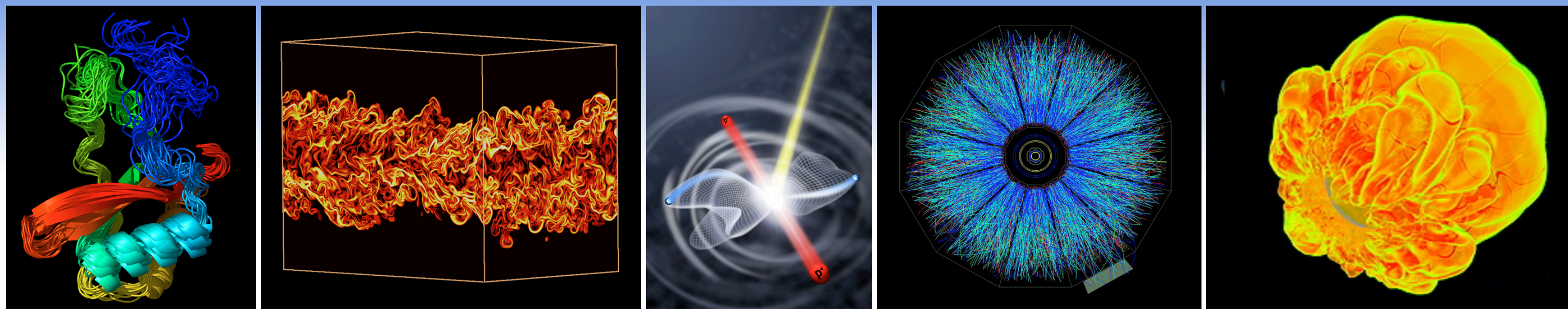


Breakthrough Computing in Petascale Applications and Petascale System Examples at NERSC

John Shalf / Harvey Wasserman
NERSC

Lawrence Berkeley National Laboratory

HPC User Forum April, 2009

Intro to NERSC

- National Energy Research Scientific Computing Center
- Mission: *Accelerate the pace of scientific discovery* by providing high performance computing, information, data, and communications services for all DOE Office of Science (SC) research.
- The production computing facility for DOE SC.
- Berkeley Lab Computing Sciences Directorate
  – Computational Research Division (CRD), ESnet
  – NERSC

ASCR* Computing Facilities

NERSC

LBNL

- High end computing
- Production computing
- DOE/SC needs
- In 2010 controlled by:
  - 5-30% ASCR
  - 60-85% other offices
  - 10% NERSC reserve
- Hundreds of projects

LCFs

ORNL and ANL

- Highest end computing
- Leading edge systems
- All Science, not just DOE
- In 2010 controlled by:
  - 60-85% ANL / ORNL process
  - 5-30% ASCR (in 2010)
  - 10% LCF reserve
- Tens of Projects

U.S. DEPARTMENT OF ENERGY | Office of Science

* OFFICE OF ADVANCED SCIENTIFIC COMPUTING RESEARCH

BERKELEY LAB

PetaScale System Examples at NERSC

- Click to add text

NERSC-6 Project Overview

- Acquire the next major NERSC computing system
  - Goal: 70-100 Sustained TF/s on representative applications (NERSC-6 SSP)
  - Fully-functional machine accepted in FY10
  - 70 TB/s *IOR* I/O bandwidth
  - RFP release September 4, 2008.
  - Today: 13-25 TF SSP on NERSC-5 (Cray XT4, ~20k-40k cores)

?

stable production environment

PetaScale Applications Performance Metric

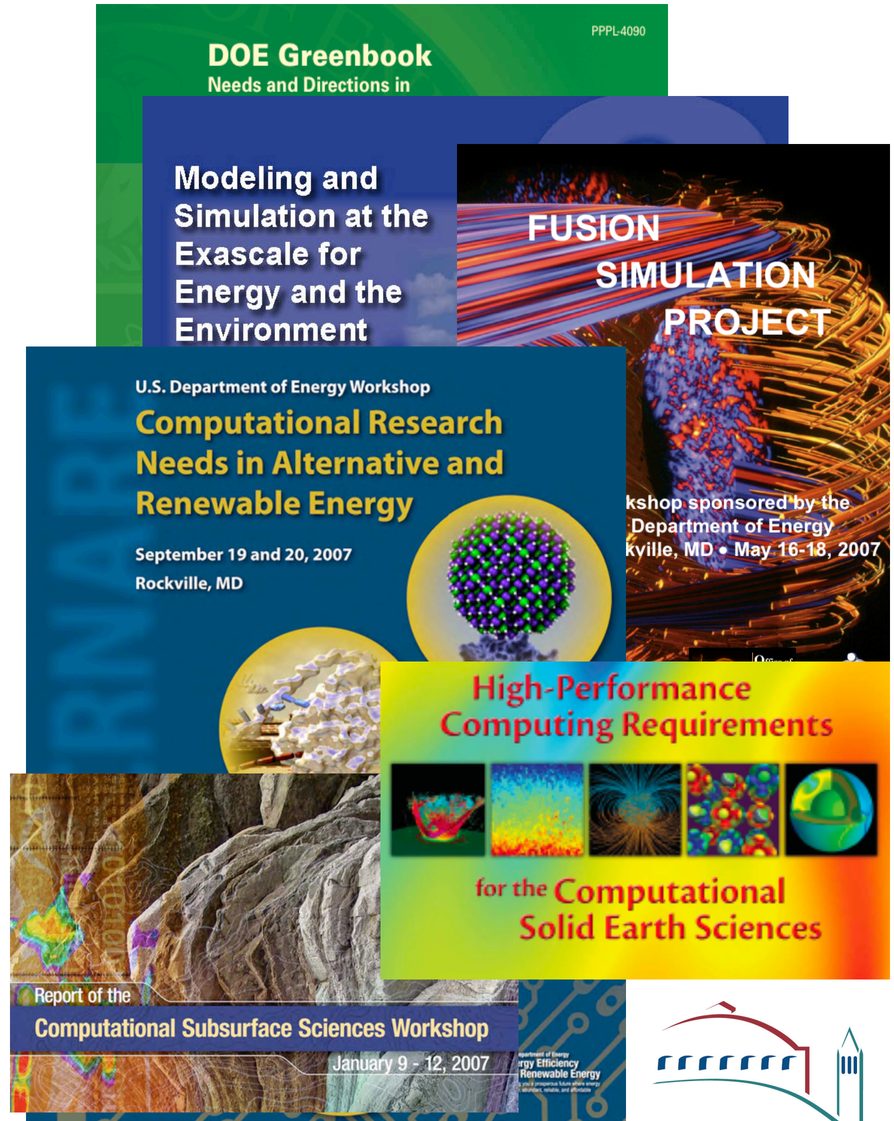
- Sustained System Performance (SSP)
  - Aggregate, un-weighted measure of sustained computational capability relevant to NERSC workload.
  - Geometric Mean of the processing rates of 7 applications multiplied by N, # of cores in the system.

$$\text{SSP in TFLOPS} = \frac{N * \sqrt[7]{\prod_i P_i}}{1000}$$

  - Key ingredient: detailed workload analysis

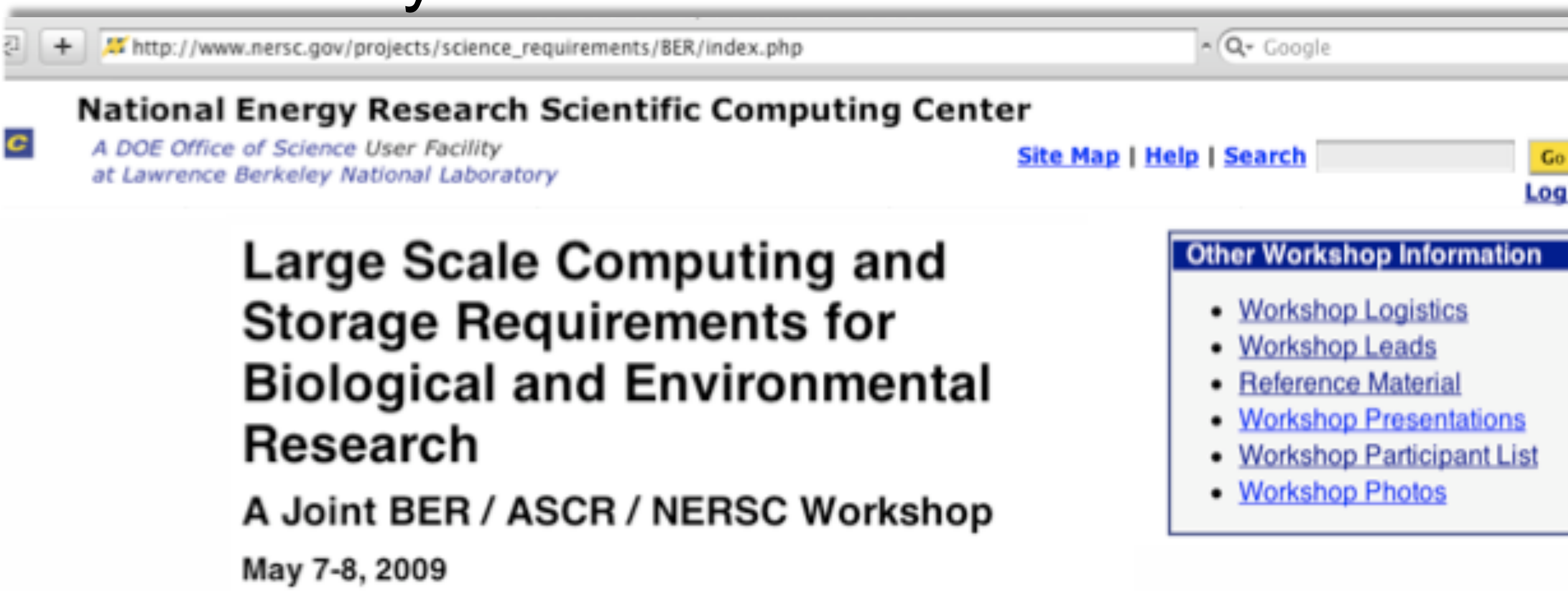
Source of Workload Information

- Documents
  - 2005 DOE Greenbook
  - 2006-2010 NERSC Plan
  - LCF Studies and Reports
  - Workshop Reports
  - 2008 NERSC assessment
- Allocations analysis
- User discussion



New Model for Collecting Requirements

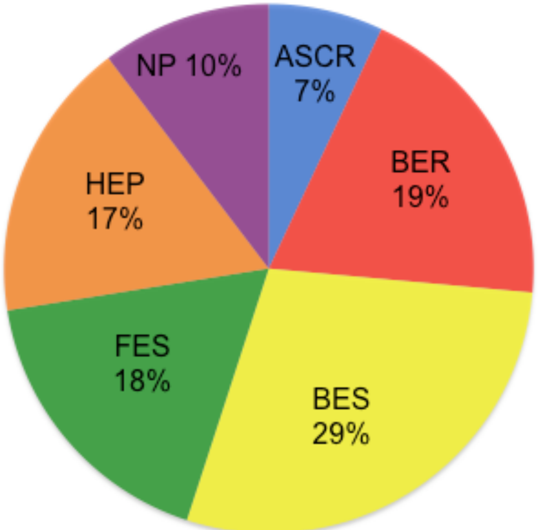
- Joint DOE Program Office / NERSC Workshops
- Modeled after ESnet method
  - Two workshops per year
  - Describe science-based needs over 3-5 years
- Case study narratives



DOE View of NERSC Workload

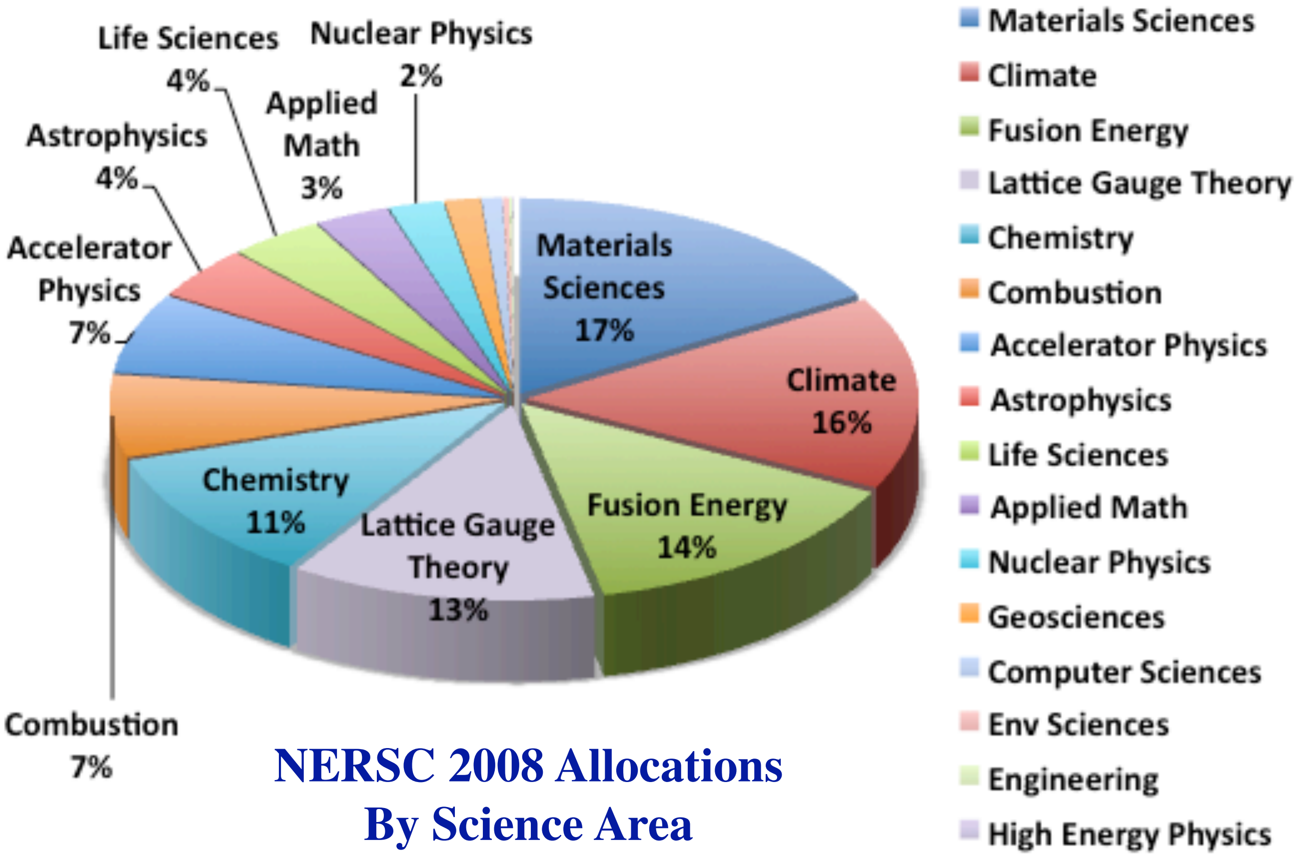
- NERSC serves a large population
  3000 users, 400 projects,
  500 code instances
- Allocations in 2009
  - 10% INCITE program
    - Open to any area, not just DOE/SC
    - Peer review process run by ASCR
  - 70% Production (ERCAP) awards:
    - From 10K hour (startup) to 5M hour
    - Controlled by DOE program offices
  - 10% each NERSC and DOE/SC reserve
    - Includes NEH and NOAA, JBEI, other Climate
- Focus is high end computing, data services (not mid-range)

2009 Allocations



ASCR	**Advanced Scientific Computing Research**
BER	**Biological & Environmental Research**
BES	**Basic Energy Sciences**
FES	**Fusion Energy Sciences**
HEP	**High Energy Physics**
NP	**Nuclear Physics**

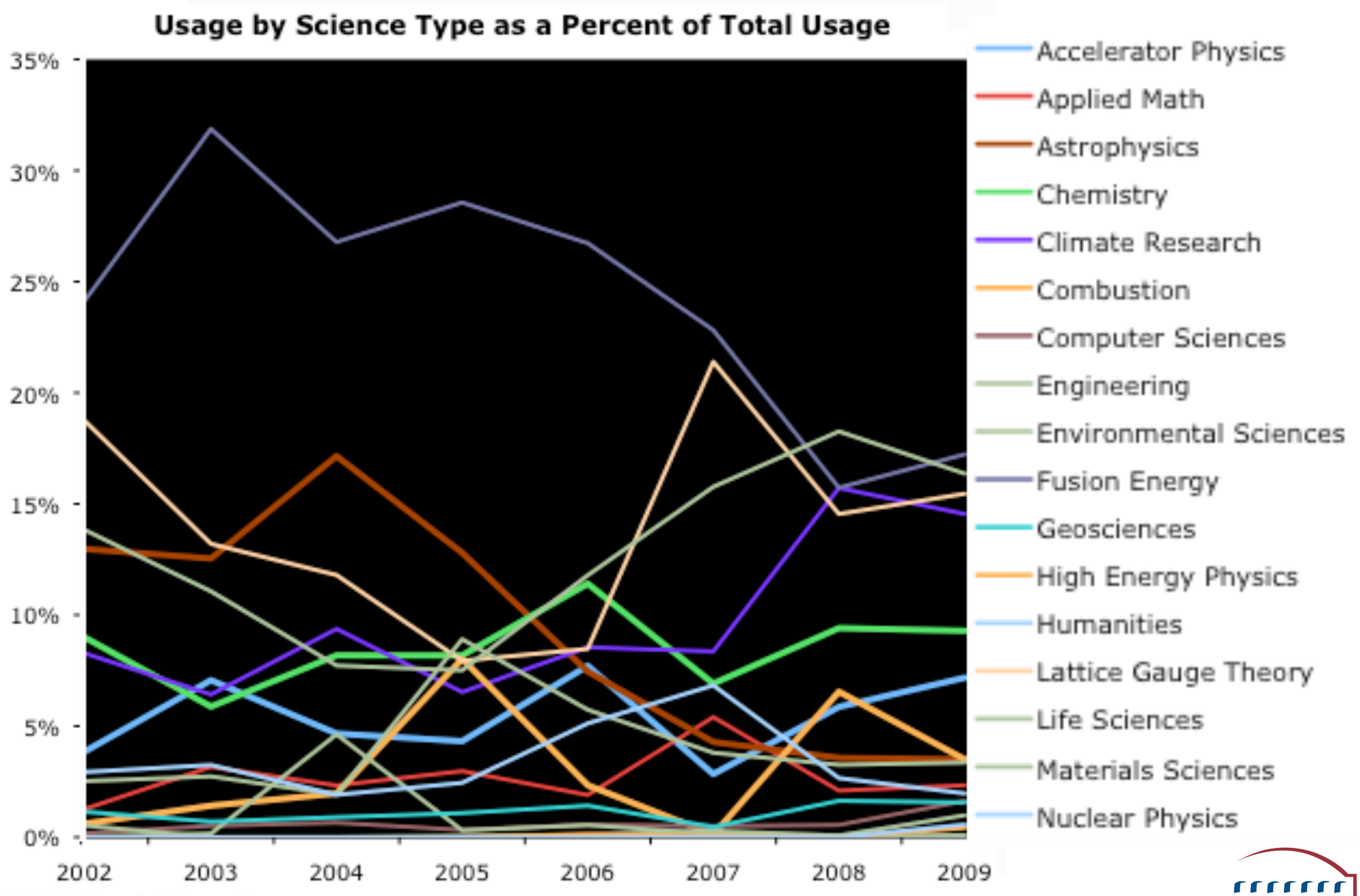
Science View of Workload



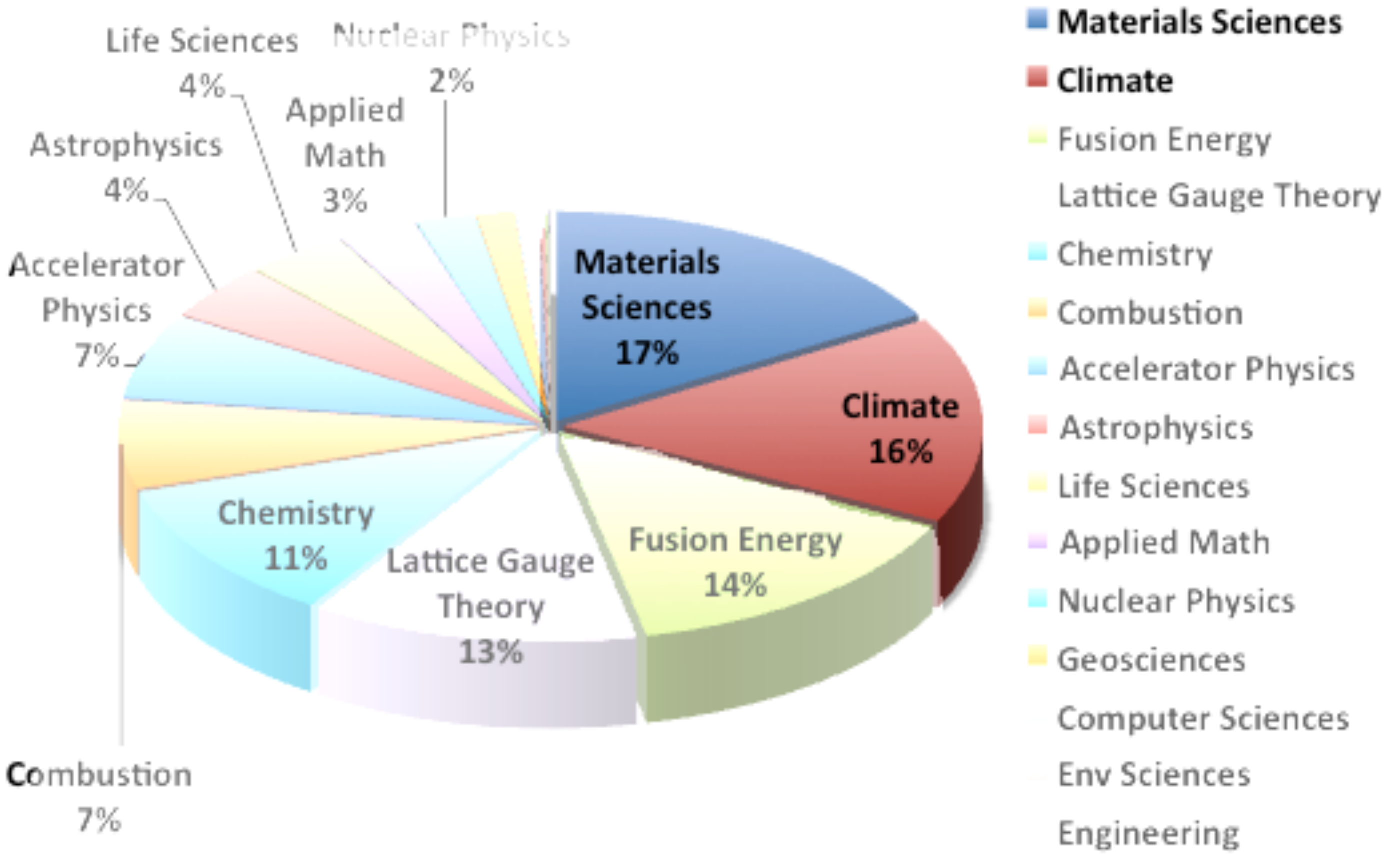
NERSC 2008 Allocations
By Science Area

NERSC Serves Broad DOE Science Priorities

DOE Science Priorities Vary



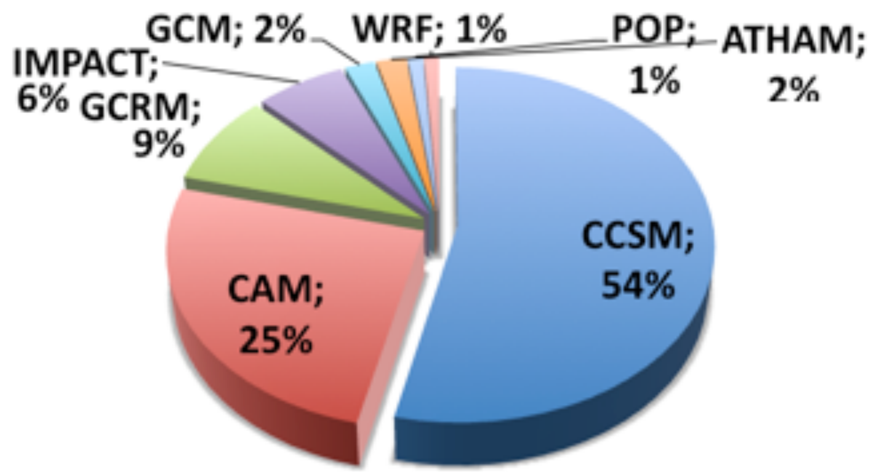
Workload Examples



Example: Climate Modeling

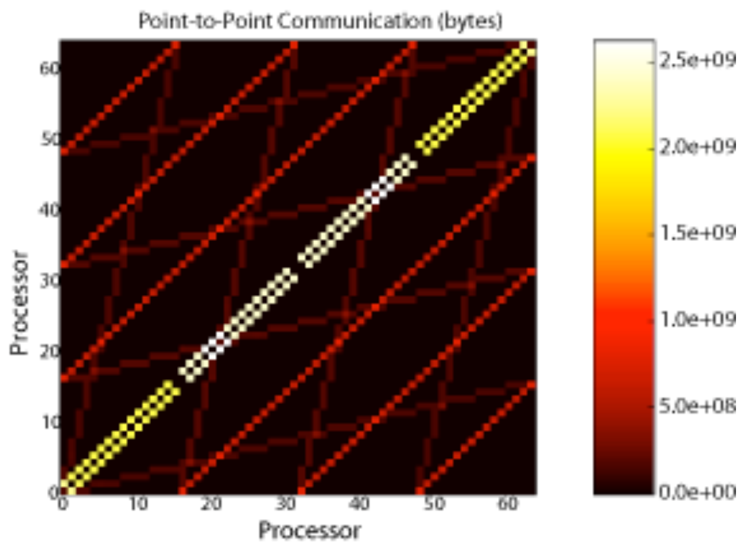
- CAM dominates CCSM3 computational requirements.
- FV-CAM increasingly replacing Spectral-CAM in future CCSM runs.
- Drivers:
  - Critical support of U.S. submission to the Intergovernmental Panel on Climate Change (IPCC).
  - V & V for CCSM-4
- 0.5 deg resolution tending to 0.25

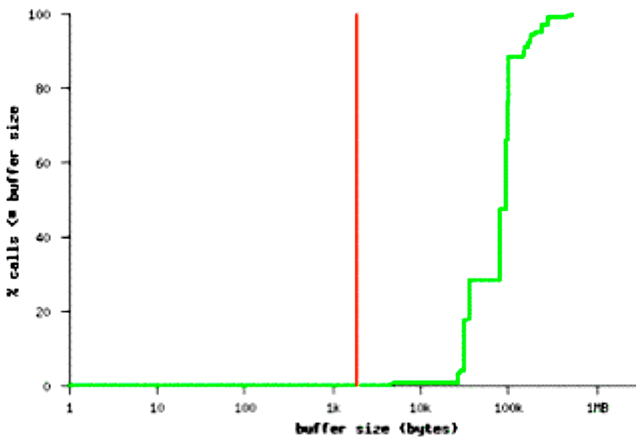
Climate Without INCITE



- Focus on ensemble runs - 10 simulations per ensemble, 5-25 ensembles per scenario, relatively small concurrencies.

FV-CAM Characteristics

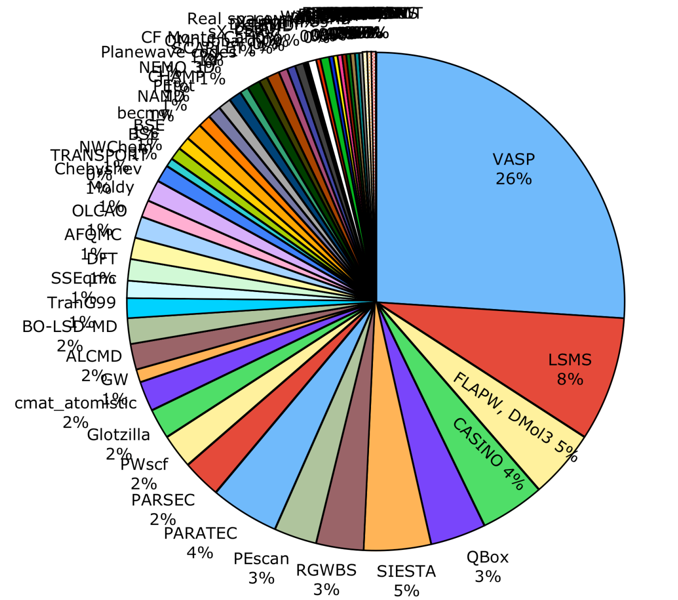


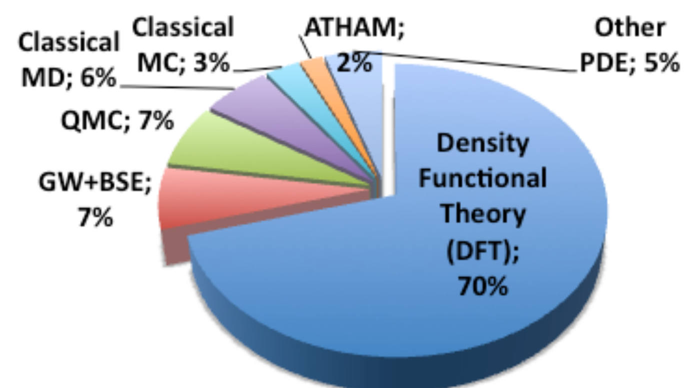


- Unusual interprocessor communication topology – stresses interconnect.
- Relatively low computational intensity – stresses memory subsystem.
- MPI messages in bandwidth-limited regime.
- Limited parallelism.

Example: Material Science

- 62 codes, 65 users.
- Drivers: nanoscience, ceramic xtals, novel materials, quantum dots…
- DFT dominates, usually PW
- VASP, PARATEC, PETOT, QBox
- Libraries: SCALAPACK / FFTW / MPI
- Dominant phases of PW DFT algorithm:
  - 3-D FFT
    - Real / reciprocal space transform via 1-D FFTs
    - O(Natoms2) complexity
  - Subspace Diagonalization
    - O(Natoms3) complexity
  - Orthogonalization
    - dominated by BLAS3
    - ~O(Natoms3) complexity
  - Compute Non-local pseudopotential
    - O(Natoms3) complexity
- Various choices for parallelization

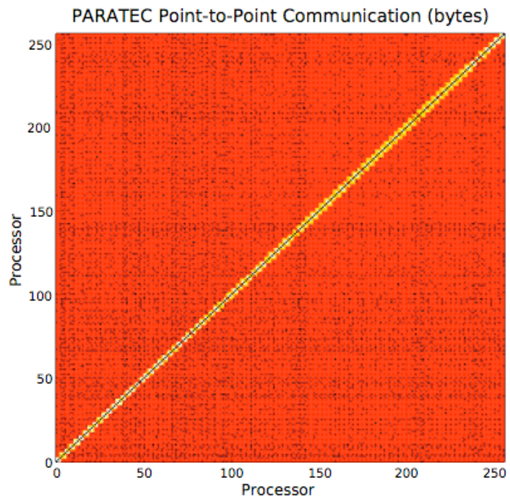


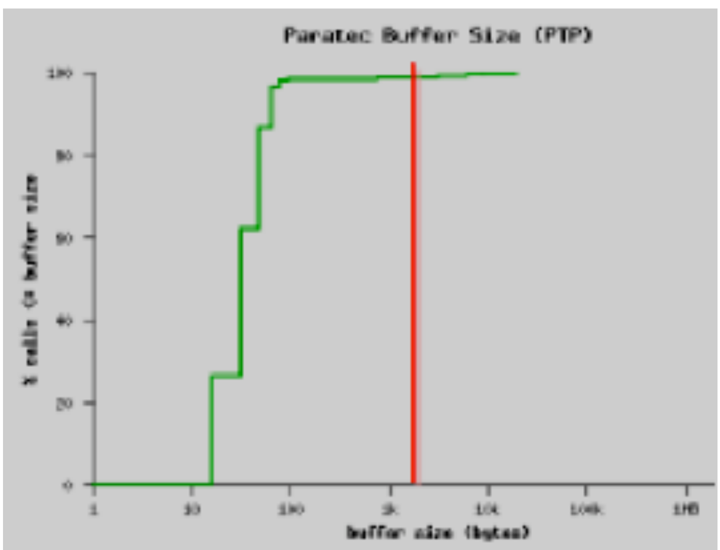


Andrew Canning

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

PARATEC Characteristics





- All-to-all communications
- Strong scaling emphasizes small MPI messages.
- Overall rate dominated by FFT speed and BLAS.
- Achieves high per-core efficiency on most systems.
- Good system discrimination.
- Also used for NSF Trac-I/II benchmarking.

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC-6 Application Benchmarks

Benchmark	Science Area	Algorithm Space	Base Case Concurrency	Problem Description	Lang	Libraries
CAM	Climate (BER)	Navier Stokes CFD	56, 240 Strong scaling	D Grid, (~.5 deg resolution); 240 timesteps	F90	netCDF
GAMESS	Quantum Chem (BES)	Dense linear algebra	256, 1024 (Same as Ti-09)	DFT gradient, MP2 gradient	F77	DDI, BLAS
GTC	Fusion (FES)	PIC, finite difference	512, 2048 Weak scaling	100 particles per cell	F90	
IMPACT-T	Accelerator Physics (HEP)	PIC, FFT component	256,1024 Strong scaling	50 particles per cell	F90	FFTW
MAESTRO	Astrophysics (HEP)	Low Mach Hydro; block structured-grid multiphysics	512, 2048 Weak scaling	16 32^3 boxes per proc; 10 timesteps	F90	Boxlib
MILC	Lattice Gauge Physics (NP)	Conjugate gradient, sparse matrix; FFT	256, 1024, 8192 Weak scaling	8x8x8x9 Local Grid, ~70,000 iters	C, assem.	
PARATEC	Material Science (BES)	DFT; FFT, BLAS3	256, 1024 Strong scaling	686 Atoms, 1372 bands, 20 iters	F90	Scalapack, FFTW

Challenges for Computing Centers

- Power density is the problem, parallelism is the solution
  - (unless you're content with 2008 application speed).
- Little consensus on parallel programming model.
- Fault tolerance at scale
- Efficient algorithms vs. efficient parallelism
- Balancing systems for broad workload, including data-rich computing

Source: "The Landscape of Parallel Computing Research: A View From Berkeley," http://view.eecs.berkeley.edu/

What Does it Mean for NERSC?

- Short term:
  - Immediate need to select best future machine.
    - Anticipate some bids with accelerators, limited memory
    - 3.5 MW power limit for Oakland Scientific Facility
- Longer term:
  - Need to support existing production user base.
  - Optimizing performance-per-watt necessarily includes consideration of programmability.

What Does it Mean for NERSC?

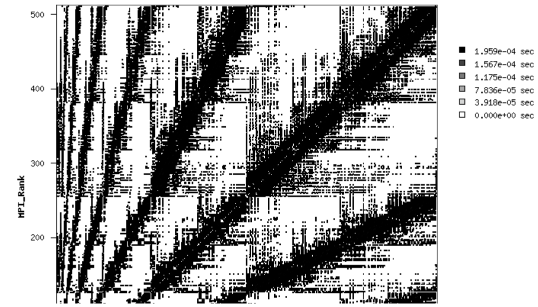
- Longer term: Can we program multicore / manycore?
  - 2 cores for video, 1 for MS Word, 1 for browser, 76 for virus / spam check? *

- Opportunity: Leverage local research in
  - Efficient Algorithms
  - Programming models / languages
  - Tuning methods
  - Power efficient architecture
  - Measurement standards and better quantitative understanding of power issues

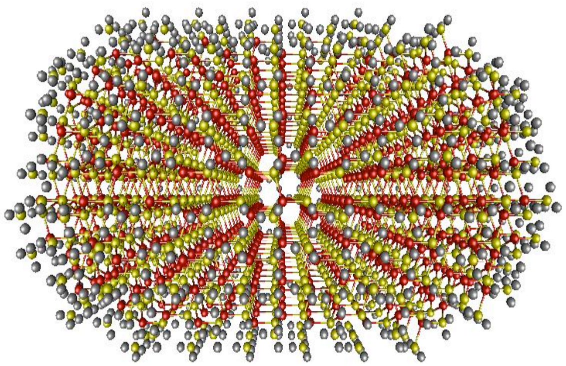
***Source: J. Kubiatowicz, 2-day short course on parallel computing," http://parlab.eecs.berkeley.edu**

Efficient Algorithms

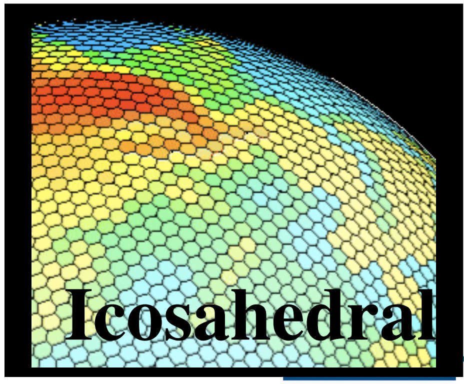
- Astrophysics/Combustion: AMR in MAESTRO
  - S. Woosley (UCSC), J. Bell (LBNL)



- Chemistry/Materials Science: O(n)-scaling codes such as LS3DF
  - L-W. Wang (LBNL)



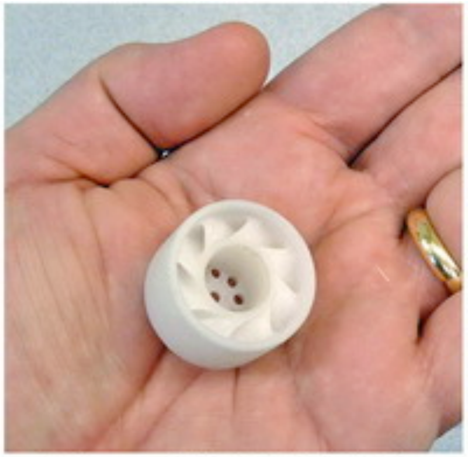
- Climate: icosohedral-grid atmospheric codes
  - D. Randal (Colo.State)



U.S. DEPARTMENT OF ENERGY | Office of Science

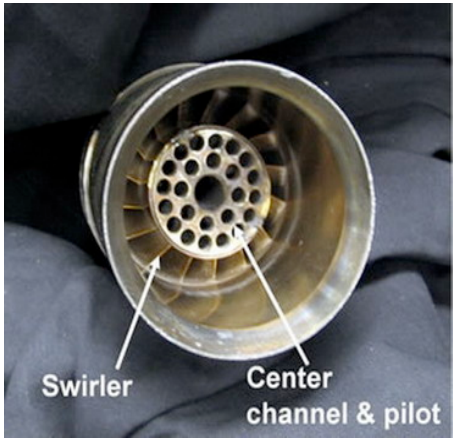
Low-Swirl Burner Simulation

- Low-Swirl Burners invented in 1991 at LBNL.
- Now being developed for near-zero-emission gas turbines (2007 R&D 100 Award)
- Could dramatically reduce pollutants by using special “lean premixed” fuels in power generation and transportation.
- But combustion with these fuels can be highly unstable, making robust systems hard to design.



1" burner (5 kW, 17 KBtu/hr)

28" burner (44 MW, 150 MBtu/hr)



Low-Swirl Burner Simulation

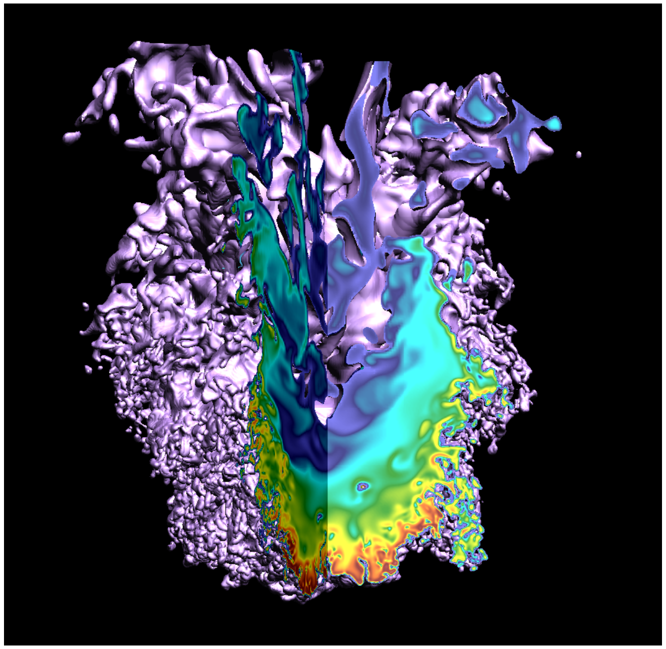
- Numerical simulation of a lean premixed hydrogen flame in a laboratory-scale low-swirl burner. Uses a low Mach number formulation (LMC code), adaptive mesh refinement (AMR) and detailed chemistry and transport.
- PI: John Bell, LBNL

Science Result:

- **Simulations capture cellular structure of lean hydrogen flames and provide a quantitative characterization of enhanced local burning structure**

NERSC Results:

- **LMC dramatically reduces time and memory.**
- **Scales to 4K cores, typically run at 2K**
- **Used 2.2M hours on Franklin in 2007, allocated 3.4M hours in 2008**



J B Bell, R K Cheng, M S Day, V E Beckner and M J Lijewski, Journal of Physics: Conference Series 125 (2008) 012027

Scalable Nanoscience Algorithms

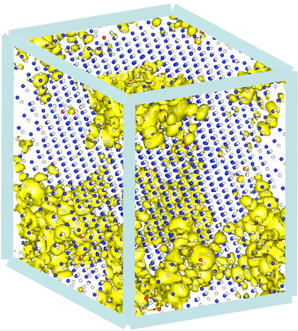
- Calculation: Linear Scaling 3D Fragment (LS3DF). Density Functional Theory (DFT) calculation numerically equivalent to more common algorithm, but scales with O(n) in number of atoms rather than $O(n^3)$
- Lin-Wang Wang, Zhengji. Zhao, LBNL

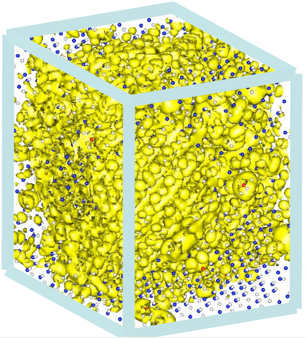
Science Results

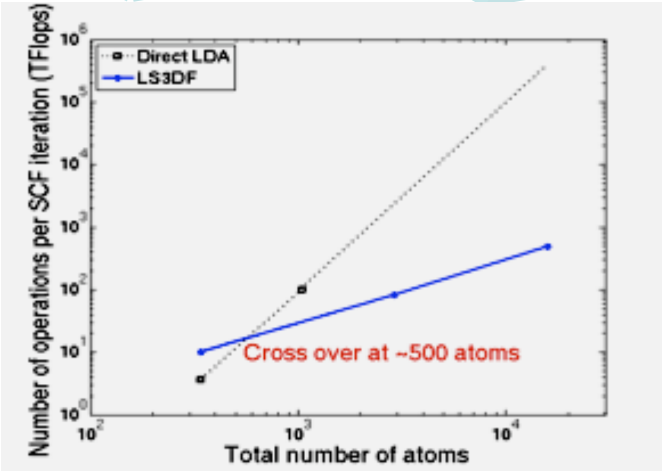
- Calculation of 3500 atom ZnTeO alloy to predict efficiency of a new solar cell material.

Scaling Results

- 36k – 160k cores, XT4, XT5, BG/P
- Took 1 hour vs ~months (est.) for previous $O(n^3)$ algorithm
- Good efficiency (40% of peak)
- **Gordon Bell Prize at SC08**

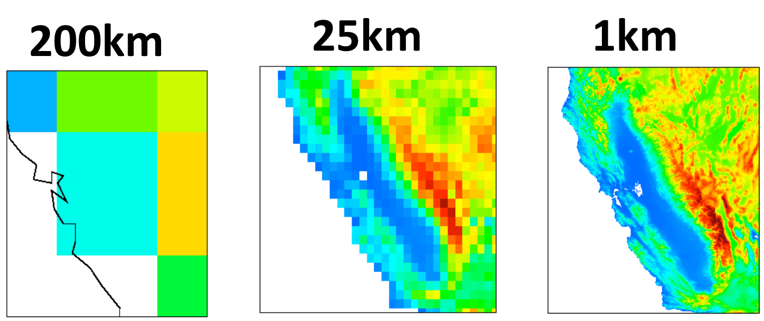


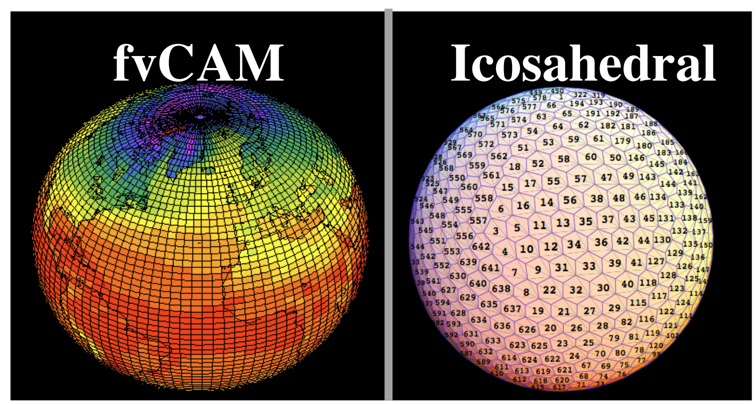




New Approach for Climate Modeling

- Goal: 1-km cloud-resolving model, 1000X real time
- Existing Lat.-long. grid, advection algorithm breaks down before 1km
  - Grid cell aspect ratio at the pole is 10000; time step is problematic at this scale
- Requires new discretization for atmosphere model
  - Partner with Dave Randall (CSU) to use the Icosahedral grid code
  - Uniform cell aspect ratio across globe
  - ~2 million horizontal subdomains, ~20 million total
  - ~5 MB memory per subdomain, ~100 TB total
  - Requires ~10PF sustained, 200 PF peak
- New approach: Green Flash





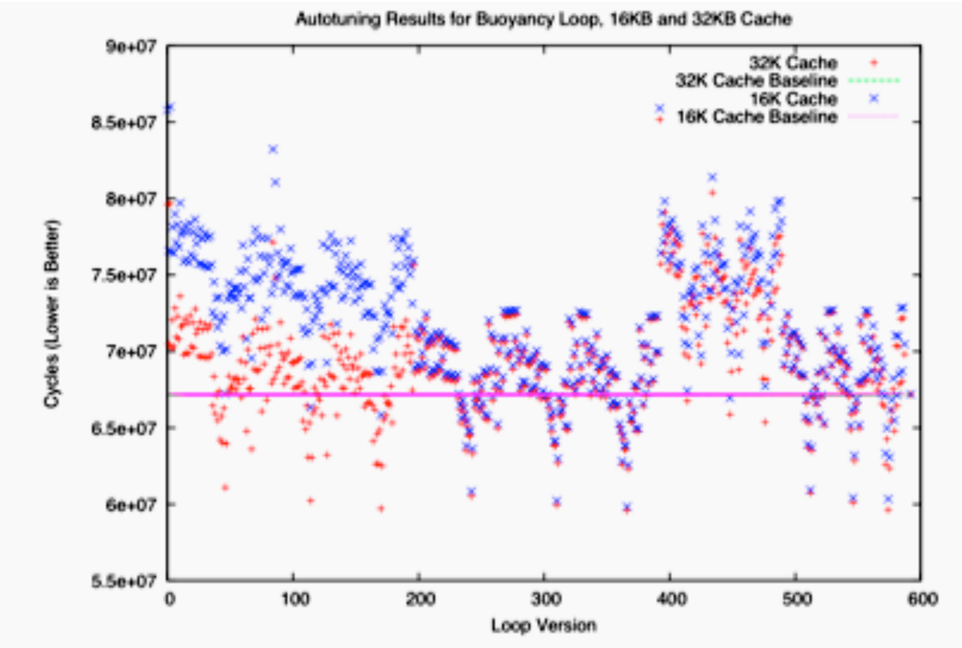
Green Flash Overview

- Research effort, feasibility study
  - Target: 100x better power efficiency; reject Opteron, BG/P approach
- Elements of the approach
  - Choose science target first (climate, now), design machine for it
  - Design (simplified) hardware, software, scientific algorithms together using hardware emulation and auto-tuning
- What is new about this approach
  - Investigate commodity processes used to design power-efficient embedded devices (redirect the tools to benefit scientific computing!)
  - Auto-tuning to map algorithm to complex hardware
  - RAMP: Fast hardware-accelerated emulation of new chip designs

M. Wehner, L. Oliker, and J. Shalf, "Towards Ultra-High Resolution Models of Climate and Weather," Int. J. High Perf. Comp. App, May 2008, 22, No. 2

Current Status: SC08 Demo

- BEE3 board emulating Tensilica Xtensa processor running CSU code
  - 1km scale SubDomain
- Autotuning framework for Tensilica architecture
  - Stencil autotuner can apply ~dozen optimizations



- Moving on multi-core emulation, to explore CMP design trade-offs
  - pack fewer cores in socket to minimize memory bandwidth
  - maximize cores in socket to minimize surface-to-volume ratio

Summary

- GF -> TF highly disruptive (vector to MPI)
- TF -> PF not as disruptive (Fortran/MPI)
- PF -> EF going to very disruptive
  - Uncertain programming model
  - Million-way parallelism
  - Much less memory and lower memory BW
  - Accelerators, unconventional memory hierarchies
  - Must ensure a migration path from current programming approaches to new ones
  - More efficient algorithms, HW, approaches to writing

Questions?

- Please visit
  - the NERSC Website http://www.nersc.gov
  - Green Flash: http://www.lbl.gov/CS/html/greenflash.html
  - O(N) electronic structure: https://hpcrd.lbl.gov/~linwang/
  - NERSC Science Driven System Architecture http://www.nersc.gov/projects/SDSA