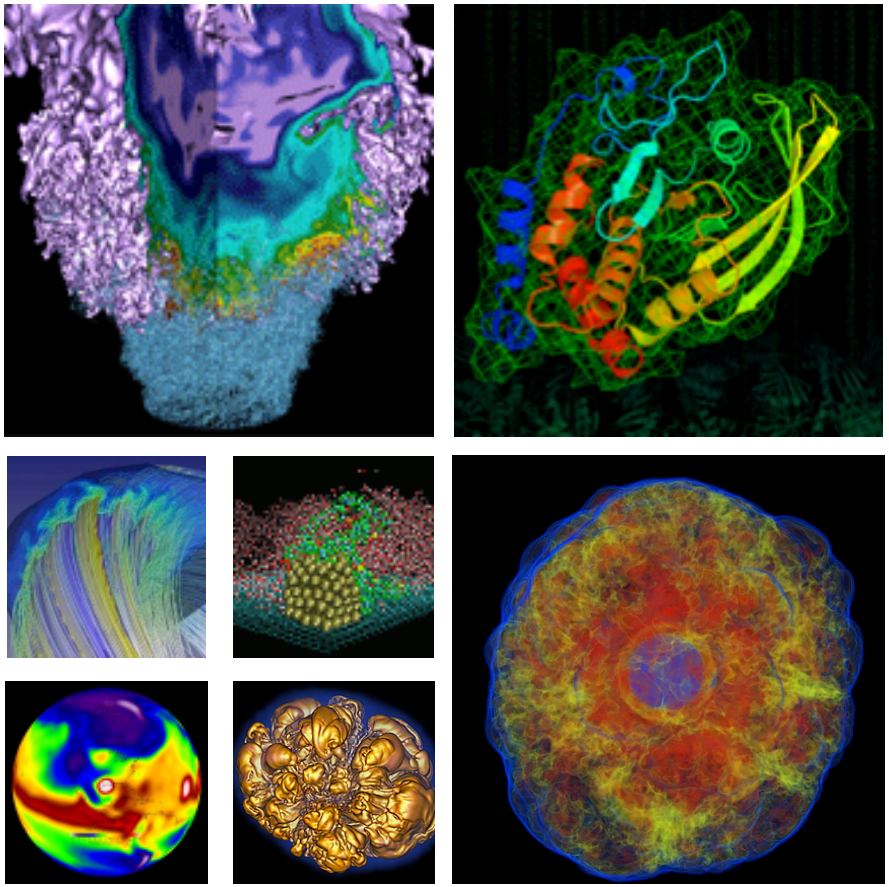
NERSC-8 Market Survey



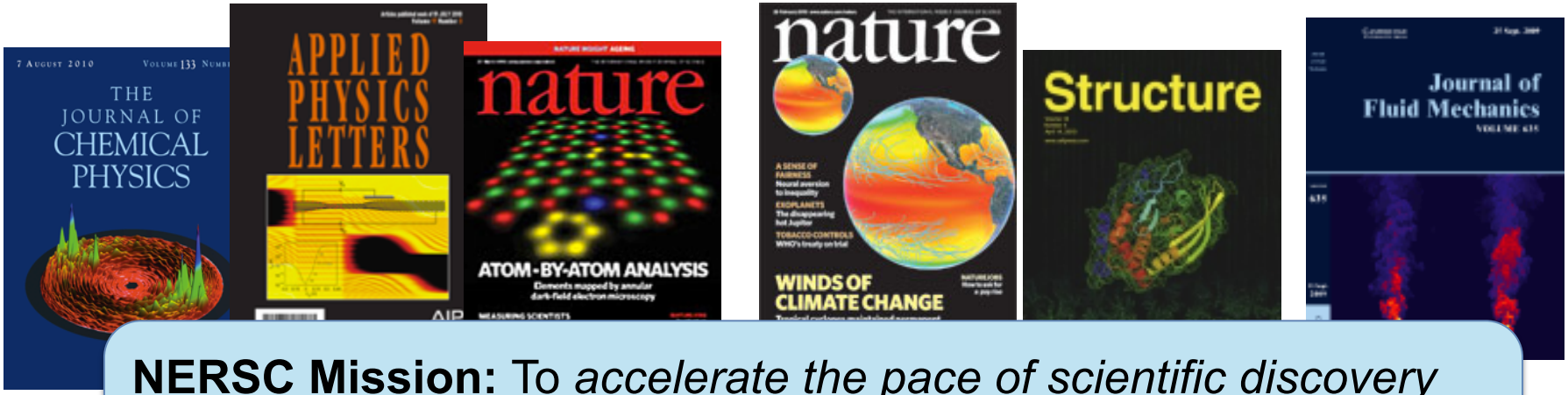
Katie Antypas
NERSC-8 Project Lead

November 15, 2012

We are starting our next procurement, NERSC-8, with a round of market surveys

- **Seek vendor input to optimize timing, requirements and business practices**
- **Opportunity for vendors to provide input prior to formal procurement process**

NERSC's mission is to enable science



NERSC Mission: To *accelerate the pace of scientific discovery* by providing high-performance computing, data systems and services to the DOE Office of Science community.

NERSC has over 4500 users in 650 projects that produce about **1500 publications per year!**

NERSC's Long Term Strategy

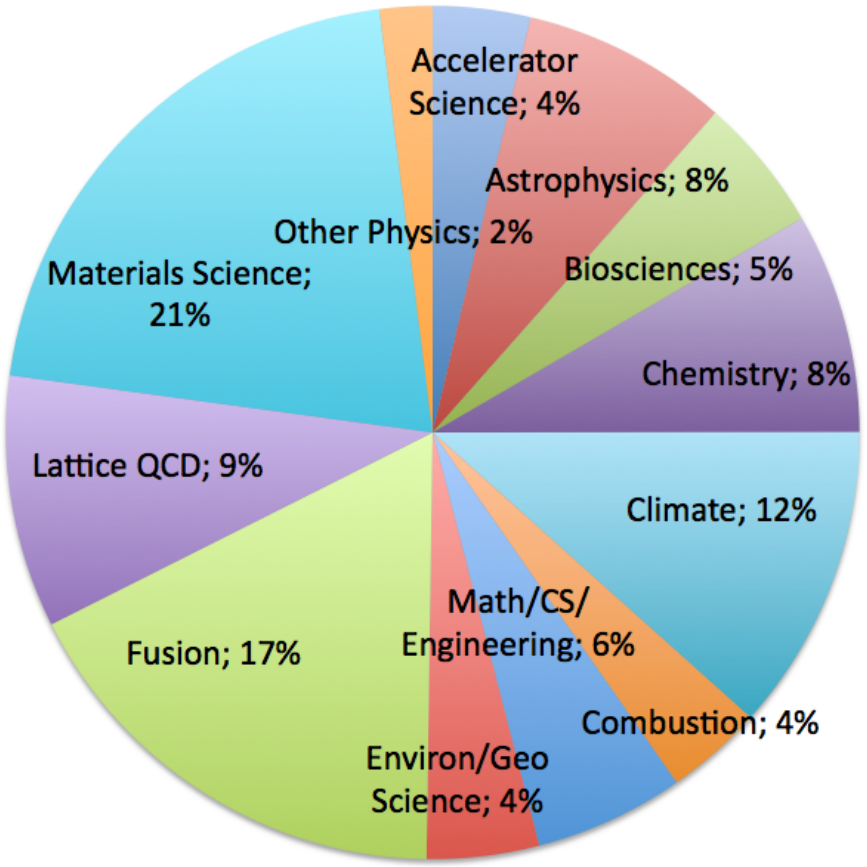
- **New system every ~3 years, run for 5-6 years**
  - Maximizes stability rather than peak / machine
  - Single system for DOE/SC HPC workload (minimize TCO)
  - Testbeds, e.g., GPU cluster, for technology exploration / users
- **NERSC-5 decommissioned May 1, 2012**
  - Franklin 25 TF on applications, 356 TF peak
- **NERSC-6 installed in phases, 2009-2011**
  - Hopper is 144 TF on applications; 1.3 PF peak
- **NERSC-7**
  - Cray Cascade system; 236 TF on applications; > 2PF peak
  - 2 phased system, installed 2013
- **NERSC-8 planned for 2015/2016**
- **Plan to reach Exascale in next 10 years**

NERSC-8 Design Targets and Limitations

- 10-30x application performance of Hopper XE6 system as measured by SSP metric (Hopper 1.3 PFs peak)
- Transition DOE users to more energy efficient manycore architectures
- Support Office of Science workload
- Delivery 2015/2016
- Maximum power – 6MW
- More specifications in draft RFP – release expected Dec. 2013

NERSC supports a broad range of science applications

2011 Allocation Breakdown

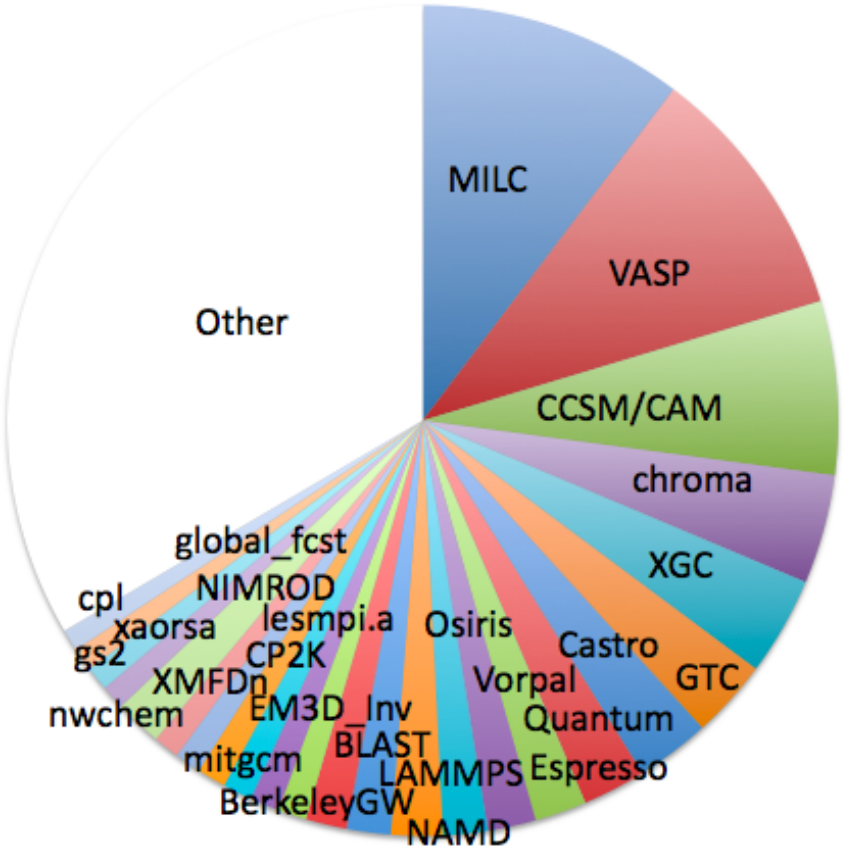


- **Over 4500 users**
- **Over 650 projects**
- **Hundreds of users logged in each day**
- **~1500 publications per year**

Challenge is to procure a system that satisfies our diverse workload while preparing users for more energy-efficient future architectures

Although NERSC has a broad user base, the workload is highly concentrated and unevenly distributed

Breakdown of NERSC Workload



- **Two-thirds of NERSC workload is concentrated in ~25 application codes**
- **While porting applications to any new architecture will be challenging, concentration of applications, makes task less daunting.**

U.S. DEPARTMENT OF ENERGY | Office of Science

Code analysis from Hopper from Jan – Aug 2012

BERKELEY LAB

Approximate NERSC-8 timeline

- Mission Need approval Q3 2012
- RFP release ~Q2 2013
- Vendor selection and negotiations ~Q3/Q4 2013
- Delivery late 2015/early 2016

The ACES Trinity team and the NERSC-8 team are collaborating

- **Teams worked together on Hopper/Cielo and found interactions useful**
- **Strengthen alliance between SC/NNSA on road to exascale**
- **Share technical expertise between Labs**

Plans are for a joint Trinity/NERSC-8 RFP calling for two distinct systems of similar technology with the intention to award both systems to the same vendor.

Current NERSC Systems

Large-Scale Computing Systems

Hopper (NERSC-6): Cray XE6

- 6,384 compute nodes, 153,216 cores
- 144 Tflop/s on applications; 1.3 Pflop/s peak

Edison (NERSC-7): Cray XC30 (Cascade)

- Arrives in 2 phases – Dec 2012 with full system Q2 2013
- 236 Tflop/s on applications; > 2 Pflop/s peak

Midrange

140 Tflops total

Carver

- IBM iDataplex cluster
- 9884 cores; 106TF

PDSF (HEP/NP)

- ~1K core cluster

GenePool (JGI)

- ~5K core cluster
- 2.1 PB Isilon File System

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

- 8.5 PB capacity
- 15GB/s of bandwidth

HPSS Archival Storage

- 240 PB capacity
- 5 Tape libraries
- 200 TB disk cache

Analytics & Testbeds

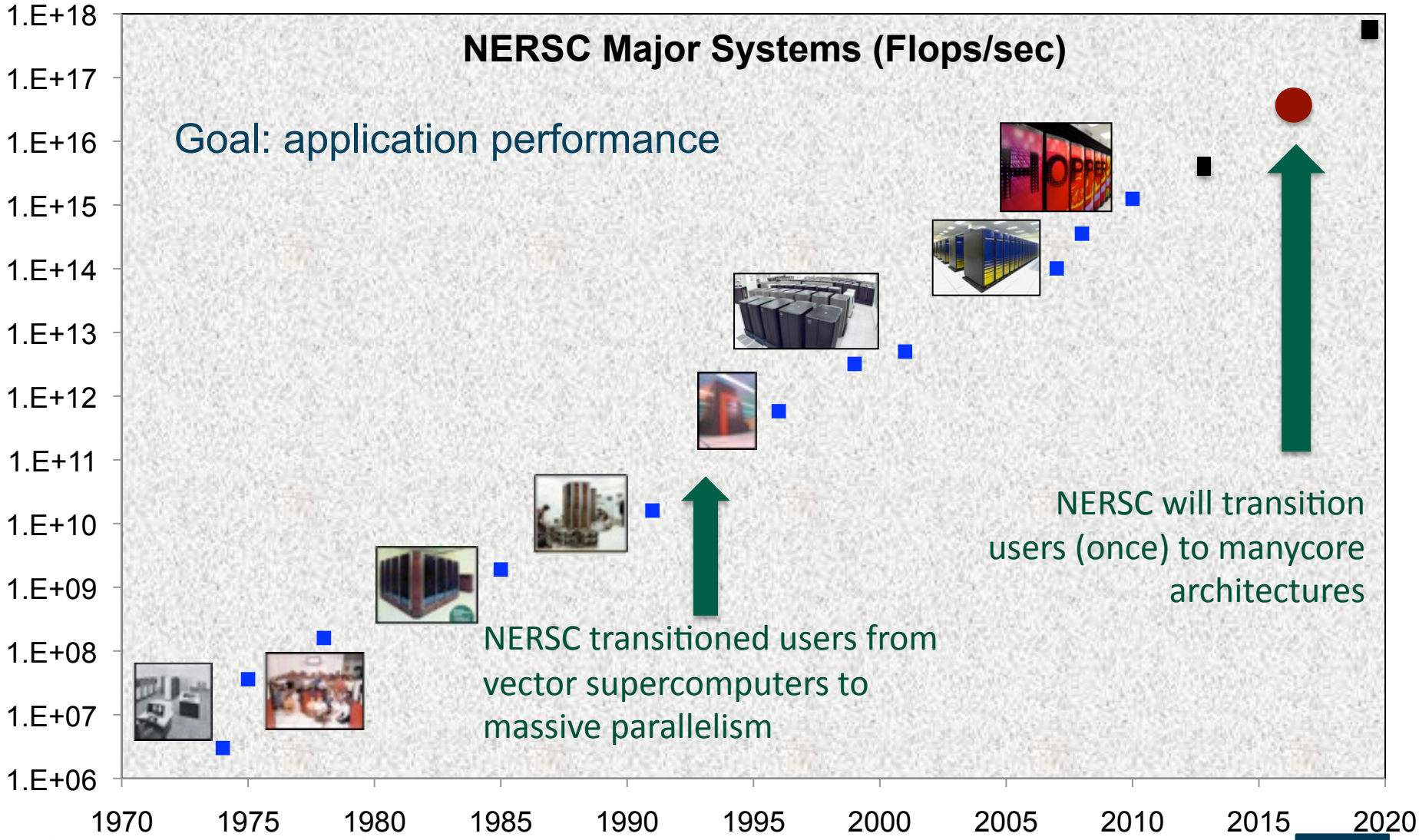
Euclid

(512 GB shared memory)

Dirac 48 Fermi GPU nodes

Magellan Hadoop

NERSC plan will take scientists through technology transition

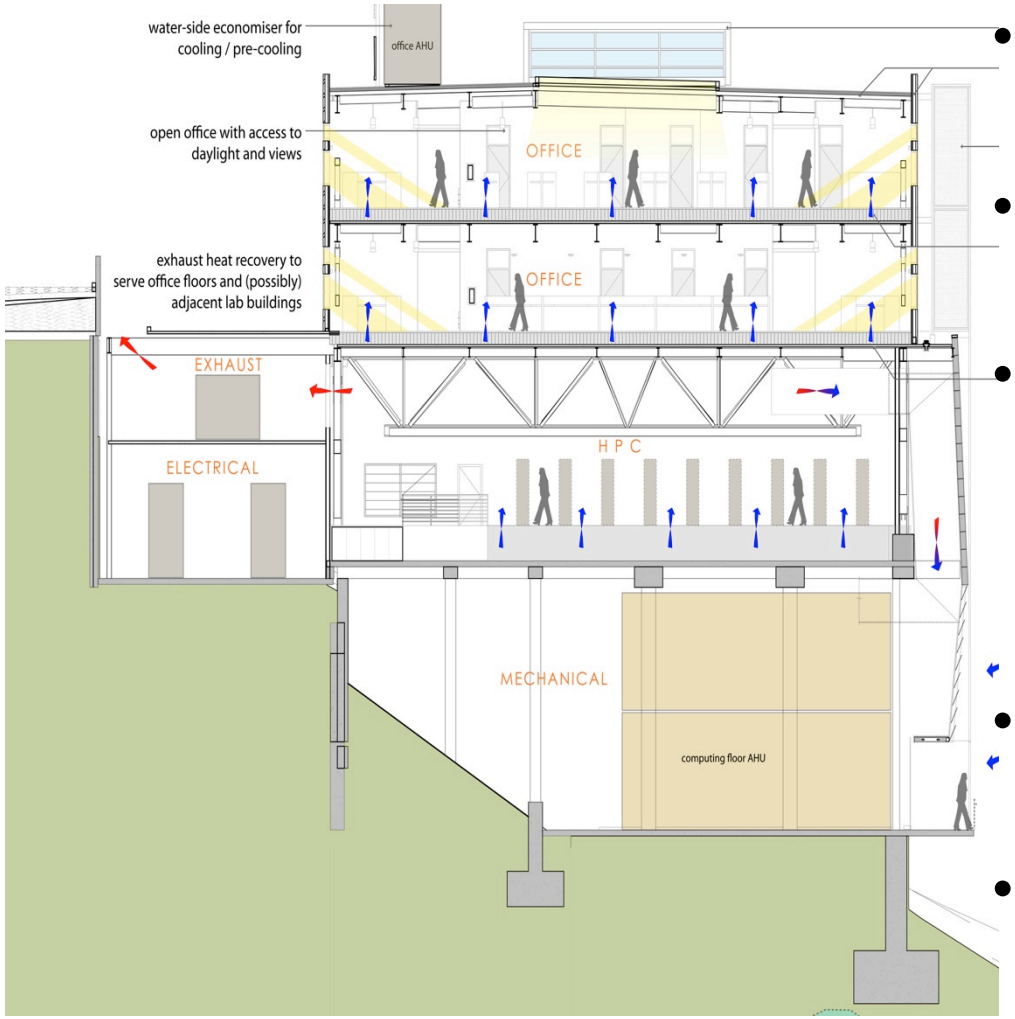


NERSC-8 will be housed in the new CRT facility on the main LBL campus

- **Four story, 140,000 GSF**
  - Two 20k SF office floors, 300 offices
  - 28k SF HPC floor
  - Mechanical floor
- **Occupancy Fall 2014**
- **12.5 MW power**
- **System plans**
  - N6 (Hopper) remains at OSF
  - N7 installed at OSF → CRT
  - N8 installed in CRT

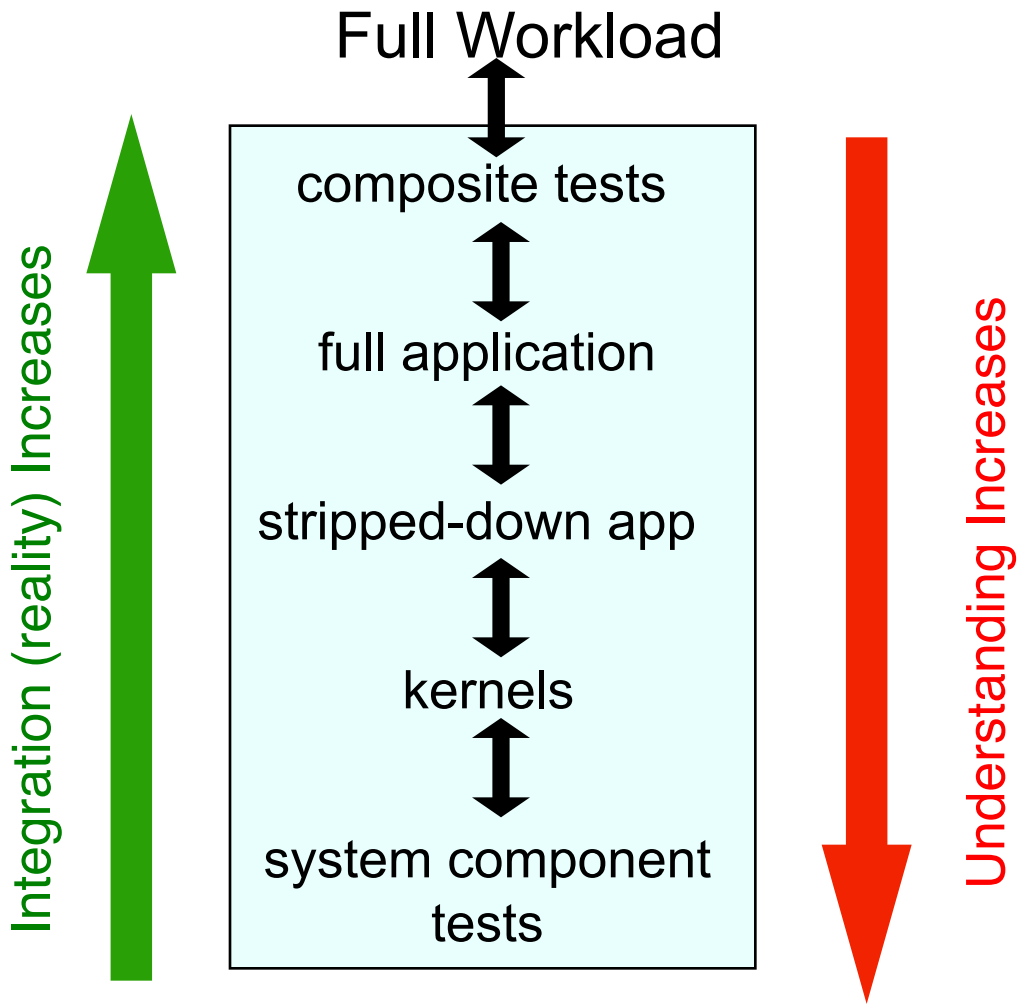
Berkeley's climate along with the CRT design enable an extremely energy efficient building

Building Section



- **Power Usage Effectiveness (PUE): 1.1**
- **Air cooling with 75F degree air year round without chillers**
- **Liquid cooling**
  - 74F degree water year round without chillers
  - 65F degree water using chillers for 560 hours/year (6%)
- **Computer room exhaust heat used to warm office floors**
- **LEED Gold**

In the past NERSC has released benchmarks of various levels of complexity with the RFP

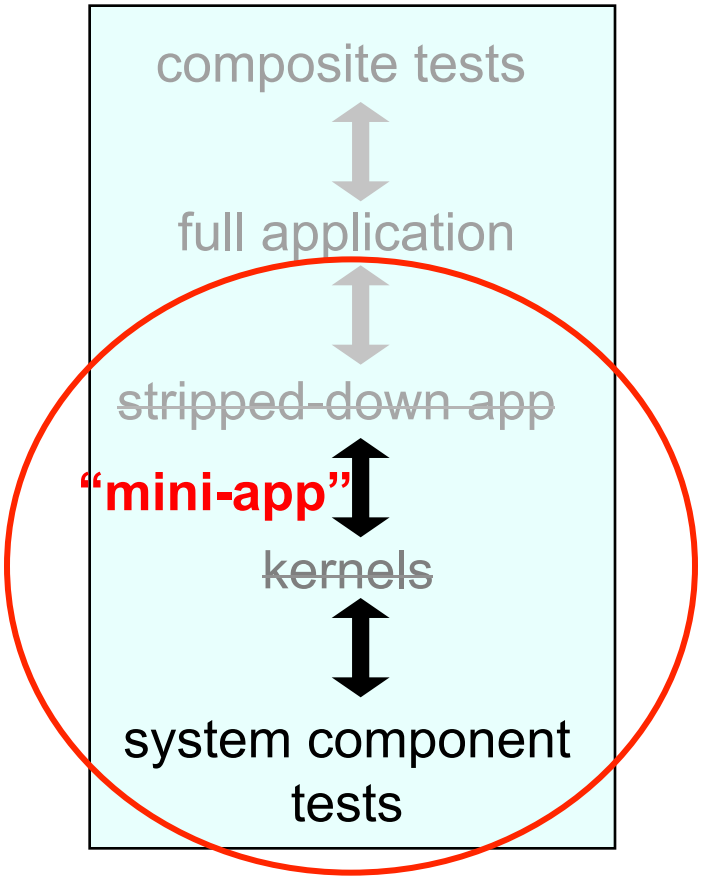


Current technology landscape makes benchmark packaging and selection challenging

- Variety of chip architectures (CPU, GPU, MIC)
- Uncertain programming model (MPI, OpenMP, OpenACC, CUDA, CILK)
- Limited staff to assemble benchmarks, infeasible to release benchmarks compatible for all programming models
- Cognizant of effort required by vendors to run benchmarks
- Collaboration with Trinity means we must find overlap in benchmark selection

NERSC-8/Trinity plan to use "mini-apps" and micro-benchmarks for system evaluation

- **Releasing full benchmarks for all architectures and programming models infeasible**
- **Plan is to release mini-apps with MPI+OpenMP**
- **Full applications and NERSC SSP will remain a part of our acceptance testing**



Benchmarks are critical part of the NERSC-8 procurement

- Distinguish performance of systems
- Compare price and performance metrics such as Flops/$ and Flops/Watt
- Represent scientific workload on system
- Give confidence that chosen system will perform well for NERSC workload
- Used throughout lifetime of the system

Proposed NERSC-8/Trinity MiniApp Benchmarks

MiniApp	Descirption
miniFE	Unstructured implicit finite element
miniGhost	Finite difference stencil
miniContact	Contact search
AMG	Algebraic Mult-Grid linear system solver for unstructured mesh physics packages
UMT	Unstructured-Mesh deterministic radiation Transport
miniPartisn	Structured Particle Transport Surrogate
miniDFT	Density Functional Theory (DFT)
GTC	Particle-in-cell magnetic fusion
MILC	Lattice Quantum Chromodynamics (QCD). Sparse matrix inversion, CG

Note: this list is not final and could change before final RFP release

Other Important Criteria

- Focus on performance for real applications
- Energy efficiency
- Application portability
- Ease of programming
- Scalable interconnect
- System resilience and reliability
- Active power management
- System facility integration
- System software, management software
- Support model for systems