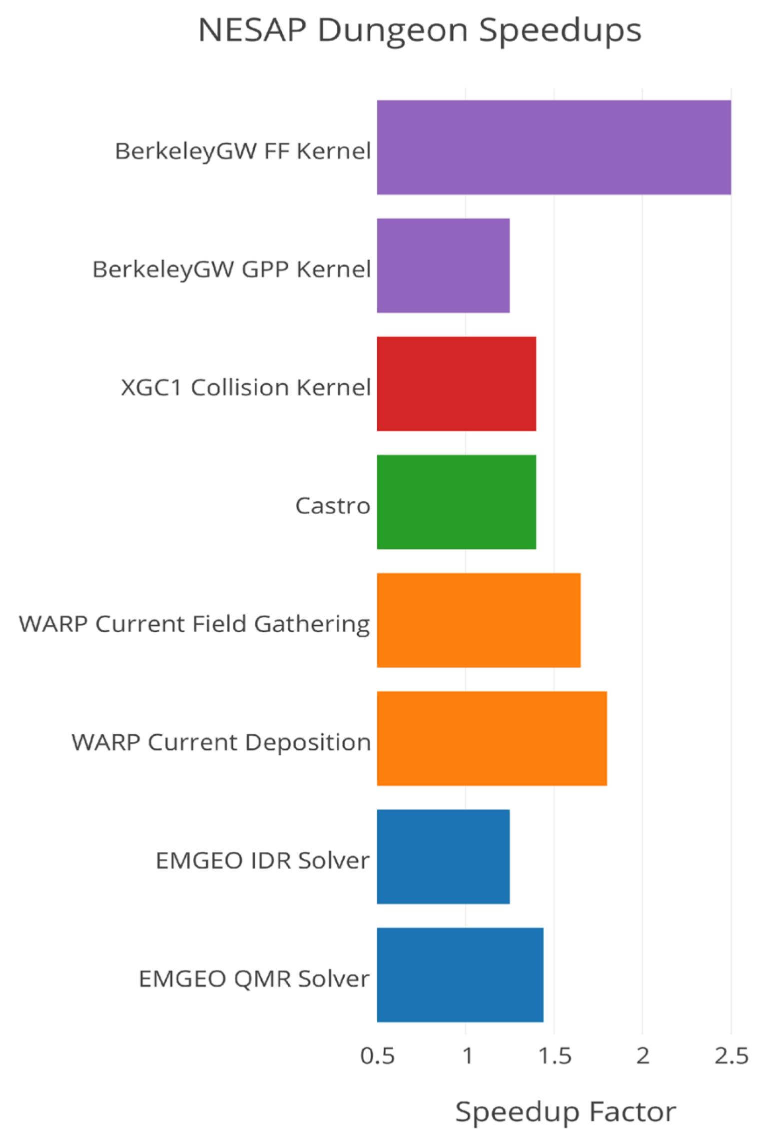
- EMGEO
- Boxlib
- WARP
- XGC1
- VASP
- ESPRESSO

Moving

- PARSEC
- Chombo
- MFDN
- Meraculous
- NWChem

Need Lots of Work

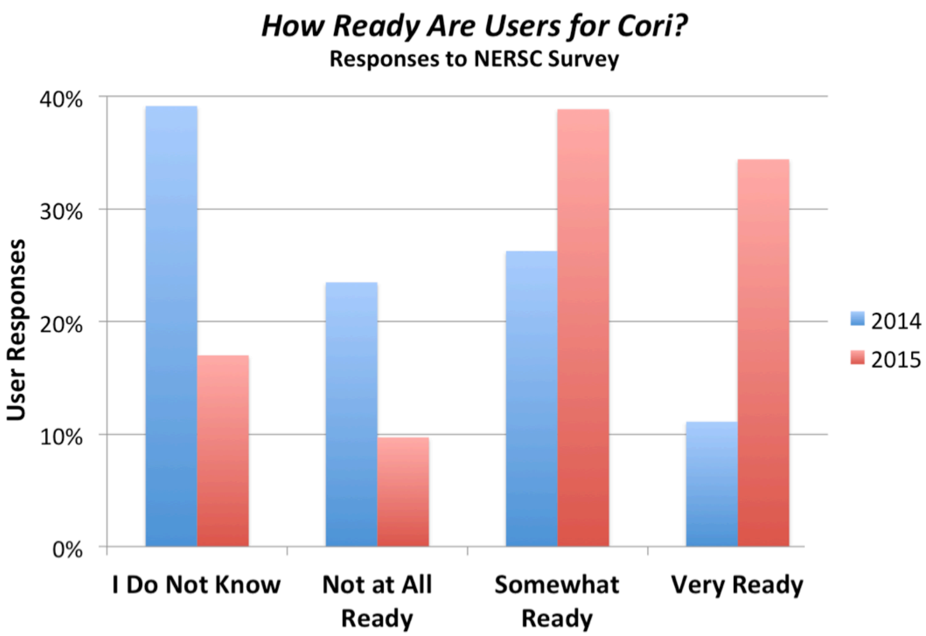
- CESM
- ACME
- MPAS

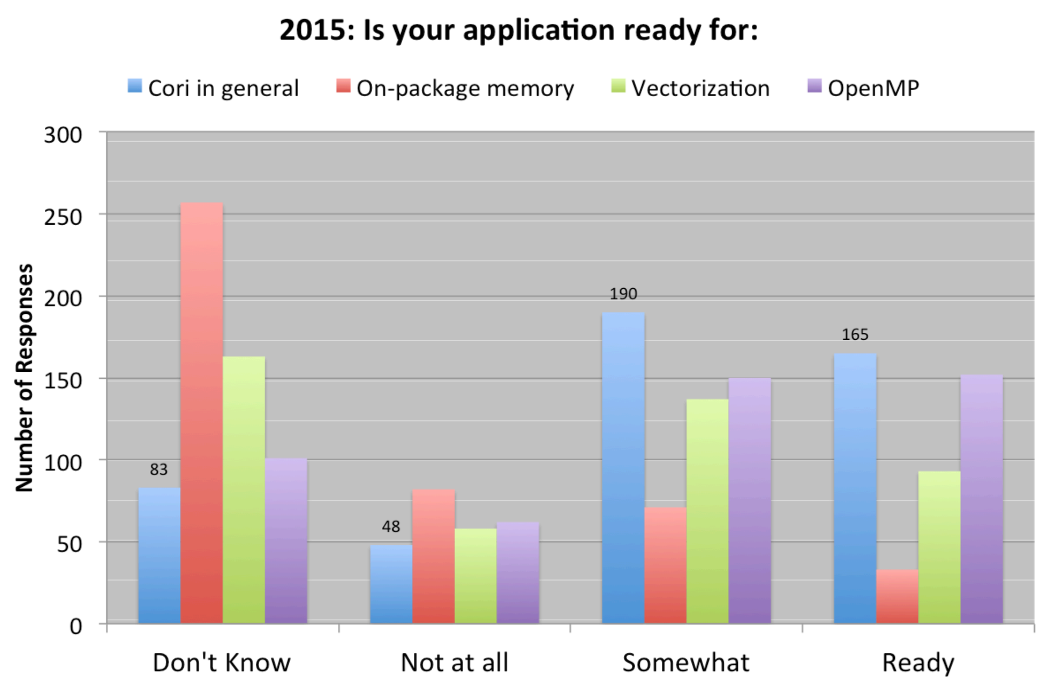


What has gone well

- Setting requirements for Dungeon Session (Dungeon Session Worksheet).
- Engagement with IXPUG and user communities (DFT, Accelerator Design for Exascale Workshop at CRT)
- Learned a massive amount about tools and architecture
- Large number of NERSC and vendor training events (Vectorization, OpenMP, Tools/Compilers)
- Cray COE VERY helpful to work with. Very pro-active.
- Pipelining code work via Cray and Intel experts
- Case studies on the web to transfer knowledge to larger community

Have our training sessions, outreach and case studies made a difference?





Users report significant increase in readiness and awareness of Cori architecture

NESAP Plans

- Increase excitement and effort in 2016 with extra training events, on-site hackathons and more dungeon sessions with KNL hardware in first 9 months of year .
- Continue successful Cray+Intel pipelining approach.
- Continue App-Readiness (and post-doc program) as an ongoing center effort through 2025 (exascale).
- Maintain a community database of lessons learned and programming “pearls” for many-core that is searchable by keywords like “vectorization”, “latency”, “stencil” as a standalone portal

Application Portability

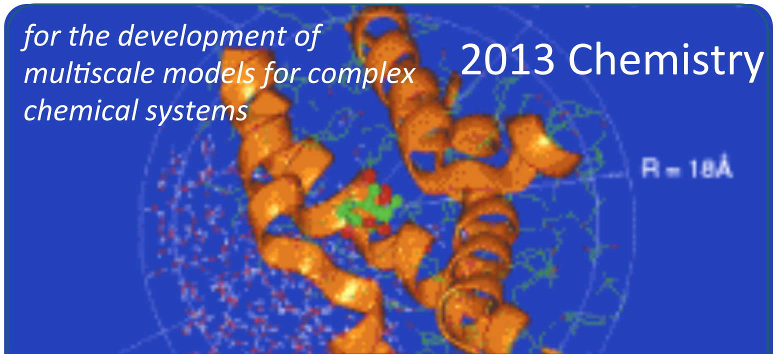
NERSC, OLCF and ALCF have been partnering on application portability and have held a number of workshops in the past 18 months.

Workshop/Meeting	Topic	Date	Location
Application Portability Kick-off meeting	Briefing NERSC, OLCF and ALCF on the other's architectures	Mar. 2014	Oakland, CA NERSC Facility
Application Portability	Programming models for each next generation system. Coordinating NESAP, CAAR and ESP projects	Sept. 2014	Oakland, CA NERSC Facility
Application Portability II	Vendors briefing on tools and programming models for portability (NNSA participants included)	Jan. 2015	Oak Ridge, TN
HPCOR (HPC Operational Review) on Application Performance Portability	Workshop with ~100 participants discussing best practices, emerging practices and opportunities for application performance portability	Sept. 2015	Bethesda, MD
HPC Portability Workshop at SC15	Papers presented on application portability, followed by discussion.	Nov. 2015	SC 15 in Austin, TX

Next steps – New collaboration running from March 2016 – March 2018

- NERSC has new hire identified to work on Application Portability
- Each facility will choose a 'mini-app' and optimize it for their home system
- Then will work with cross facility team, to test portable options on different architecture
- Project timed to correspond with availability of prototype hardware and new testbeds

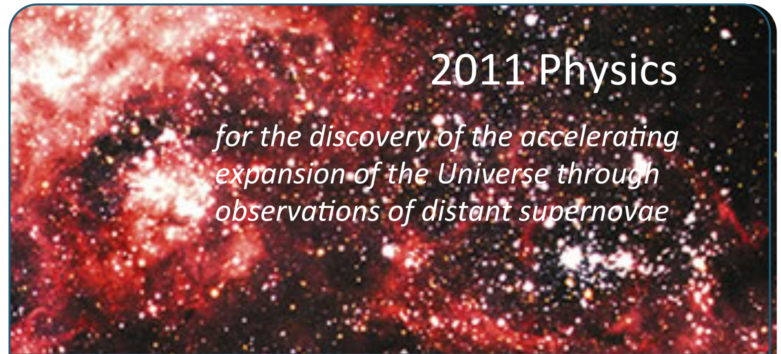
Extreme Data Science Plays Key Role in Scientific Discovery



for the development of multiscale models for complex chemical systems

2013 Chemistry

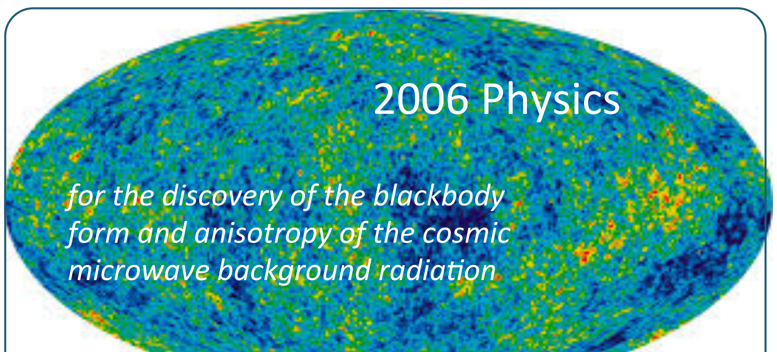
Martin Karplus



2011 Physics

for the discovery of the accelerating expansion of the Universe through observations of distant supernovae

Saul Perlmutter



2006 Physics

for the discovery of the blackbody form and anisotropy of the cosmic microwave background radiation

George Smoot

2007 Peace

for their efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change

Warren Washington

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Nobel Prize in Physics 2015

Scientific Achievement

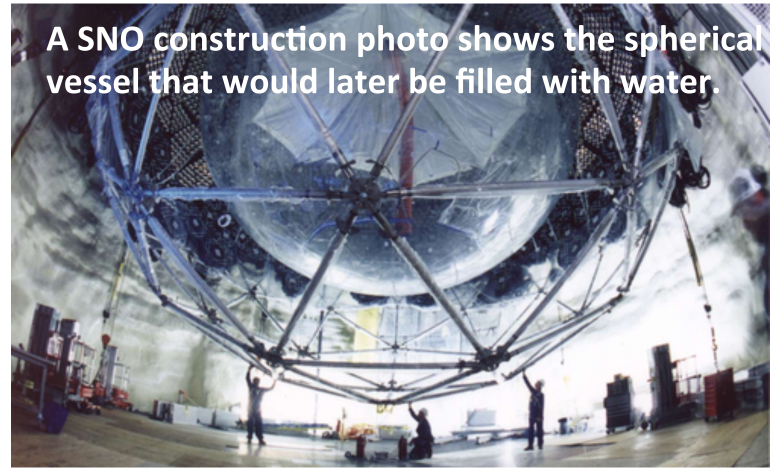
The discovery that neutrinos have mass and oscillate between different types

Significance and Impact

The discrepancy between predicted and observed solar neutrinos was a mystery for decades. This discovery overturned the Standard Model interpretation of neutrinos as massless particles and resolved the “solar neutrino problem”

Research Details

The Sundbury Neutrino Observatory (SNO) detected all three types (flavors) of neutrinos and showed that when all three were considered, the total flux was in line with predictions. This, together with results from the Super Kamiokande experiment, was proof that neutrinos were oscillating between flavors and therefore had mass



A SNO construction photo shows the spherical vessel that would later be filled with water.

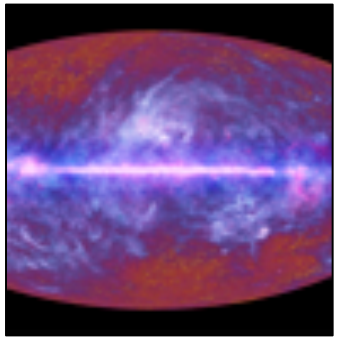
NERSC helped the SNO team use PDSF for critical analysis contributing to their seminal PRL paper. HPSS serves as a repository for the entire 26 TB data set.

Q. R. Ahmad et al. (SNO Collaboration). Phys. Rev. Lett. 87, 071301 (2001)

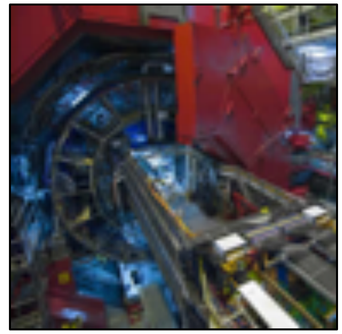
Nobel Recipients: Arthur B. McDonald, Queen’s University (SNO)
Takaaki Kajita, Tokyo University (Super Kamiokande)

NERSC has been supporting data intensive science for a long time

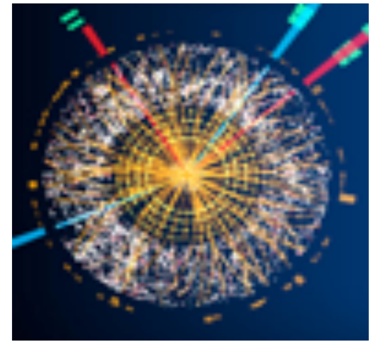
Palomar Transient Factory
Supernova



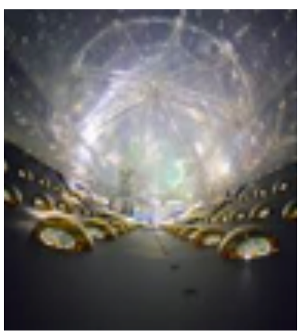
Planck Satellite
Cosmic Microwave Background Radiation



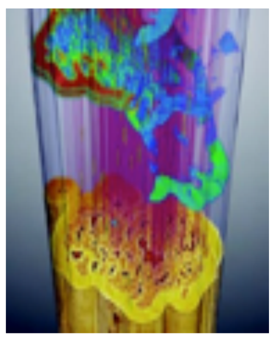
Alice
Large Hadron Collider



Atlas
Large Hadron Collider



Dayabay
Neutrinos

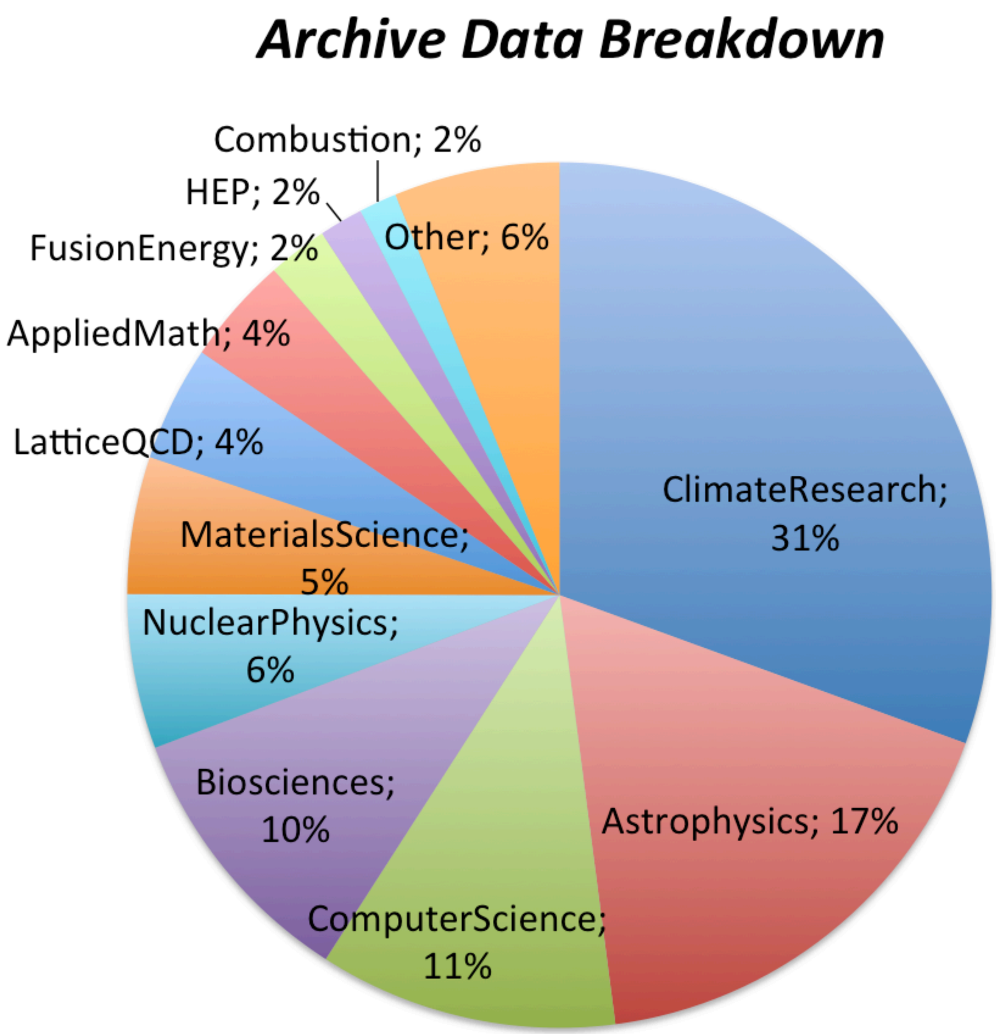


ALS
Light Source

LCLS
Light Source

Joint Genome Institute
Bioinformatics

NERSC archives An Enormous Amount of Data for the Scientific Community



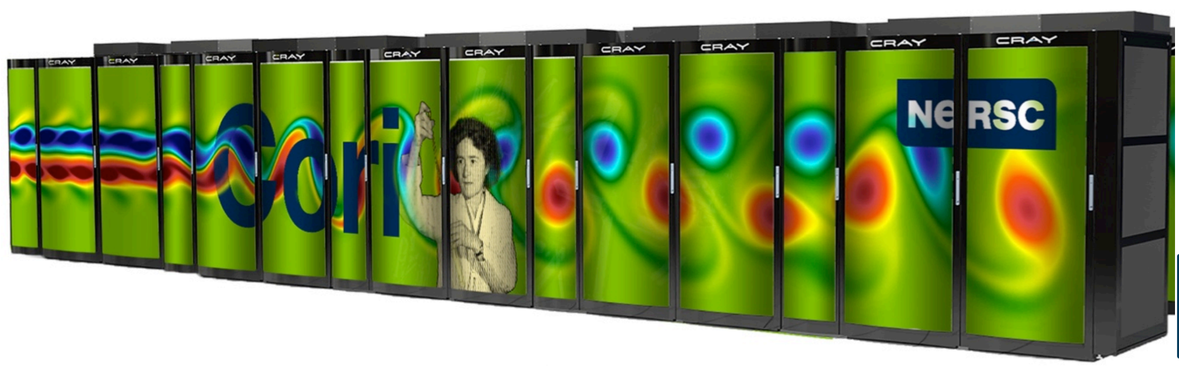
60+ PB of data are stored in NERSC's HPSS Archive

NERSC's Goal for Data Initiative

Increase the productivity, usability, and impact of DOE's experimental user facilities and other data-intensive science by providing comprehensive data systems and services to store, analyze, manage, and share data.

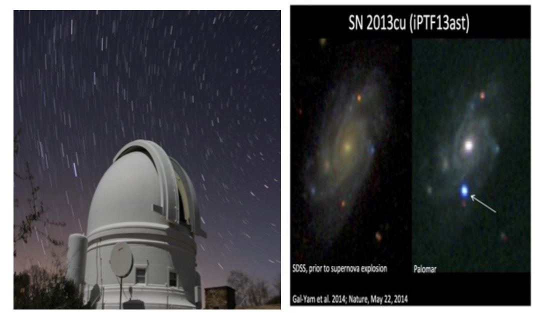
NERSC is making significant investments on Cori to support data intensive science

- **New queue policies: real time, and high throughput queues**
- **High bandwidth external connectivity to databases from compute nodes**
- **More (23) login nodes for managing advanced workflows**
- **Virtualization capabilities (Docker)**
- **NVRAM Flash Burst Buffer as I/O accelerator**
  - 1.5PB, 1.5 TB/sec
  - User can request I/O bandwidth and capacity at job launch time
  - Use cases include, out-of-core simulations, image processing, shared library applications, heavy read/write I/O applications

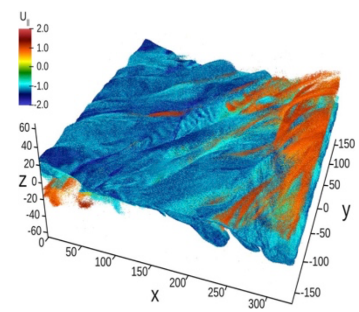


Burst Buffer Use Cases

- **Accelerate I/O**
  - Checkpoint/restart or other high bandwidth reads/writes
  - Apps with high IOP/s e.g. non-sequential table lookup
  - Out-of-core applications
  - Fast reads for image analysis
- **Advanced Workflows**
  - Coupling applications, using the Burst Buffer as interim storage
  - Streaming data from experimental facilities
- **Analysis and Visualization**
  - In-situ/ in-transit
  - Interactive visualization

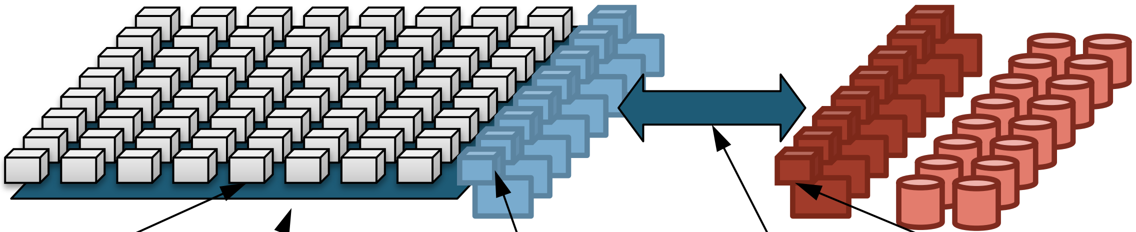


Palomar Transient Factory Pipeline:
Use Burst Buffer as cache for fast reads



VPIC – in situ visualization of a trillion particles

Burst Buffer Software Development Efforts



Create Software to enhance usability and to meet the needs of all NERSC users

- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Caching mode – data transparently captured by the BB nodes
  - Transparent to user -> no code modifications required
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale

Burst Buffer Early User Program call for proposals

- **Aug 10th: solicited proposals for BB Early Users program.**
  - Award of exclusive early use of BB on Cori P1, plus help of NERSC experts to optimise application for BB.
- **Selection criteria include:**
  - Scientific merit.
  - Computational challenges.
  - Cover range of BB data features.
  - Cover range of DoE Science Offices.
- **Great interest from the community, 29 proposals received. Good distribution across offices...**

Many great applications…

- We're very happy with the response to the program call.
- Decided to support more applications than we'd originally anticipated
- Other applications will not be supported by NERSC staff, but will have early access to Cori P1 and the BB.
- Breakdown by DoE Office:

	ASCR	BER	BES	Fusion	HEP	Nuclear	Total
NERSC Supported	1.5	2.5	2.5	1	4.5	1	13
Early Access	3	7	2.5	0	2.5	0	15

A variety of use cases are represented by the BB Early Users

Application	I/O bandwidth : reads	I/O bandwidth: writes (checkpointing)	High IOPs	Workflow coupling	In-situ / in-transit analysis and visualization	Staging inter-mediate files/ pre-loading data
Nyx/Boxlib		X		X	X	
Phoenix 3D		X		X		X
Chomo/Crunch + Visit		X		X	X	
Sigma/UniFam/Sipros	X	X	X			X
XGC1	X	X				X
PSANA				X	X	X
ALICE	X					
Tractor			X	X		X
VPIC/IO					X	X
YODA			X			X
ALS SPOT/TomoPy	X			X	X	X
kitware						

Realtime access to HPC systems

- We've heard from a number of users that the lack of 'realtime' access to the system is a barrier to scientific productivity
- With NERSC's new batch scheduler, SLURM, we have the capability to offer 'immediate' or 'real-time' access on Cori Phase 1, for projects and users with requirements for fast turn around
- We added a question to ERCAP about realtime needs to assess demand and size realtime resources.

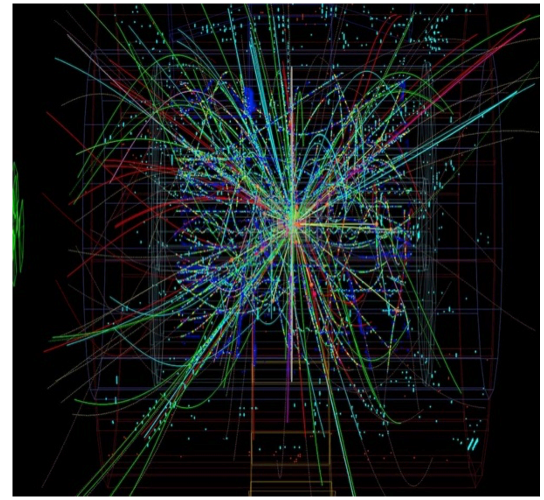
Immediate Queue – ERCAP Requests

- **19 responses (out of > 700) a small fraction of our workload**
- **Responses from 5 of 6 Offices, SBIR and EERE, demonstrating need is not confined to one scientific domain**
- **Expected responses:**
  - ALS, Palomar Transient Factory, CRD workflow research, MyGreenCar, OpenMSI, KBASE, Materials analysis, 2 PDSF projects
- **A few surprises**
  - 3 similar Fusion responses (MIT, GA, LLNL) noting DIII-D (tokomak fusion reactor run by General Atomics) can be adjusted by real-time codes
  - Industry response, Vertum partners, Predictive Power Grid performance, run simulations daily 12 hours apart.

Shifter brings user defined images to supercomputers

NERSC

- **Shifter, a container for HPC, allows users to bring a customized OS environment and software stack to an HPC system.**
- **Use cases**
  - High energy physics collaborations that require validated software stacks
  - Cosmology and bioinformatics applications with many 3rd party dependencies
  - Light source applications that with complicated software stacks that need to run at multiple sites



SHIFTER

Upgrading Cori's External Connectivity

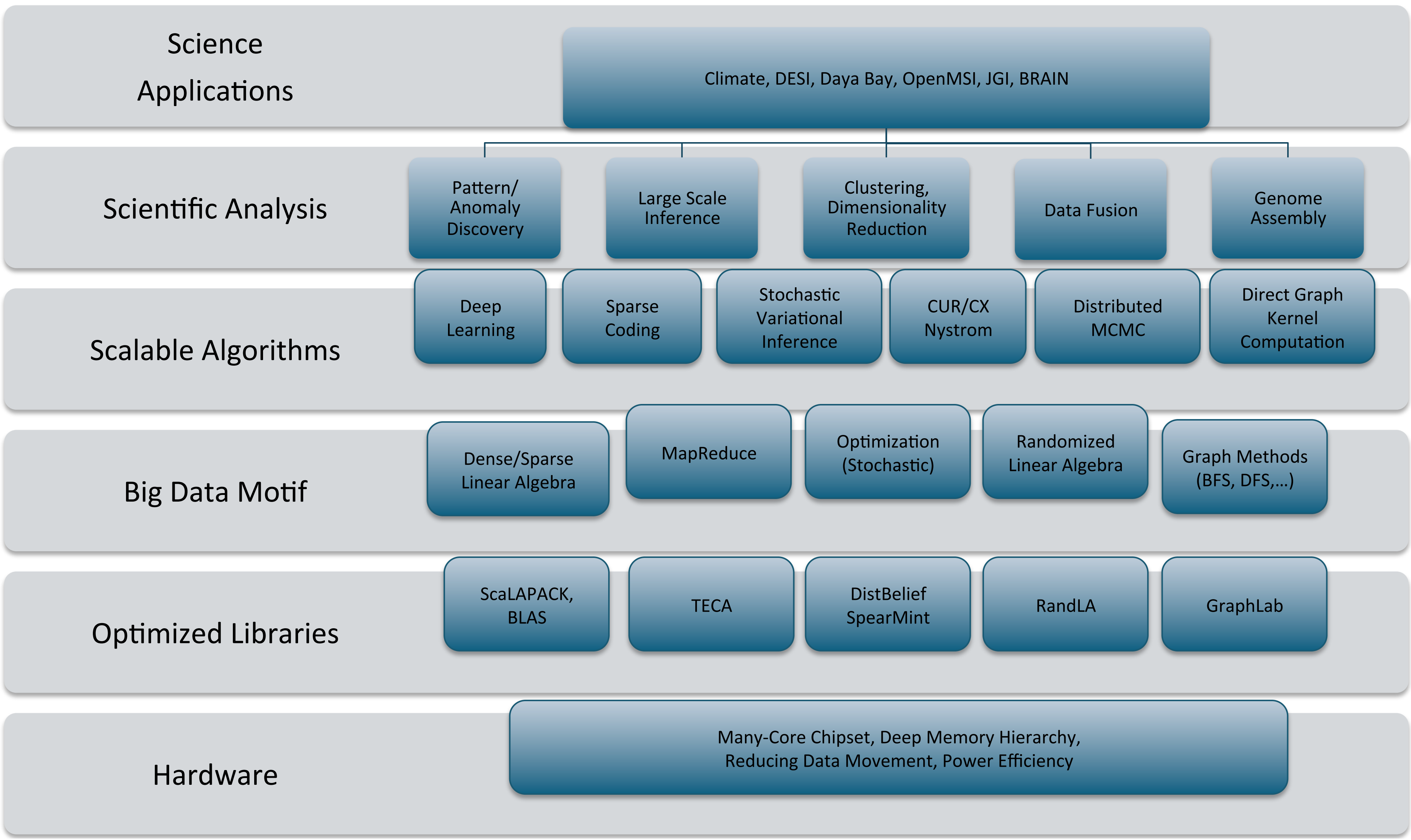
Enable 100Gb+ Instrument to Cori

- Streaming data to the supercomputer allows for analytics on data in motion
- Cori network upgrade provides SDN (software defined networking) interface to ESnet. 8 x 40Gb/s bandwidth.
- Integration of data transfer and compute enables workflow automation

Cori Network Upgrade Use Case:

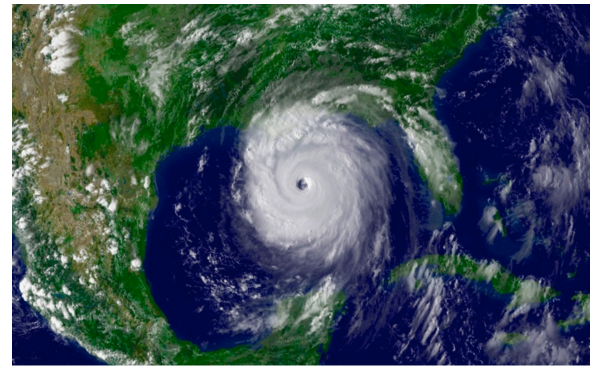
- X-ray data sets stream from detector directly to Cori compute nodes, removing need to stage data for analysis.
- Software Defined Networking allows planning bandwidth around experiment run-time schedules
- 150TB bursts now, LCLS-II has 100x data rates

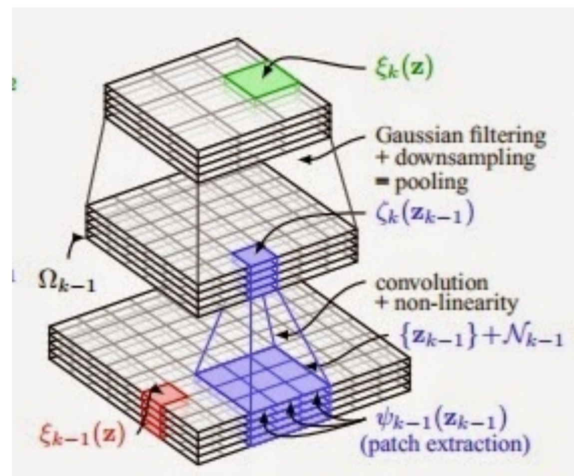
Data Analytics Research Strategy



Deep Learning for Pattern Detection in Climate Simulations

- **Scientific Achievement**
  - Extreme events/patterns can now be automatically extracted from massive climate simulation datasets
  - Tropical Cyclones, Atmospheric Rivers and Weather fronts can be analyzed in an automated manner
- **Significance and Impact**
  - One of the first successful demonstrations of Deep Learning for solving a scientific pattern recognition problem
- **Research Details**
  - 87%-99% accuracy obtained in labeling Tropical Cyclones, Atmospheric Rivers and Weather Fronts patterns
  - Convolutional auto-encoders used with hyper-parameter tuning on Edison and Cori
  - Paper submitted to KDD 2016

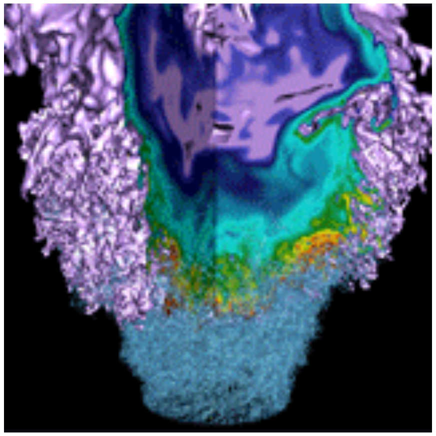


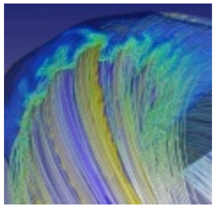


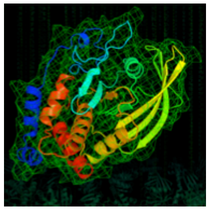
NERSC is actively exploring Deep Learning

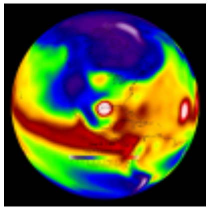
- **Collaboration with Intel and UCB on Deep Learning**
  - Recent Intel ImageNet submission used Edison and Cori platforms; *Caffe+PCL_DNN* deployed
  - Efforts to port *FireCaffe* underway w/ Kurt+Forrest
- **Collaboration with startups**
  - Nervana Systems (*Neon*)
  - WhetLabs (*Spearmint*)
- **MANTISSA is funding leading statistics and machine learning researchers**
  - UC Berkeley EECS, Statistics, Redwood Center
  - Harvard, MIT

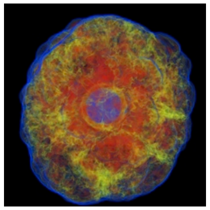
Superfacility Concept

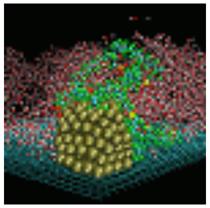




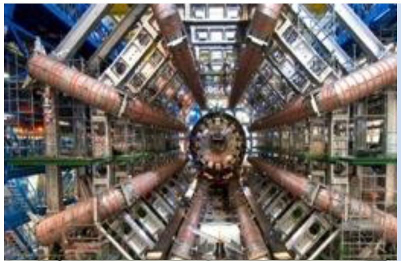
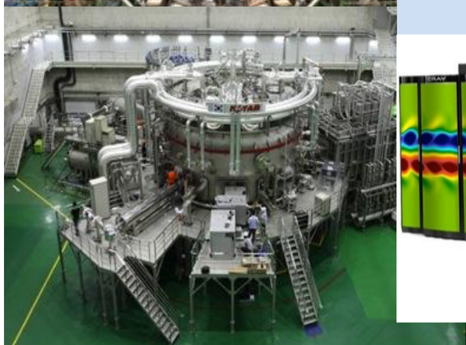
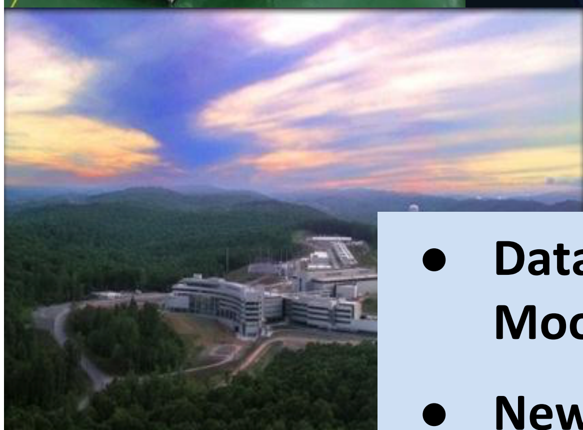


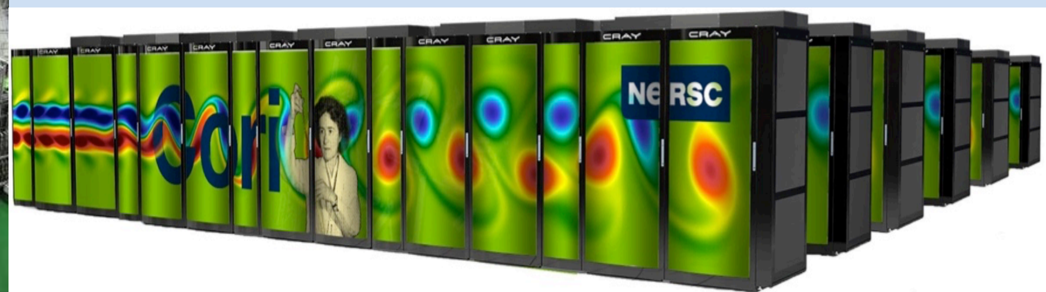




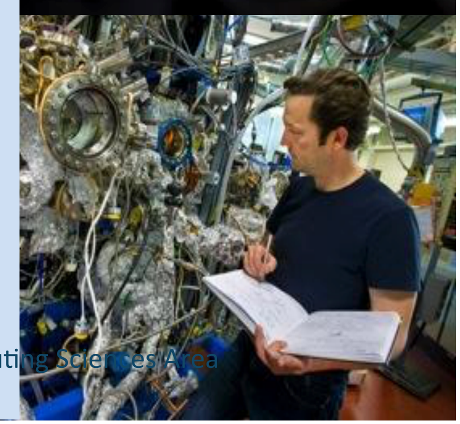


Experimental and observational science is at crossroads





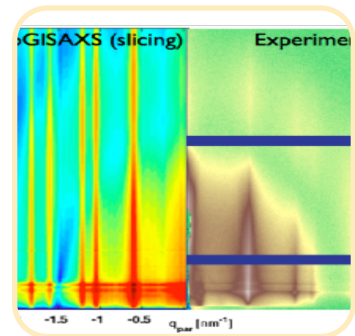



- Data volumes are increasing faster than Moore's Law
- New algorithms and methods for analyzing data
- Infeasible to put a supercomputing center at every experimental facility

Superfacility Prototype and Use Case : Process of science transformed

‘Eliminate boundaries between the Scientist and the world’s best algorithms running on the best architecture for that code’

Real-time analysis of ‘slot-die’ technique for printing organic photovoltaics, using ALS + NERSC (SPOT Suite for reduction, remeshing, analysis) + OLCF (HipGISAXS running on Titan w/ 8000 GPUs).



http://www.es.net/news-and-publications/esnet-news/2015/esnet-paves-way-for-hpc-superfacility-real-time-beamline-experiments/ Results presented at March 2015 meeting of American Physical Society by Alex Hexemer. Additional DOE contributions: **GLOBUS** (ANL), **CAMERA** (Berkeley Lab)

Thank you!

Popular features of a data intensive system can be supported on Cori

Data Intensive Workload Need	Cori Solution
Fast I/O	NVRAM 'burst buffer', configurable bandwidth and capacity on a per job basis
Large memory nodes	128 GB/node on Haswell; Option to purchase fat (1TB) login node
Flexible queues supporting real-time access and massive numbers of jobs	New real-time, serial and high throughput queues on Cori
Complex workflows	More (23) external login nodes;
High bandwidth communication and streaming to compute nodes from external sources	We have worked with Cray to increase bandwidth to Cray system and we have funding for hardware to enable streaming directly to compute nodes.
Easy to customize environment	New Shifter solutions allows customized images