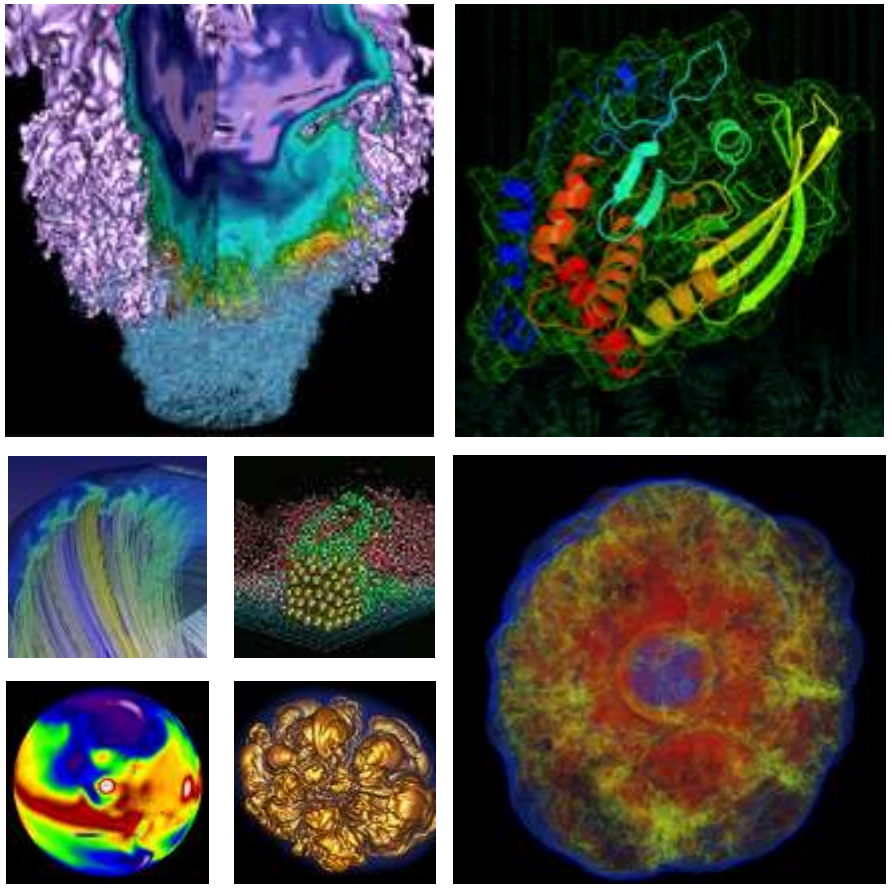
Cori: The NERSC-8 System



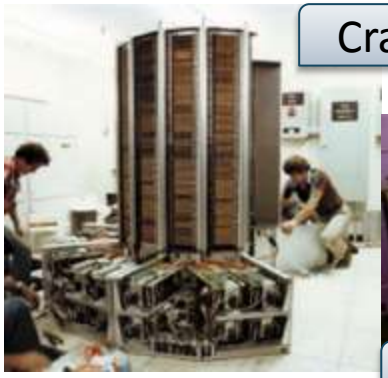
Jay Srinivasan
Group Lead, Computational Systems
NERSC-8 Project Deputy

September 17, 2014, HPC User Forum

Topics

- **NERSC at 40 years**
- **User Needs and insatiable demand for cycles**
- **Cori's unique features**
  - Procurement model
  - System and Chip architecture
  - Timeline
  - Using Cori
  - Free cooling

NERSC's 40th Anniversary!



Cray 1 - 1978

Cray 2 – 1985

Cray T3E Mcurie - 1996

IBM Power3 Seaborg - 2001

1974	Founded at Livermore to support fusion research with a CDC system
1978	Cray 1 installed
1983	Expanded to support today's DOE Office of Science
1986	ESnet established at NERSC
1994 - 2000	Transitioned users from vector processing to MPP
1996	Moved to Berkeley Lab
1996	PDSF data intensive computing system for nuclear and high energy physics
1999	HPSS becomes mass storage platform
2006	Facility wide filesystem
2010	Collaboration with JGI
2013	Petascale Cray HPCS system

U.S. DEPARTMENT OF ENERGY | Office of Science

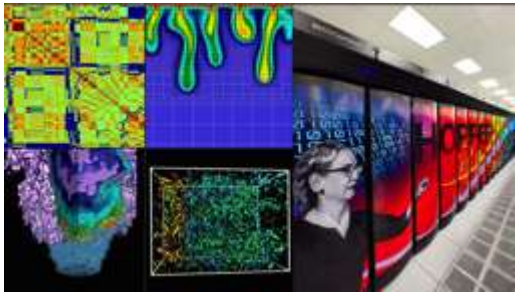
BERKELEY LAB

Facility for DOE Office of Science Research

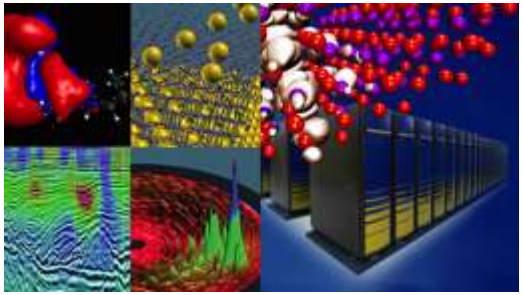
U.S. DEPARTMENT OF ENERGY | Office of Science

Largest funder of physical science research in U.S.

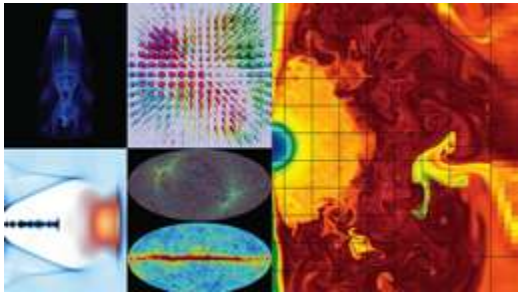
Bio Energy, Environment



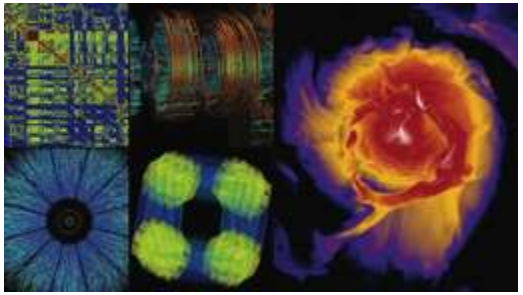
Computing



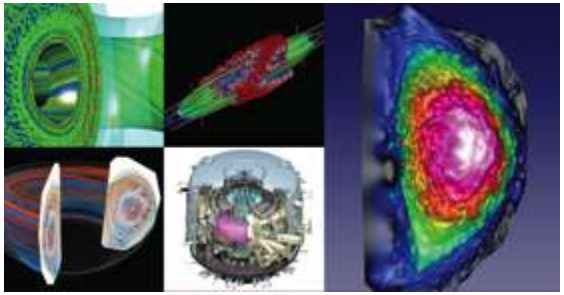
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



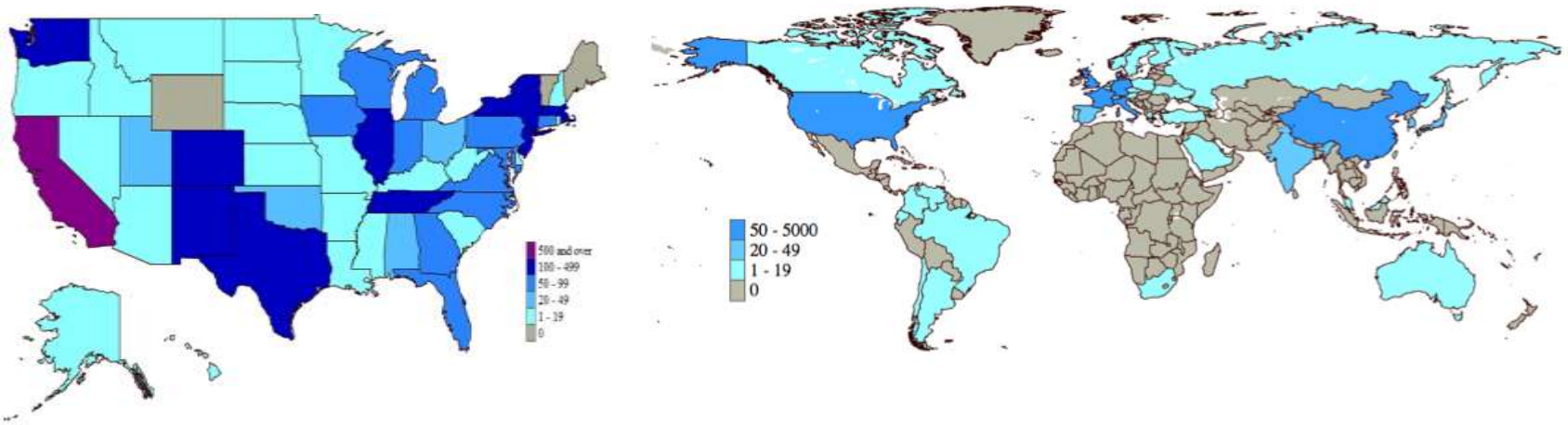
Nuclear Physics



Fusion Energy, Plasma Physics

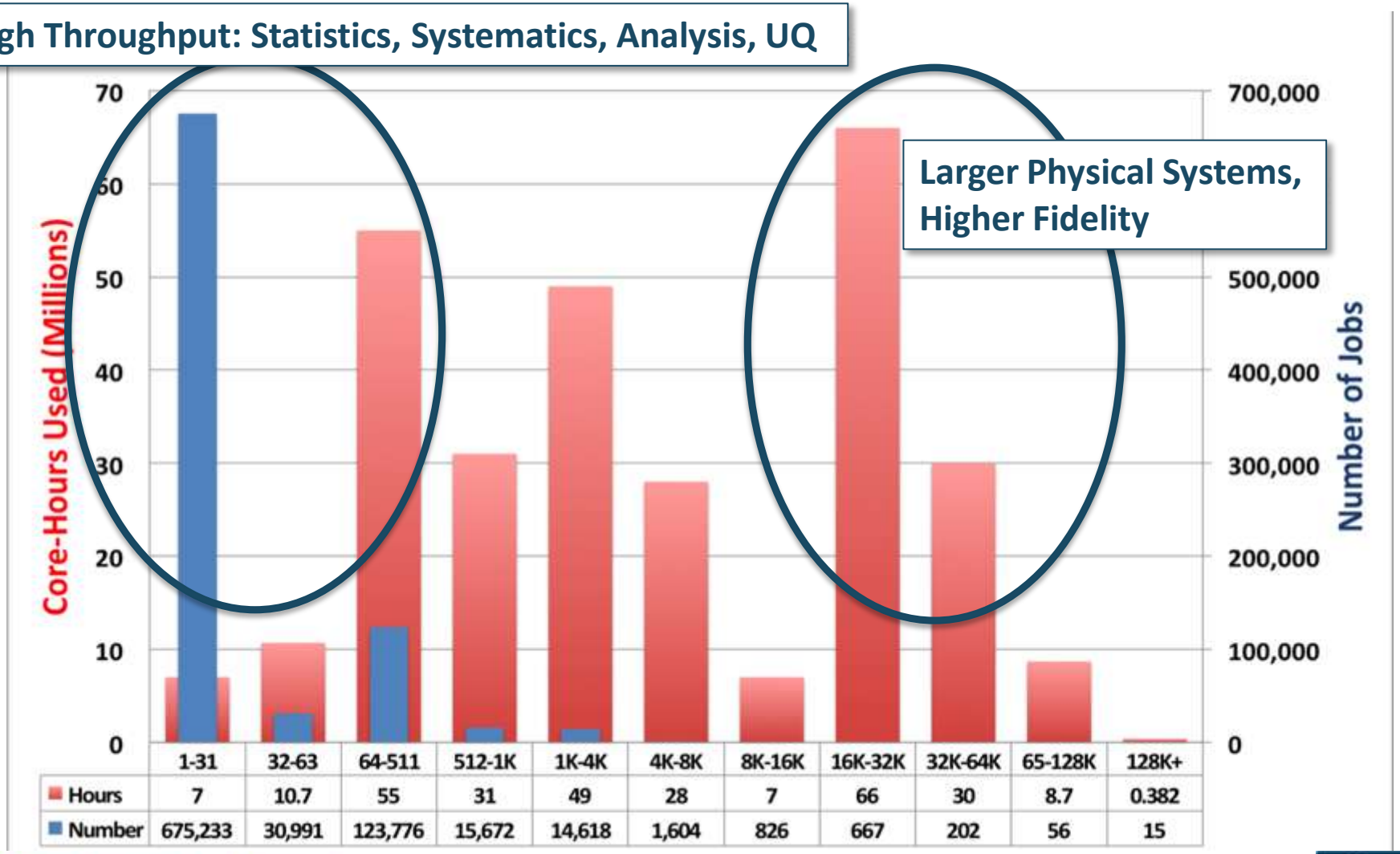
We support a broad user base

- **5000 users, and we typically add 350 per year**
- **Geographically distributed: 47 states as well as multinational projects**



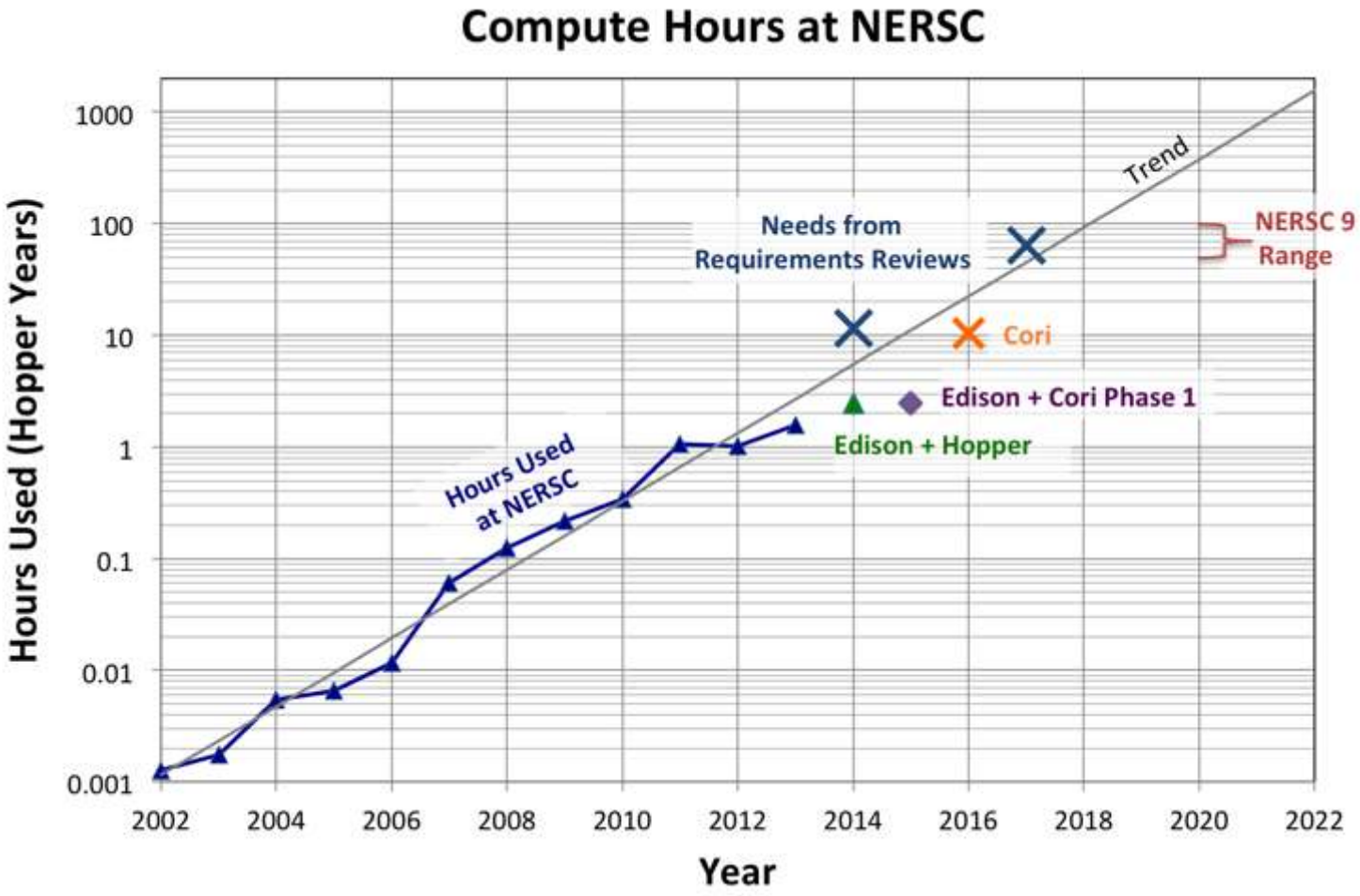
NERSC Supports Science Needs at Many Difference Scales and Sizes

High Throughput: Statistics, Systematics, Analysis, UQ



	1-31	32-63	64-511	512-1K	1K-4K	4K-8K	8K-16K	16K-32K	32K-64K	65-128K	128K+
Hours	7	10.7	55	31	49	28	7	66	30	8.7	0.382
Number	675,233	30,991	123,776	15,672	14,618	1,604	826	667	202	56	15

Increasing User Needs



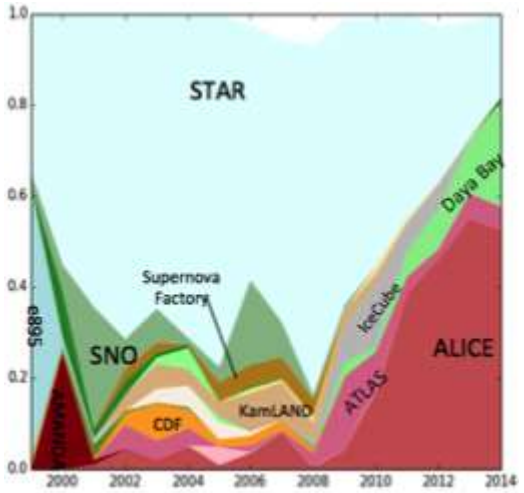
We currently deploy separate Compute Intensive and Data Intensive Systems

Compute Intensive

Data Intensive

Carver

Genepool



PDSF

We surveyed our users about why they use our data intensive systems:

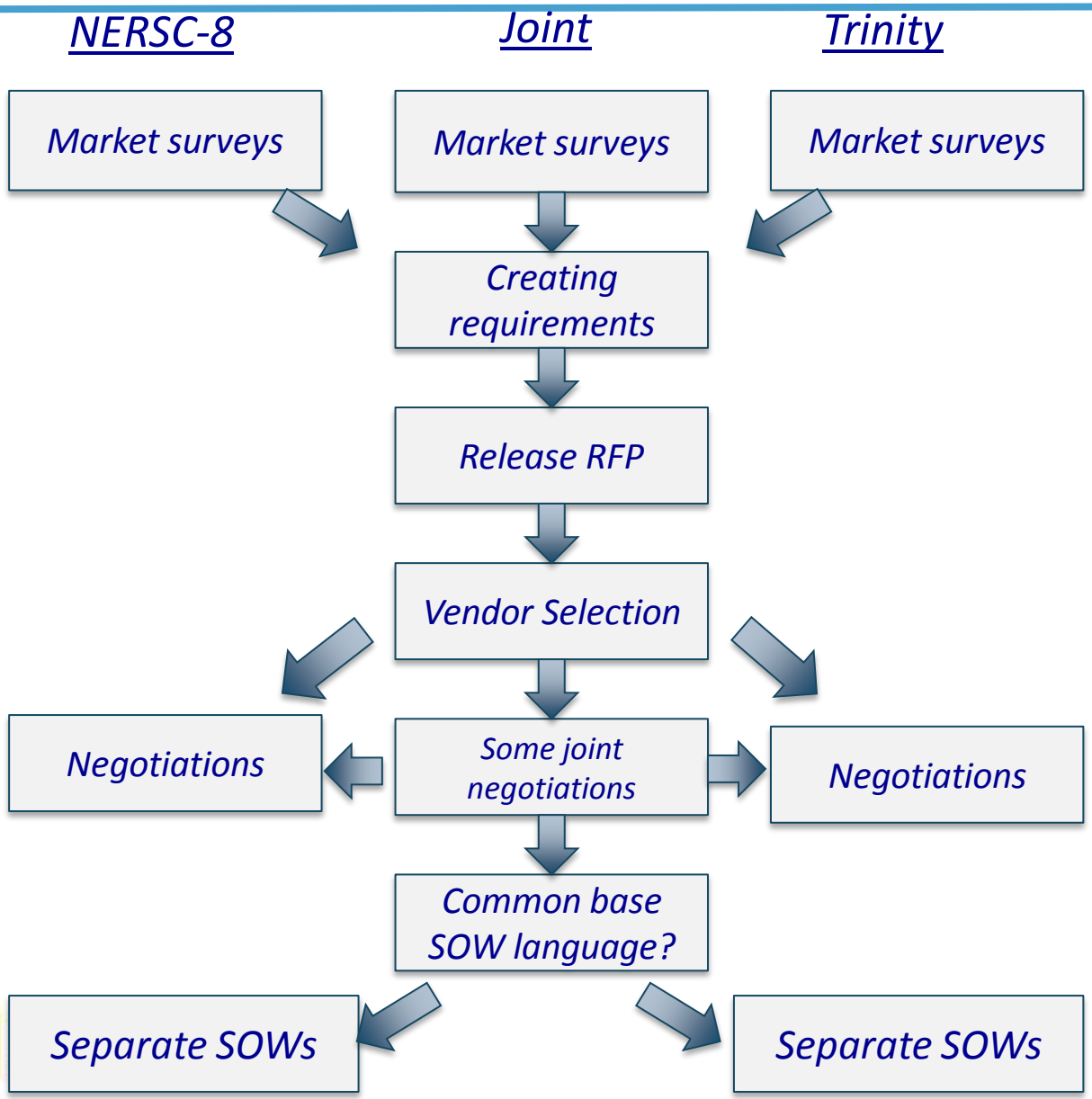
- **Complex workflows (including High Throughput Computing - HTC)**
- **Policy flexibility**
- **Local disk**
- **Very large memory**
- **Massive serial jobs (~100K)**
- **Communicate with databases / host databases**
- **Stream data from Observational/Experimental Facilities**
- **Easy to customize environment and the environment is familiar**

You'll notice most of this list is not hardware…

odds, Data & Simulation are intertwined

- **The Department of Energy has Big Data needs**
  - E.g., Office of Science's Experimental Facilities
- **Big data requires simulation**
  - Missing data; models of what the data means
- **Big data requires big computing**
  - And memory and storage
- **Heroic computing generates big data**
  - Petabytes in many science areas
- **Large volume computing generates big data**
  - Screening materials, proteins, UQ, etc.
- **Data improves science impact**
  - Community data sets enable science

SC and NNSA cooperation on a procurement: NERSC-8 and Trinity



NERSC-8 (Cori) Mission Need

The Department of Energy Office of Science requires an HPC system to support the rapidly increasing computational demands of the entire spectrum of DOE SC computational research.

- Provide a significant increase in computational capabilities, at least 10 times the sustained performance of the Hopper system on a set of representative DOE benchmarks
- Delivery in the 2015/2016 time frame
- Provide high bandwidth access to existing data stored by continuing research projects.
- Platform needs to begin to transition users to more energy-efficient many-core architectures.

The Cori System: Pre-exascale computing and big data capabilities

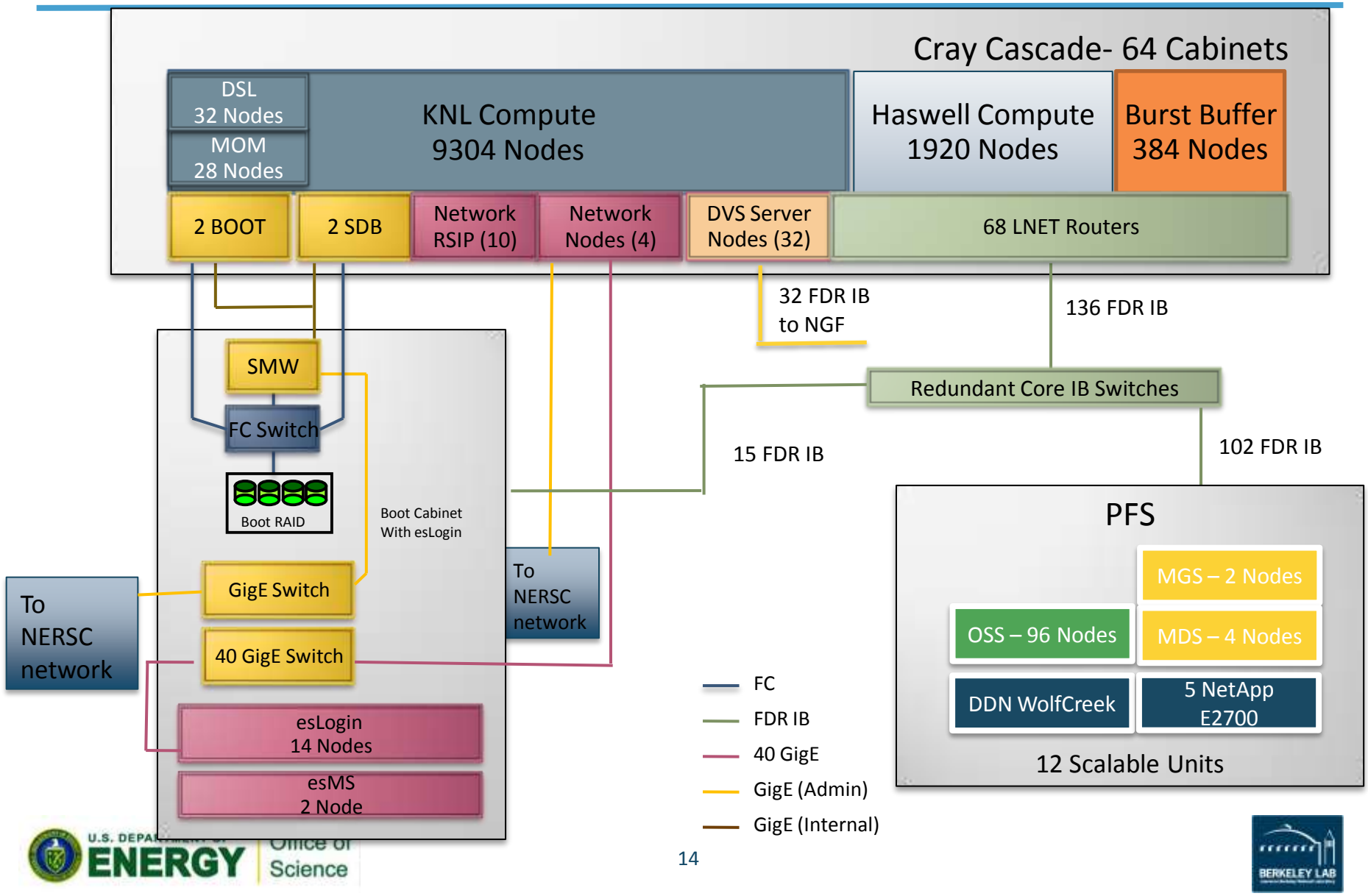
- **Cray XC system with over 9300 Intel Knights Landing compute nodes – 2016**

- **NERSC Exascale Science Applications Program will prepare users for Cori**
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams
- **Data Partition with ~2000 Haswell nodes - 2015**
  - NVRAM Burst Buffer to accelerate data intensive applications
  - 28 PB of disk, 432 GB/sec I/O bandwidth
- **Cray Aries interconnect**

Image source: Wikipedia

System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.

The Cori System



Knights Landing

- **Next generation Xeon-Phi >3TF peak (3X single-thread performance over current Xeon-Phi)**
- ***Self-hosted, not a co-processor or accelerator***
- **>60 cores per processor with four hardware threads each**
- **512b vector units (32 flops/clock – AVX 512)**
- **16GB High B/W on-package memory (5X B/W of DDR4 DRAM)**
- **64-128 GB of DRAM/node**

Cache Model

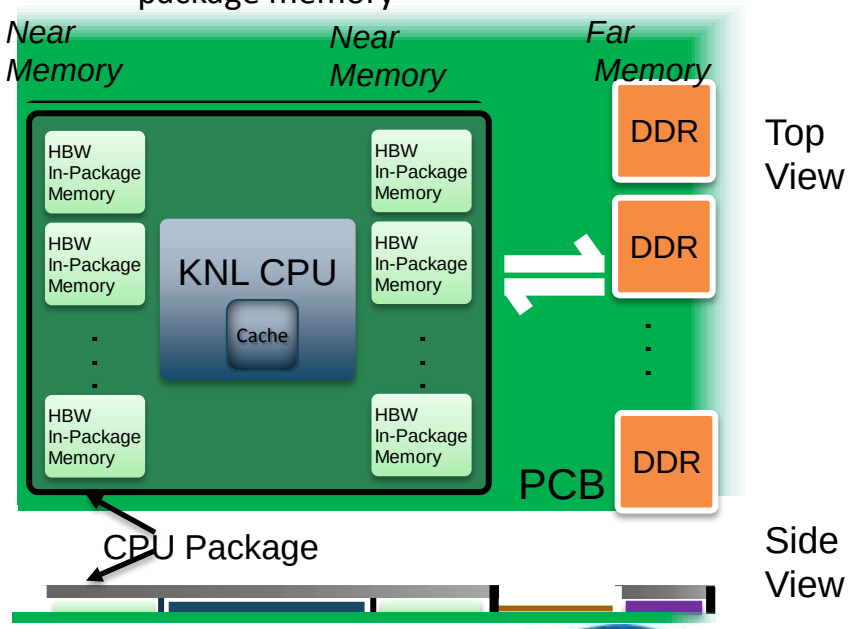
Let the hardware automatically manage the integrated on-package memory as an "L3" cache between KNL CPU and external DDR

Flat Model

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

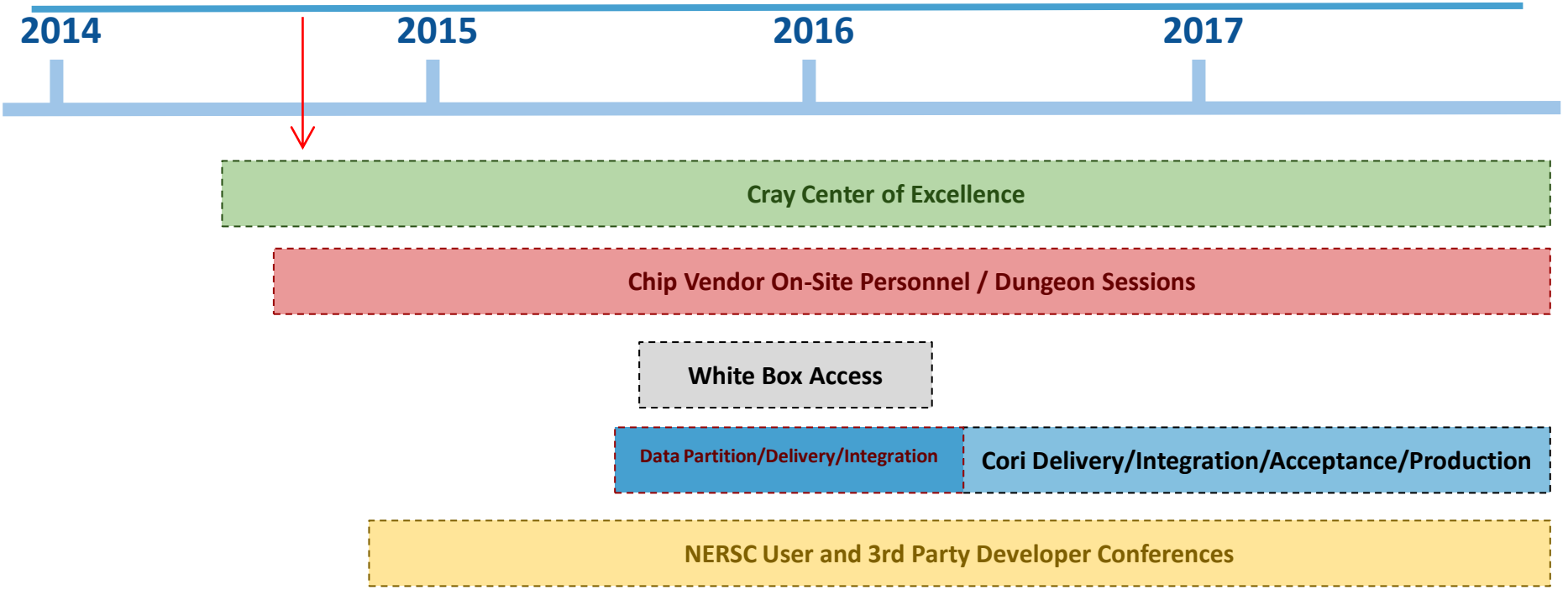
Hybrid Model

Harness the benefits of both cache and flat models by segmenting the integrated on-package memory



Slide from Intel

Cori Timeline



Cori Contract Includes Significant Vendor Support

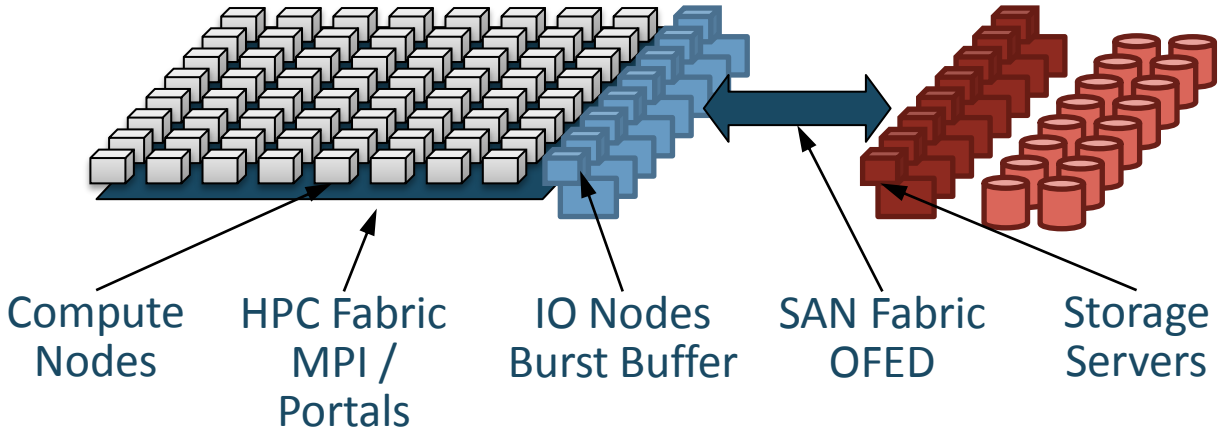
- **Cray**
  - 5 FTE years of application and optimization support
- **Intel**
  - Remote access to an early KNL system
  - KNL white boxes @ NERSC before arrival of Cori
  - 4 Training sessions – 2 per year
  - Quarterly Dungeon sessions – 16 in total
  - Intel associate on-site 1 week/month for 4 years
- **Cray/NERSC Center of Excellence (COE):**
  - Study KNL performance in detail and help migrate codes

A single system offers significant benefits to science

- **NERSC has two major strategic thrusts**
  - Useable Exascale
  - Enabling Data Intensive Computing
- **Cori plus the Cori Phase 1 Data Partition will enable NERSC to make an impact on both**
- **Success depends on enabling the Cori Data Partition to meet the needs of Data Intensive Users**

Goals are to enable the analysis of large experimental data sets and in-situ analysis coupled to Petascale simulations

Burst Buffer Software NRE Efforts



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

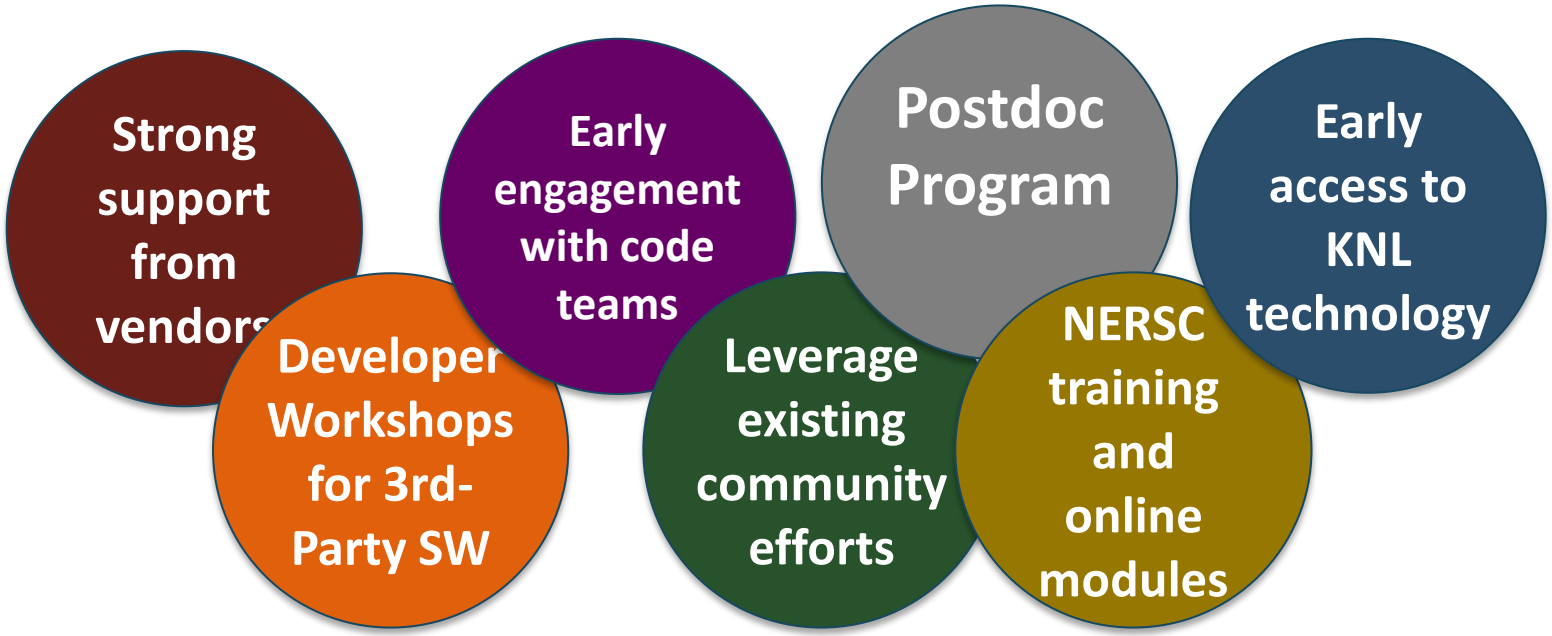
- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale
- Caching mode – data transparently captured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

Programming Model Considerations

- **Knight's Landing is a self-hosted part**
  - Users can focus on adding parallelism to their applications without concerning themselves with PCI-bus transfers
- **MPI + OpenMP preferred programming model**
- **Why OpenMP?**
  - Expect between 1-2GB memory *per core*
  - With 2 threads/core memory/thread drops to less than 1 GB
  - Will need to use HW threads to get optimal performance on KNL
- **MPI-only will work – performance may not be optimal**
- **On package MCDRAM**
  - How to optimally use ? Explicitly or Implicitly?

NERSC Exascale Science Applications Program (NESAP)

- **NESAP components:**

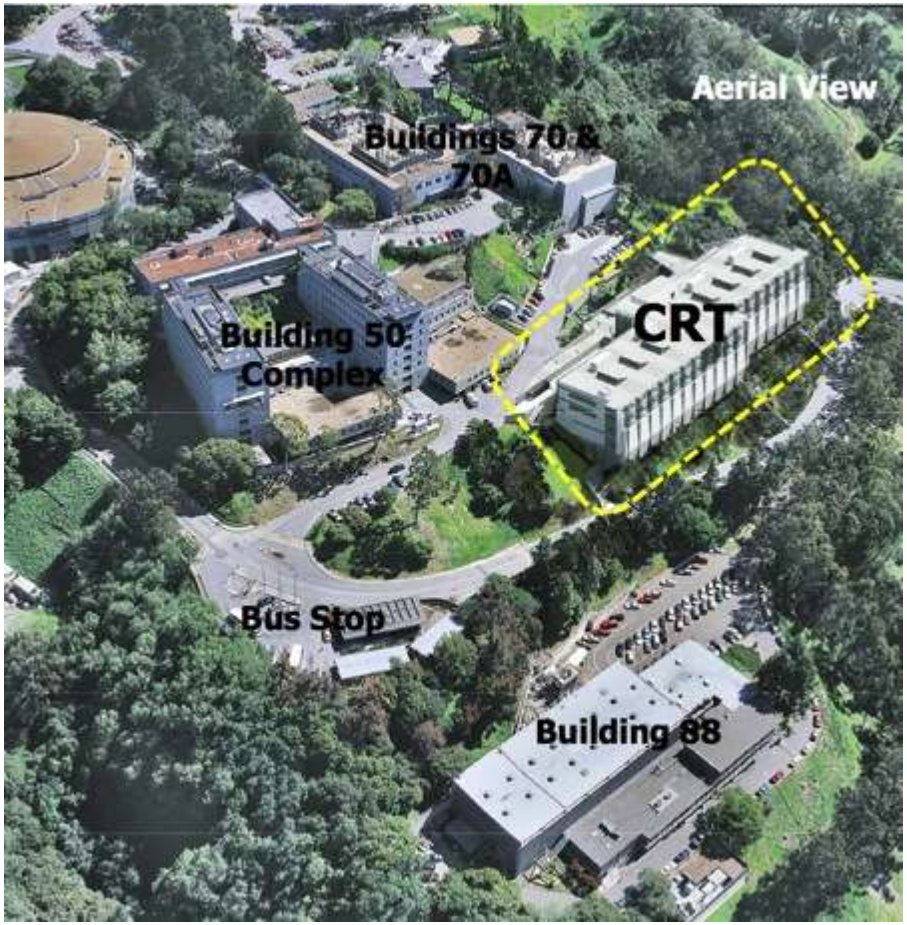


Community involvement is important

- **Many applications run across multiple systems**
- **The transition to manycore will be challenging – We need to share lessons learned and best practices**
- **We want to encourage portability as much as possible**
- **Xeon Phi user group is coming soon**
- **Intel is sponsoring Intel Parallel Computing Centers at many institutions**

NERSC will move to CRT in Spring 2015

- **Mixed office and data center building**
  - 300 offices on two floors
  - 20->30Ksf HPC floor
  - Seismic isolation raised floor
- **Extremely energy efficient**
  - LEED gold design
  - "Free" cooling for air and water
  - No chillers
  - Heat recovery
- **Expandable power**
  - 12.5MW at move-in
  - 42MW capacity to building
- **Cori will be installed at CRT**



Summary

- **NERSC's goals are:**
  - To provide *usable* Exascale computing **and**
  - Enable Data Intensive Computing
- **Cori is the first step**
  - Exciting technology!
    - Processor
    - Burst buffer
    - Data Partition
  - Lots of vendor and community support
  - Programming environment that enables users to transition to energy efficient architectures through robust and portable coding

Thank you!

Acknowledgements: NERSC-8 Team (Allie Andrews, Katie Antypas, Brian Austin, Nick Cardo, Matt Cordery, Chris Daley, Jack Deslippe, Scott French, Richard Gerber, Helen He, Larry Pezzaglia, , Harvey Wasserman, Nick Wright, Woo-Sun Yang,Zhengji Zhao), CSG Staff, NERSC Management