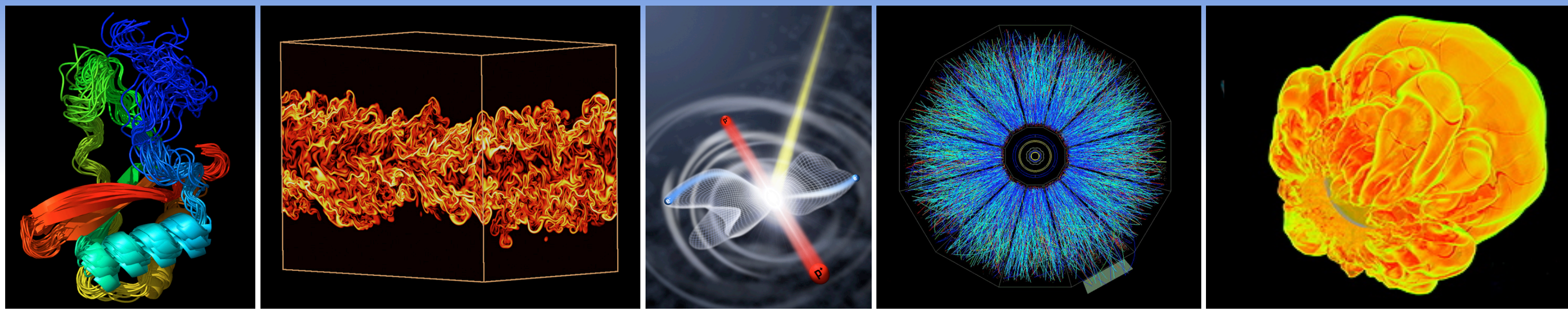


Recent Workload Characterization Activities at NERSC

Harvey Wasserman

NERSC Science Driven System Architecture Group

www.nersc.gov/projects/SDSA

October 15, 2008

Los Alamos Computer Science Symposium

Workshop on Performance Analysis of Extreme-Scale Systems and Applications

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC

BERKELEY LAB

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Acknowledgments

- **Contributions to this talk by many people:**

Kathy Yelick
NERSC Director

Bill Kramer
NERSC General Manager

Katie Antypas
NERSC USG

John Shalf
NERSC SDSA

Erich Strohmaier
LBNL FTG

Lin-Wang Wang
LBNL SCG

Esmond Ng
LBNL SCG

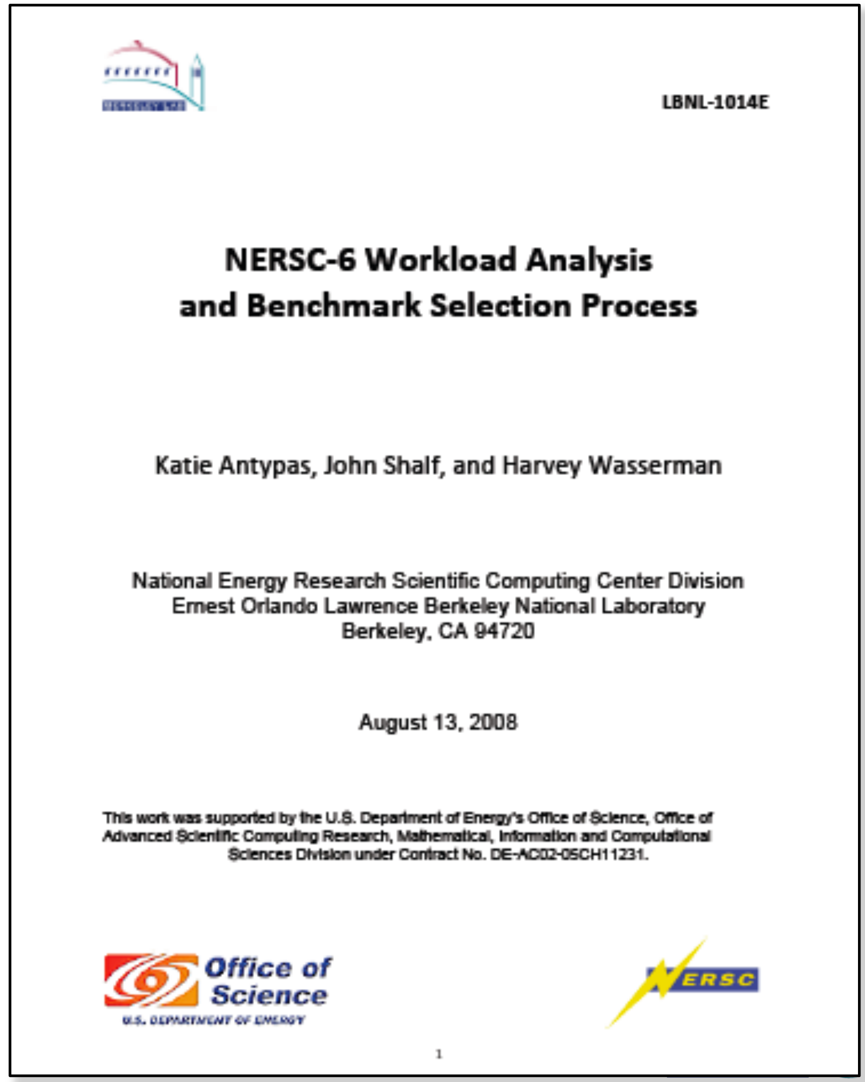
Andrew Canning
LBNL SCG

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Full Report Available

- NERSC Science Driven System Architecture Group
- www.nersc.gov/projects/SDSA/
- Analyze workload needs
- Benchmarking
- Track algorithm / technology trends
- Assess emerging technologies
- Understand bottlenecks
- Use NERSC workload to drive changes in architecture



LBNL-1014E

NERSC-6 Workload Analysis and Benchmark Selection Process

Katie Antypas, John Shalf, and Harvey Wasserman

National Energy Research Scientific Computing Center Division
Ernest Orlando Lawrence Berkeley National Laboratory
Berkeley, CA 94720

August 13, 2008

This work was supported by the U.S. Department of Energy's Office of Science, Office of Advanced Scientific Computing Research, Mathematical, Information and Computational Sciences Division under Contract No. DE-AC02-05CH11231.

Office of Science U.S. DEPARTMENT OF ENERGY

NERSC

1

NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

Motivation

"For better or for worse, benchmarks shape a field."

Prof. David Patterson, UCB CS267 2004

"Benchmarks are only useful insofar as they model the intended computational workload."

Ingrid Bucher & Joanne Martin, LANL, 1982

Science Driven Evaluation

- **Translate scientific requirements into computational needs and then to a set of hardware and software attributes required to support them.**
- **Question: how do we represent these needs so we can communicate them to others?**
  - **Answer: a set of carefully chosen benchmark programs.**

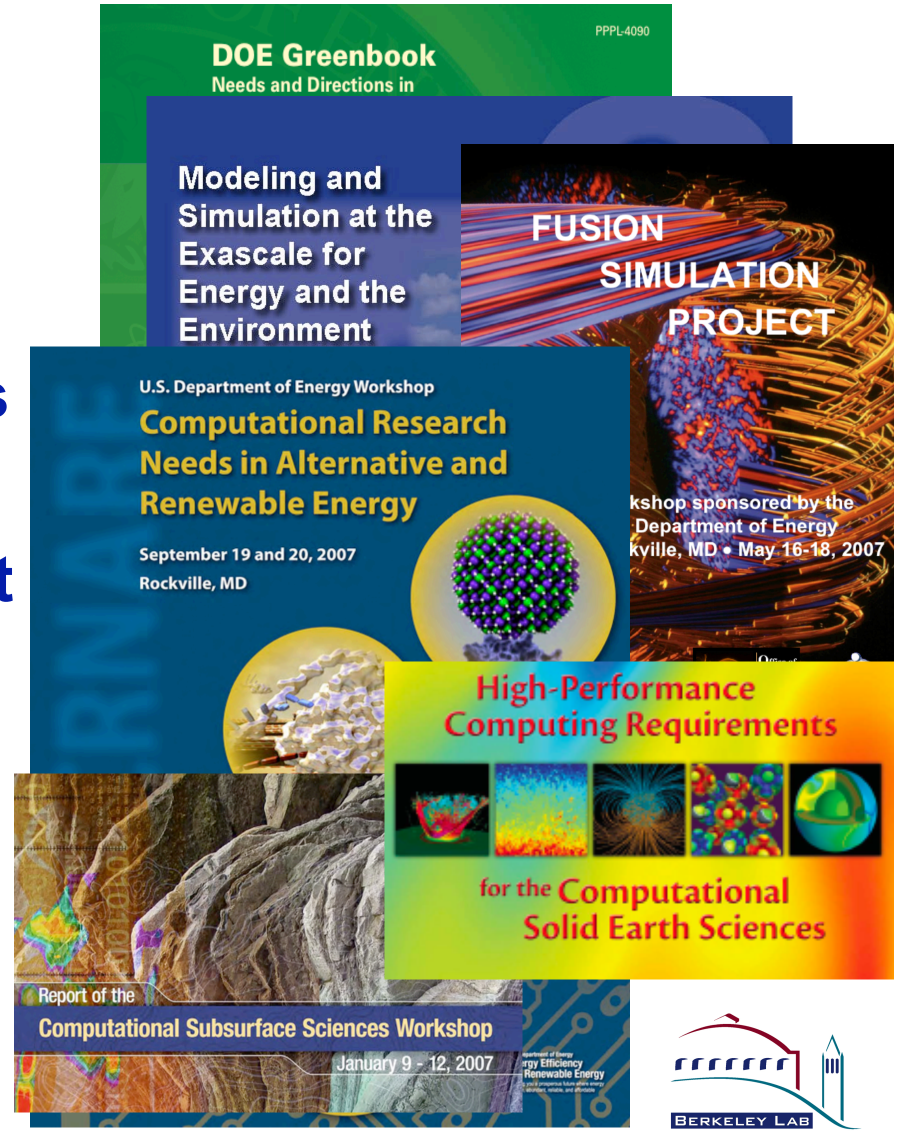
NERSC Benchmarks Serve 3 Critical Roles

- Carefully chosen to represent characteristics of the expected NERSC workload.
- Give vendors opportunity to provide NERSC with concrete performance and scalability data;
  - Measured or projected.
- Part of acceptance test and the basis of performance obligations throughout a system's lifetime.

www.nersc.gov/projects/procurements/NERSC6/benchmarks/

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

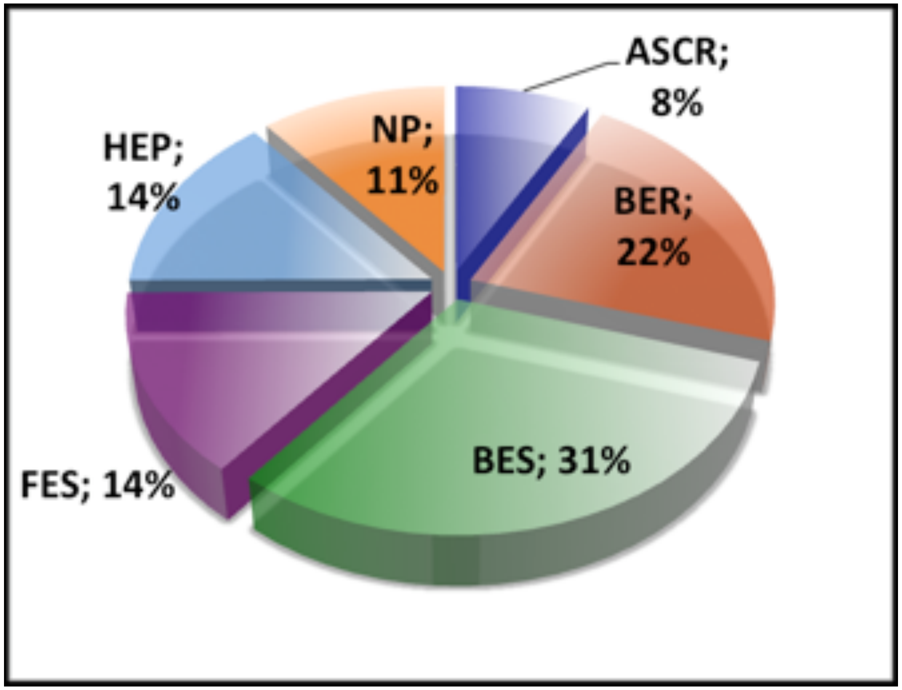
New Model for Collecting Requirements

- Modeled after ESnet activity rather than Greenbook
  - Two workshops per year, initially BER and BES
- Sources of Requirements
  - Office of Science (SC) Program Managers
  - Direct gathering through interaction with science users of the network
  - Case studies, e.g., from ESnet
    - Magnetic Fusion
    - Large Hadron Collider (LHC)
    - Climate Modeling
    - Spallation Neutron Source

NERSC is the Production Computing Facility for DOE SC

- NERSC serves a large population
  - ~3000 users, ~400 projects, nationwide, ~100 institutions
- Allocations managed by DOE
  - 10% INCITE awards: Innovative and Novel Impact on Theory and Experiment
    - Large allocations, extra service
    - Created at NERSC; now used throughout SC
    - Used throughout SC; not just DOE mission
  - 70% Annual Production (ERCAP) awards (10K-5M Hours):
    - Via Call For Proposals; DOE chooses; only at NERSC
  - 10% NERSC and DOE/SC reserve, each
- Award mixture offers
  - High impact through large awards
  - Broad impact across science domains

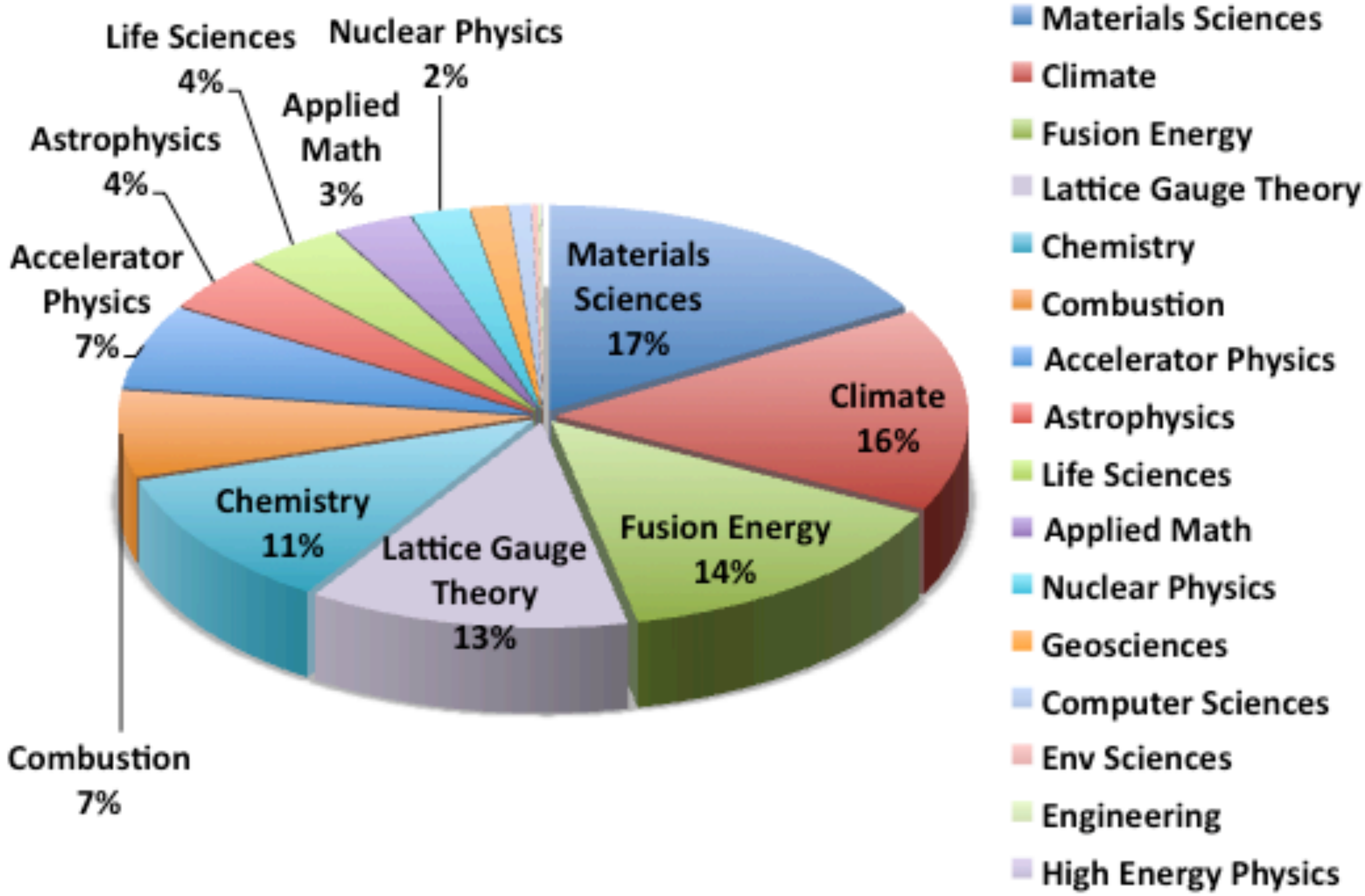
DOE View of Workload



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

NERSC 2008 Allocations By DOE Office

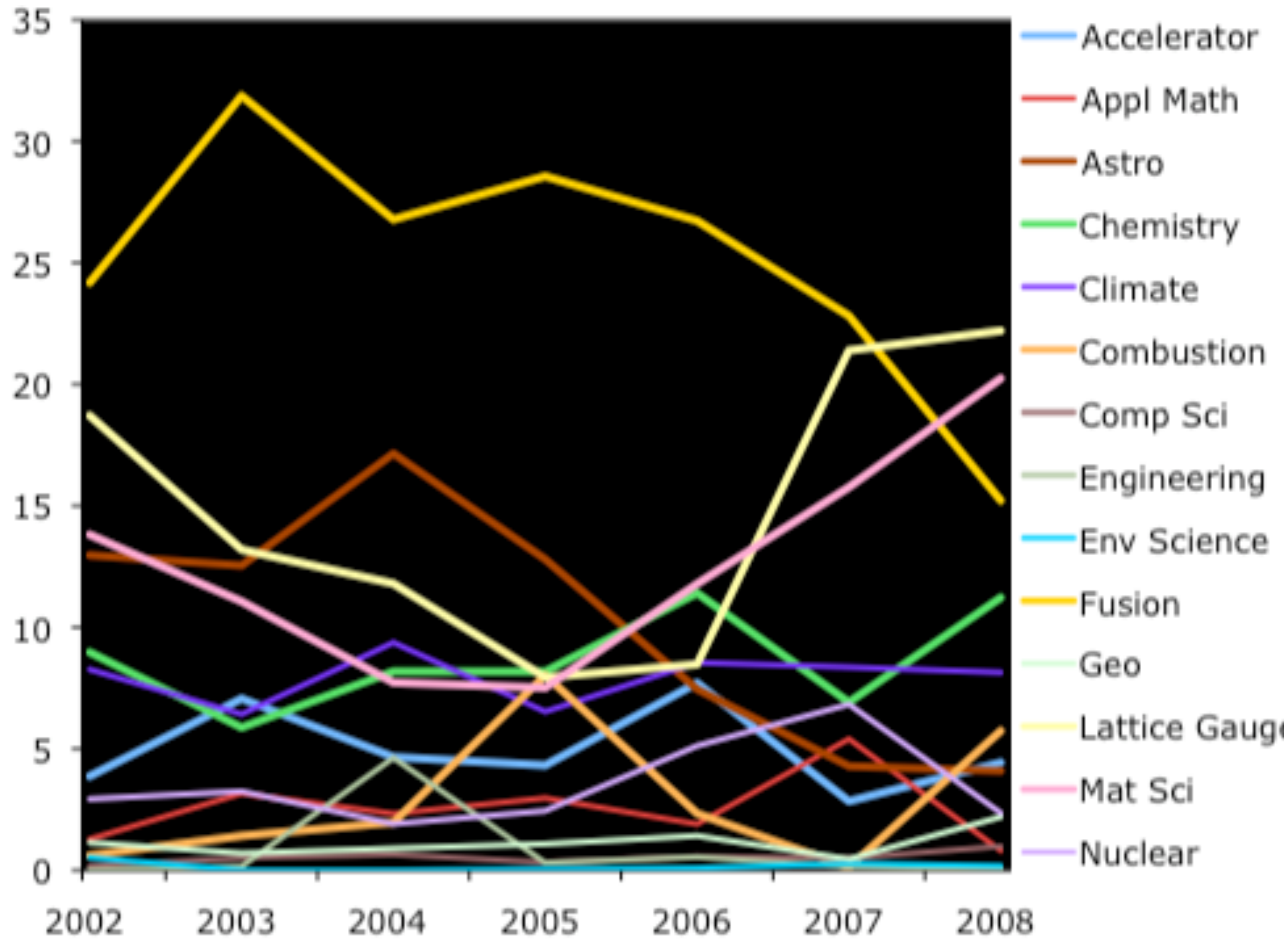
Science View of Workload



NERSC 2008 Allocations
By Science Area (Including INCITE)

Science Priorities are Variable

Usage by Science Area as a Percent of Total Usage



NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Code / Needs by Science Area

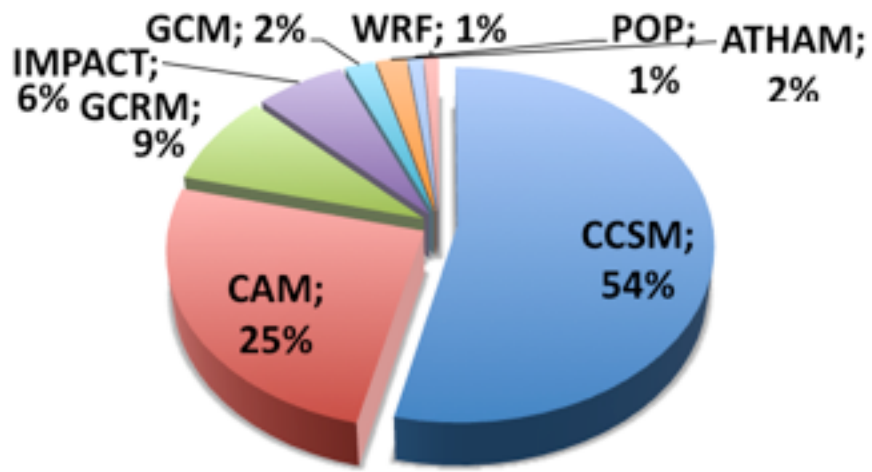
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

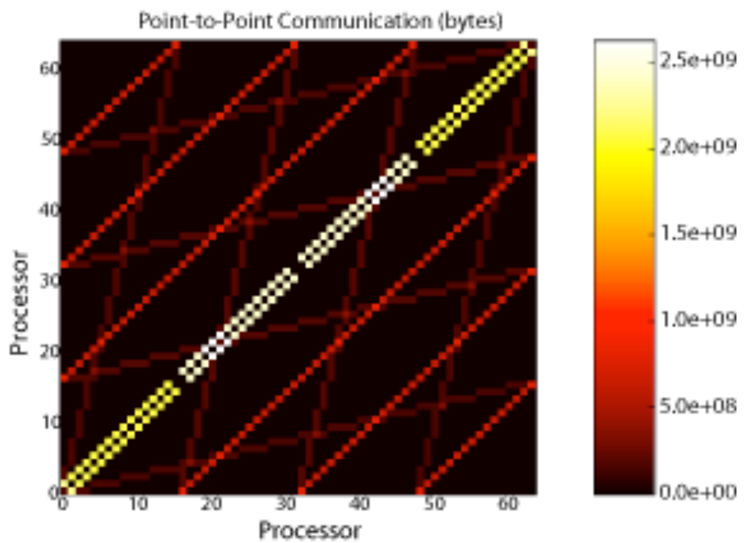
Example: Climate Modeling

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

Climate Without INCITE



fvCAM Characteristics



*Computational intensity is the ratio of # of Floating Point Operations to # of memory operations.

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

Future Climate Computing Needs

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

- **New grids**
- **Cloud resolving models –**
  - *Requires 10^7 improvement in computational speed*
- **New chemistry**
- **Spectral elements / HOMME**
- **Target 1000X real time**
- **=> all point to need for higher per-processor sustained performance**
  - **counter to current microprocessor architectural trends**

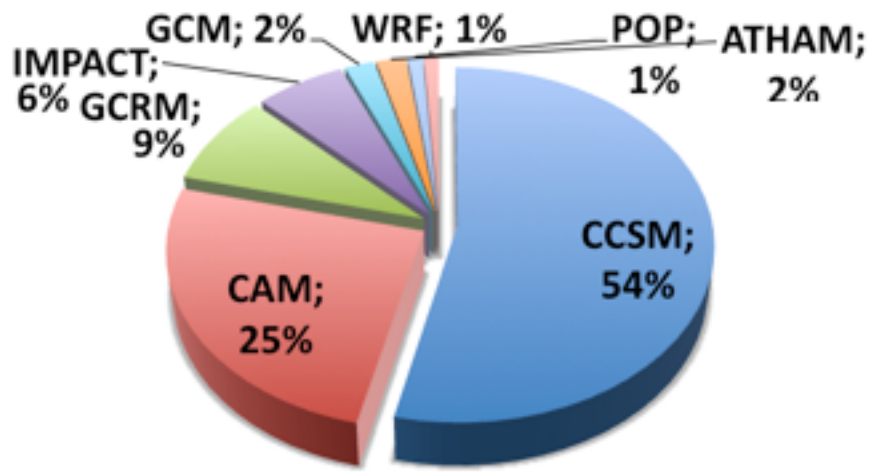
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

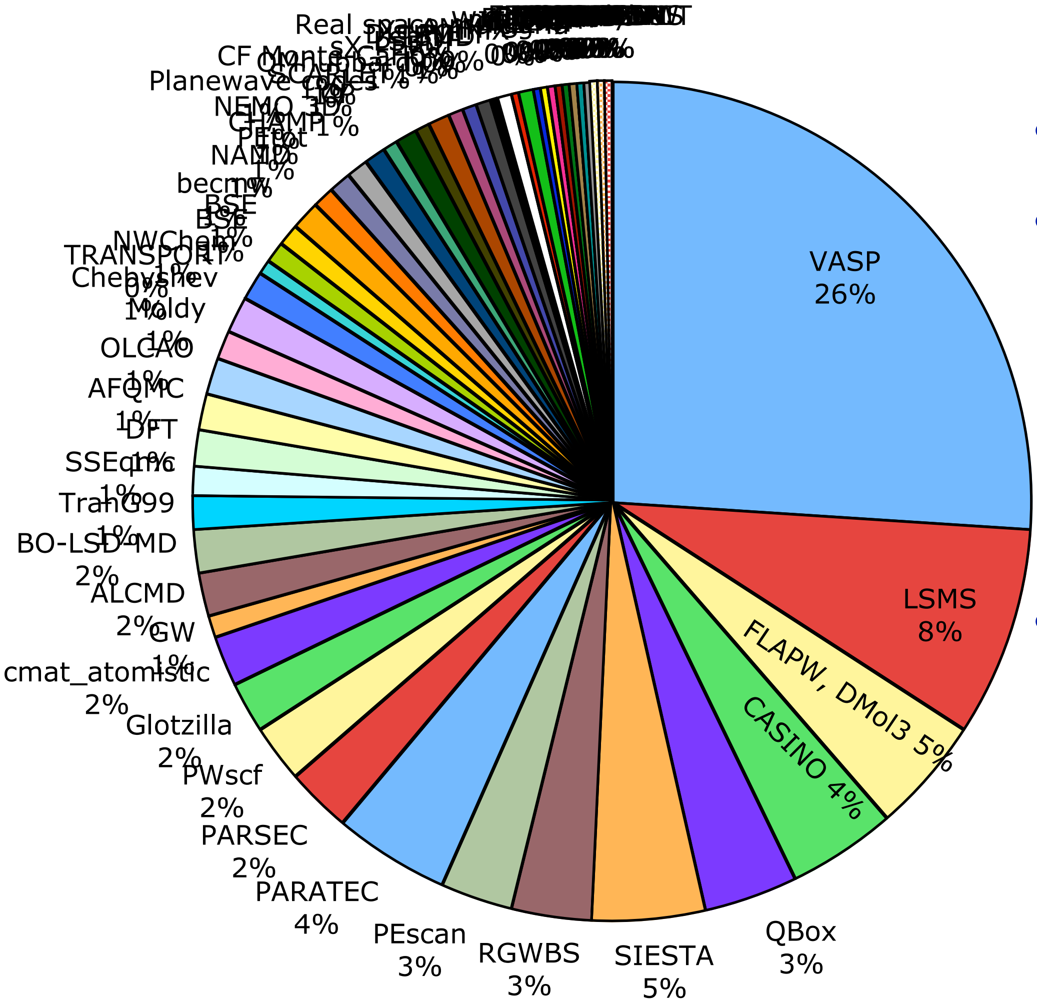
Example: Climate Modeling

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

Climate Without INCITE



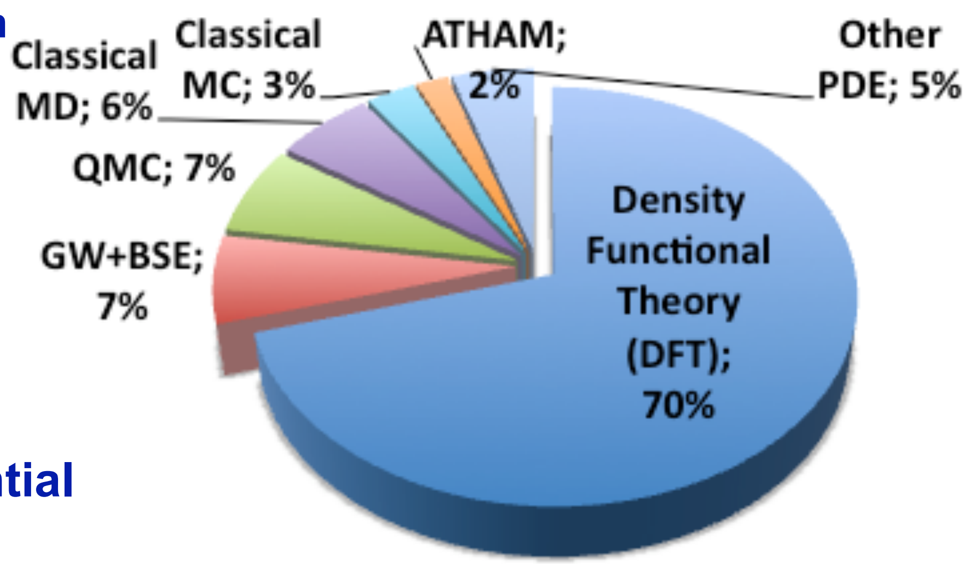
Material Science by Code



- **7,385,000 MPP hours awarded**
- **62 codes, 65 users**
- **same code used by different users => typical code used in 2.15 allocation requests**
- **Science drivers: nanoscience, ceramic crystals, novel materials, quantum dots, …**

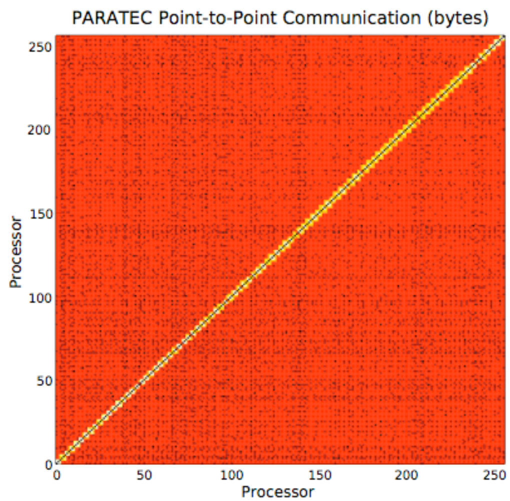
Materials Science by Algorithm

- **Density Functional Theory (DFT) dominates**
  - Most commonly uses plane-wave (Fourier) wavefunctions
  - Most common code is VASP; also PARATEC, PETOT, and Qbox
  - Libraries: SCALAPACK / FFTW / MPI
- **Dominant phases of planewave DFT algorithm**
  - **3-D FFT**
    - **Real / reciprocal space transform via 1-D FFTs**
    - **O(Natoms^2) complexity**
  - **Subspace Diagonalization**
    - **O(Natoms^3) complexity**
  - **Orthogonalization**
    - **dominated by BLAS3**
    - **~O(Natoms^3) complexity**
  - **Compute Non-local pseudopotential**
    - **O(Natoms^3) complexity**
- **Various choices for parallelization**



Analysis by Lin-Wang Wang, A. Canning, LBNL

PARATEC Characteristics



	256 cores	1024 cores
Total Message Count	428,318	1,940,665
16 <= MsgSz < 256		114,432
256 <= MsgSz < 4KB	20,337	1,799,211
4KB <= MsgSz < 64KB	403,917	4,611
64KB <= MsgSz < 1MB	1,256	22,412
1 MB <= MsgSz < 16MB	2,808	

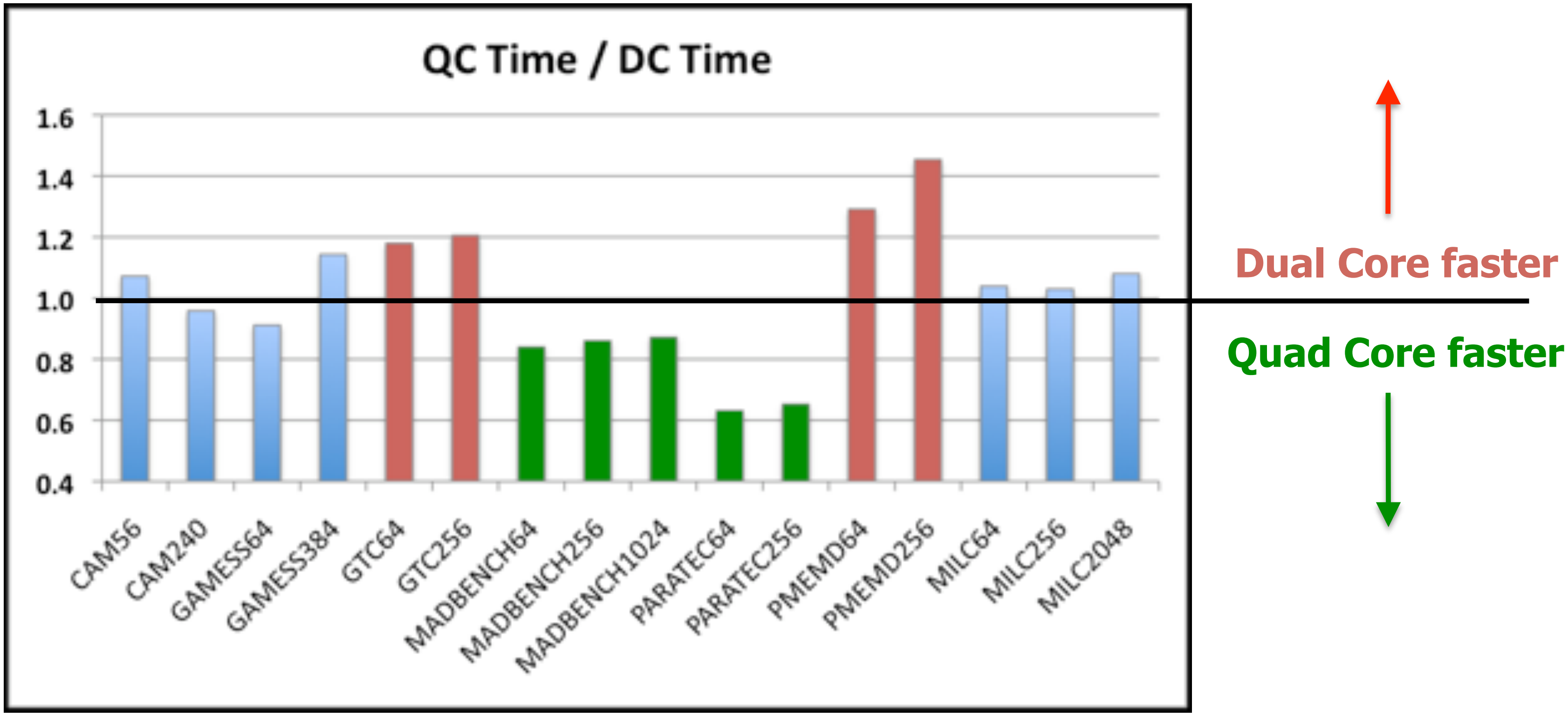
- All-to-all communications
- Strong scaling emphasizes small MPI messages.
- Overall rate dominated by FFT speed and BLAS.
- Achieves high per-core efficiency on most systems.
- Good system discrimination.
- Also used for NSF Trac-I/II benchmarking.

Performance of CRAY XT4

- NERSC “Franklin” system
- Undergoing dual-core -> quad-core upgrade
  - ~19,344 cores to ~38,688
  - 667-MHz DRAM to 800-MHz DRAM
- Upgrade done in phases “in-situ” so as not to disrupt production computing.

Initial QC / DC Comparison

NERSC-5 Benchmarks



Compare time for *n* cores on DC socket to time for *n* cores on QC socket.

Data courtesy of Helen He, NERSC USG

PARATEC: Performance

Medium Problem (64 cores)

	Dual Core	Quad Core	Ratio
FFTs[1]	425	537	1.3
Projectors[1]	4,600	7,800	1.7
Matrix-Matrix[1]	4,750	8,200	1.7
Overall[2]	2,900 (56%)	4,600 (50%)	1.6

- [1] **Rates in MFLOPS/core from PARATEC output.**
- [2] **Rates in MFLOPS/core from NERSC-5 reference count.**
- **Projector/Matrix-Matrix rates dominated by BLAS3 routines.**

=> SciLIB takes advantage of wider SSE in Barcelona-64.

PARATEC: Performance

	FFT Rate	Projector Rate	Overall
XT42.6 Dual-Core	198	4,524	671 (50%)
XT42.3 Quad-Core	309	7,517	1,076 (46%)
XT42.1 Quad-Core	270	6,397	966 (45%)
BG/P	207	567	532 (61%)
HLRB-II	194	993	760 (46%)
BASSI IBM p575	126	1,377	647 (33%)

HLRB-II is an SGI Altix 4700 installed at LRZ, dual-core Itanium with NUMAlink4 Interconnect (2D Torus based on 256/512 core fat trees)

- NERSC-5 “Large” Problem (256 cores)
- FFT/Projector rates in MFLOPS per core from PARATEC output.
- Overall rate in GFLOPS from NERSC-5 official count
- Optimized version by Cray, un-optimized for most others
- Note difference between BASSI, BG/P, and Franklin QC

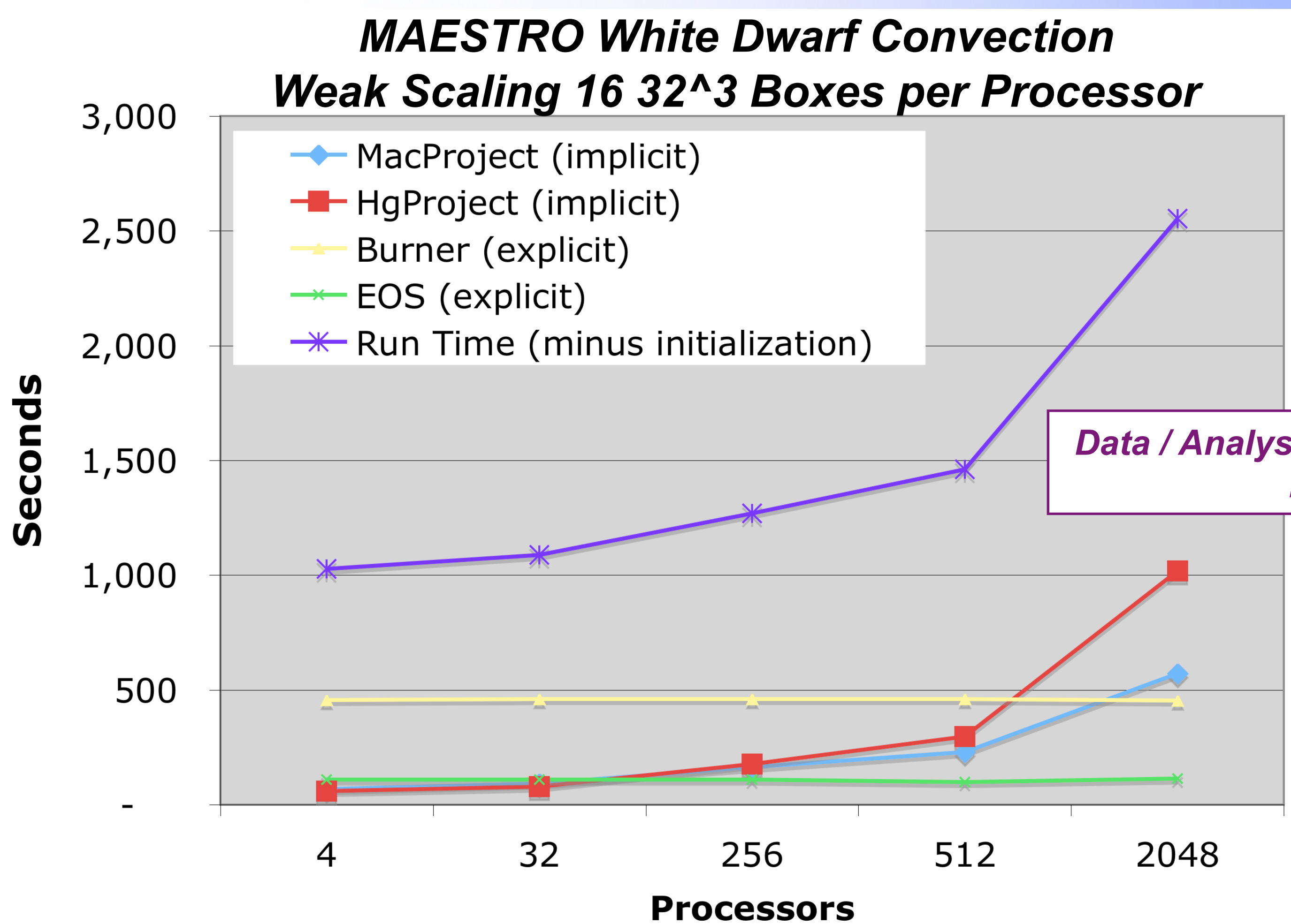
Response to Technology Trends

- **Parallel computing has thrived on weak-scaling for past 15 years**
- **Flat CPU performance increases emphasis on strong-scaling**
- **Benchmarks changed accordingly**
  - **Concurrency:** *Increased 4x over NERSC-5 benchmarks*
  - **Strong Scaling:** *Input decks emphasize strong-scaled problems*
  - **Implicit Methods***: Added MAESTRO application benchmark*
  - **Multiscale**: *Added AMR Poisson benchmark*
  - **Lightweight Messaging:** *Added UPC FT benchmark*

MAESTRO: Low Mach Number Flow

- **Authors**: LBNL Computing Research Division; SciDAC07
- **Relation to NERSC Workload:**
  - Model convection leading up to Type 1a supernova explosion;
  - Method also applicable to 3-D turbulent combustion studies.
- **Description**: Structured rectangular grid plus patch-based AMR (although NERSC-6 code does not adapt);
  - hydro model has implicit & explicit components;
- **Coding**: ~ 100,000 lines Fortran 90/77.
- **Parallelism**: 3-D processor non-overlapping decomposition, MPI.
  - Knapsack algorithm for load distribution; move boxes close in physical space to same/close processor.
    - More communication than necessary but has AMR communication characteristics.
- **NERSC-6 tests**: weak scaling on 512 and 2048 cores; 16 boxes (32^3 cells each) per processor.

MAESTRO Scaling



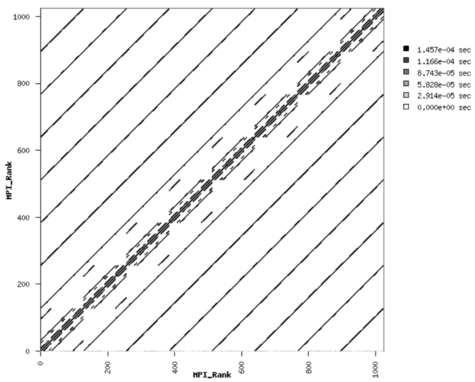
Data / Analysis by Katie Antypas, NERSC

Explicit parts of the code scale well but implicit parts of code pose more challenges due to global communications

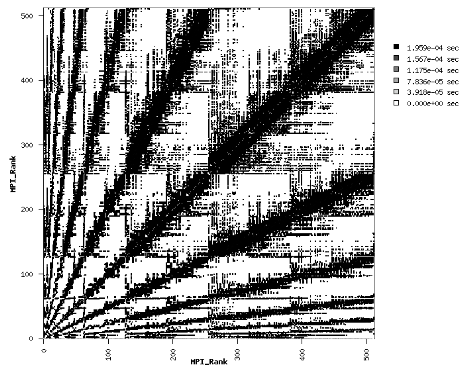
Key Tool

- NERSC's Integrated Performance Monitor (IPM)
- Portable, lightweight, and scalable tool for extracting MPI message-passing (and other) information.
- David Skinner, NERSC
- http://sourceforge.net/projects/ipm-hpc/

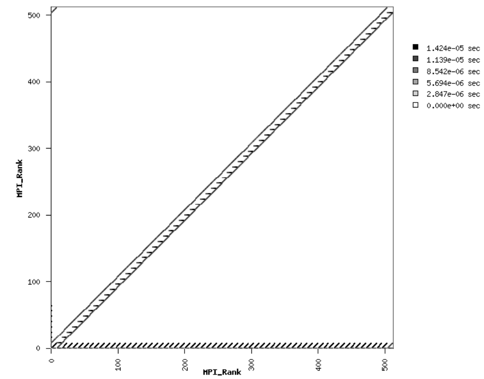
Benchmark Communication Topology from IPM



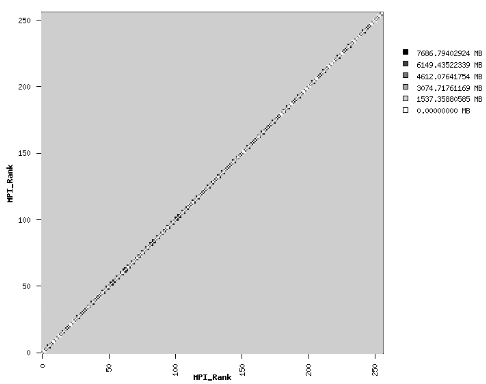
MILC



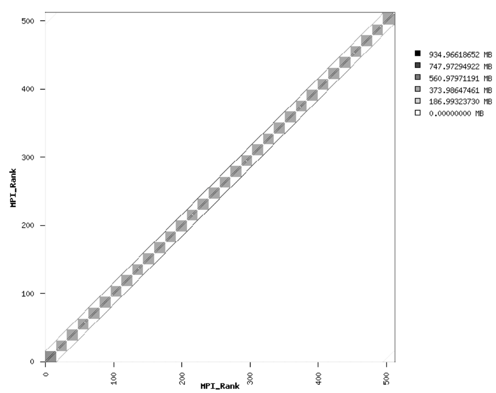
MAESTRO



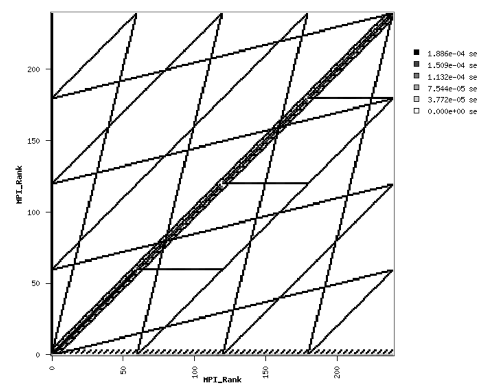
GTC



PARATEC



IMPACT-T



CAM

Other Application Areas

- **Fusion: 76 codes**
  - 5 codes account for >50% of workload: OSIRIS, GEM, NIMROD, M3D, GTC
  - Further subdivide to PIC (OSIRIS, GEM, GTC) and MHD (NIMROD, M3D) code categories
- **Chemistry: 56 codes for 48 allocations**
  - Planewave DFT: VASP, CPMD, DACAPO (already covered in MatSci)
  - Quantum Monte Carlo: ZORI
  - Ab-initio Quantum Chemistry: Molpro, Gaussian, GAMESS
- **Accelerator Modeling**
  - 50% of workload consumed by 3 codes VORPAL, OSIRIS, QuickPIC
  - Dominated by PIC codes,

code	MPP Award	Percent	Cumulative%
OSIRIS	2,112,500	11%	11%
GEM	2,058,333	11%	22%
NIMROD	2,229,167	12%	34%
M3D	1,921,667	10%	45%
GTC	1,783,333	10%	54%

Code	Award	Percent	Cumulative%
ZORI	695,000	12%	12%
MOLPRO	519,024	9%	21%
DACAPO	500,000	9%	29%
GAUSSIAN	408,701	7%	36%
CPMD	396,607	7%	43%
VASP	371,667	6%	49%
GAMESS	364,048	6%	56%

Code	MPP Award	Percent	Cumulative%
VORPAL	1,529,786	33%	33%
OSIRIS	784,286	16%	49%
QuickPIC	610,000	13%	62%
Omega3p	210,536	4%	66%
Track3p	210,536	4%	70%

Benchmark Selection Criteria

- **Coverage**
  - Cover science areas
  - Cover algorithm space
- **Portability**
  - Robust 'build' systems
  - Not an architecture specific implementation
- **Scalability**
  - Do not want to emphasize applications that do not justify scalable HPC resources
- **Open Distribution**
  - No proprietary or export-controlled code
- **Availability of Developer for Assistance/Support**

“Related Work”

- **L. Van Ertvelde, L. Eeckhout, “Dispersing Proprietary Applications as Benchmarks through Code Mutation,”**

  ASPLOS’08, March 1–5, 2008, Seattle, Washington

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° resolution); 240 timesteps	F90	netCDF
GAMESS	Quantum Chem (BES)	Dense linear algebra	384, 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	256,1024 Strong scaling	50 particles per cell	F90	
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, assemb.	
PARATEC	Material Science (BES)	DFT; FFT, BLAS3	256, 1024 Strong scaling	686 Atoms, 1372 bands, 20 iters	F90	Scalapack, FFTW

Algorithm Diversity

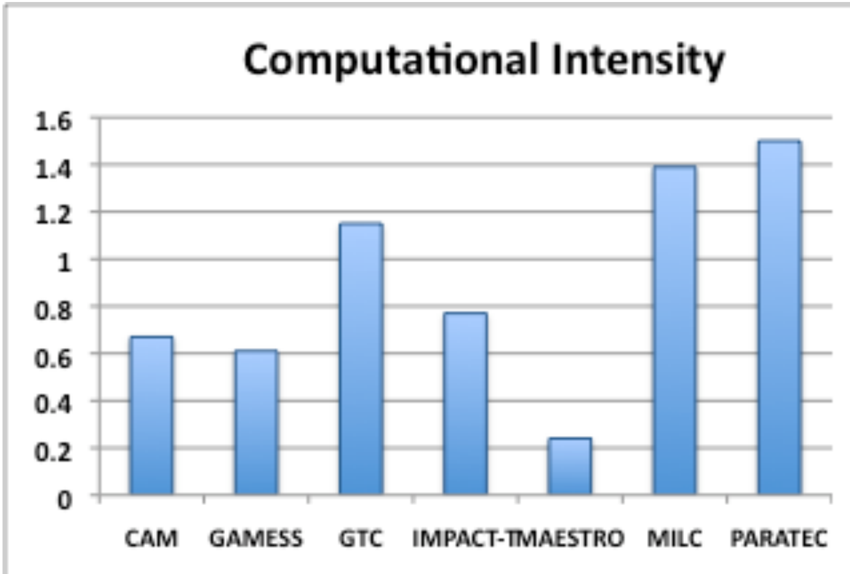
Science areas	*Dense linear algebra*	*Sparse linear algebra*	*Spectral Methods (FFT)s*	*Particle Methods*	*Structured Grids*	*Unstructured or AMR Grids*
Accelerator Science		X	X	X	X	X
Astrophysics	X	X	X	X	X	X
Chemistry	X	X	X	X		
Climate			X		X	X
Combustion					X	X
Fusion	X	X		X	X	X
Lattice Gauge		X	X	X	X	
Material Science	X		X	X	X	

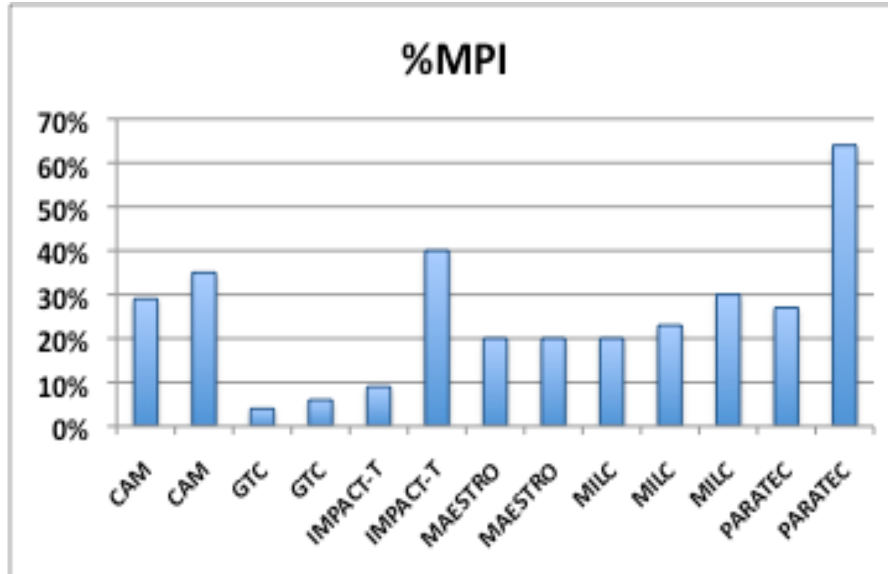
NERSC users require a system which performs adequately in all areas

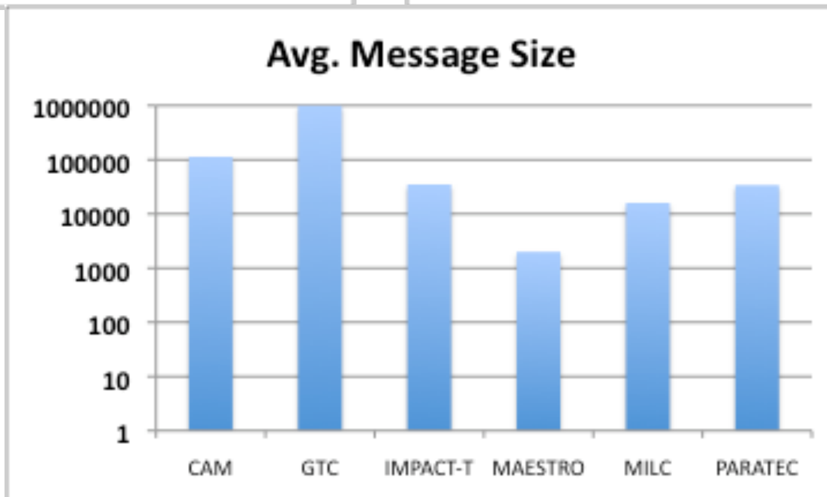
N6 Benchmarks Coverage

Science areas	*Dense linear algebra*	*Sparse linear algebra*	*Spectral Methods (FFT)s*	*Particle Methods*	*Structured Grids*	*Unstructured or AMR Grids*
Accelerator Science		X	X IMPACT-T	X IMPACT-T	X IMPACT-T	X
Astrophysics	X	X MAESTRO	X	X	X MAESTRO	X MAESTRO
Chemistry	X GAMESS	X	X	X		
Climate			X CAM		X CAM	X
Combustion					X CHOMBO	X MAESTRO
Fusion	X	X		X GTC	X GTC	X
Lattice Gauge		X MILC	X MILC	X MILC	X MILC	
Material Science	X PARATEC		X PARATEC	X	X PARATEC	

Characteristics Summary





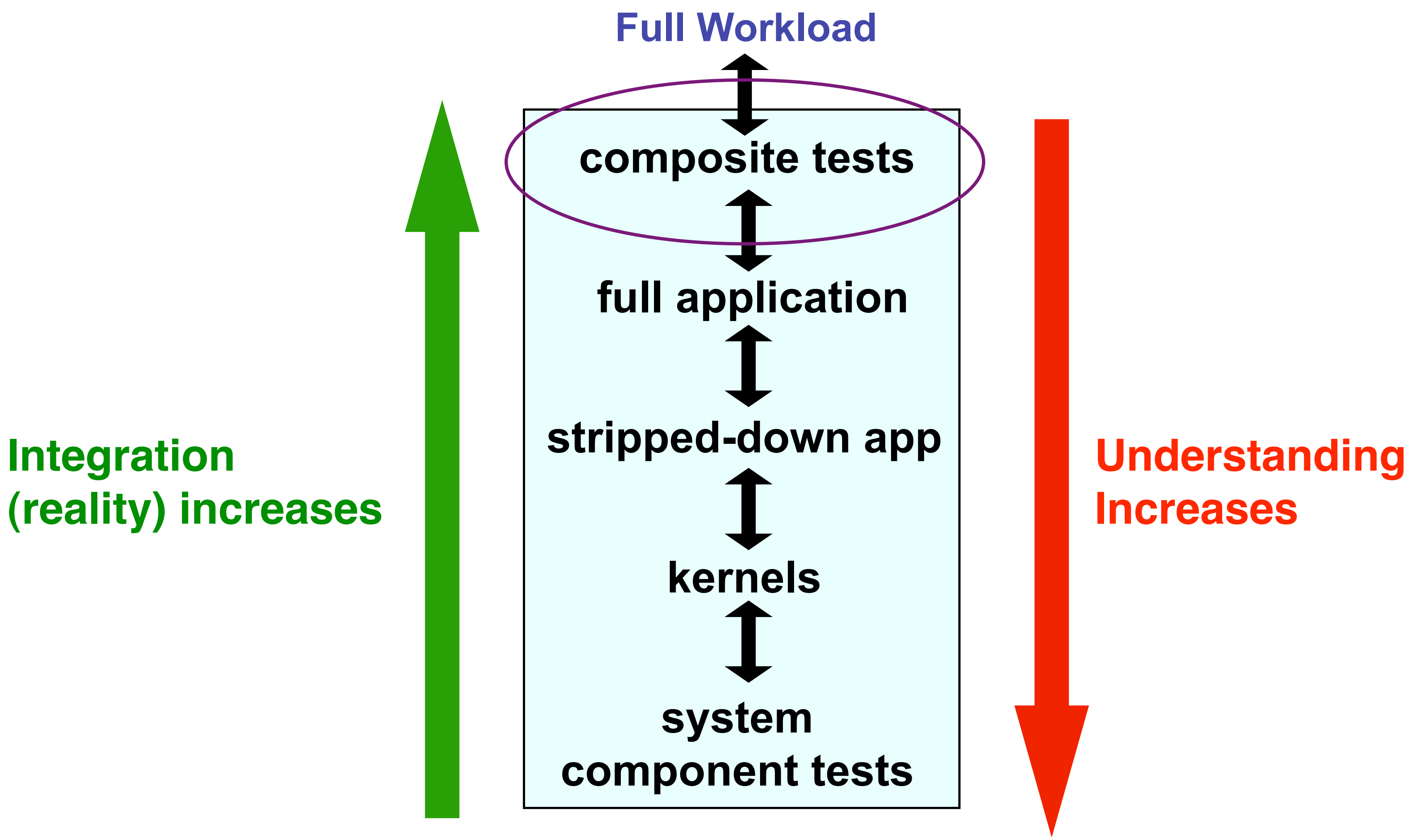


Summary So Far

- **Codes represent important science and/or algorithms <u>and</u> architectural stress points such as CI*, message type/size/topology.**
- **Codes provide a good means of system differentiation during acquisition and validation during acceptance.**
- **Strong suite of scalable benchmarks (256-8192+ cores).**

*CI = Computational Intensity, # FLOPs / Memory references

Use a Hierarchy of Tests



Sustained System Performance (SSP)

- Aggregate, un-weighted measure of <u>sustained</u> computational capability relevant to NERSC's workload.
- Geometric Mean of the processing rates of seven applications multiplied by *N*, # of cores in the system.
  - Largest test cases used.
- Uses floating-point operation count <u>pre</u>determined on a reference system by NERSC.

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

NERSC-6 Composite SSP Metric

The largest concurrency run of each full application benchmark is used to calculate the composite SSP metric

NERSC-6 SSP

CAM	GAMESS	GTC	IMPACT-T	MAESTRO	MILC	PARATEC
240p	1024p	2048p	1024p	2048p	8192p	1024p

For each benchmark measure

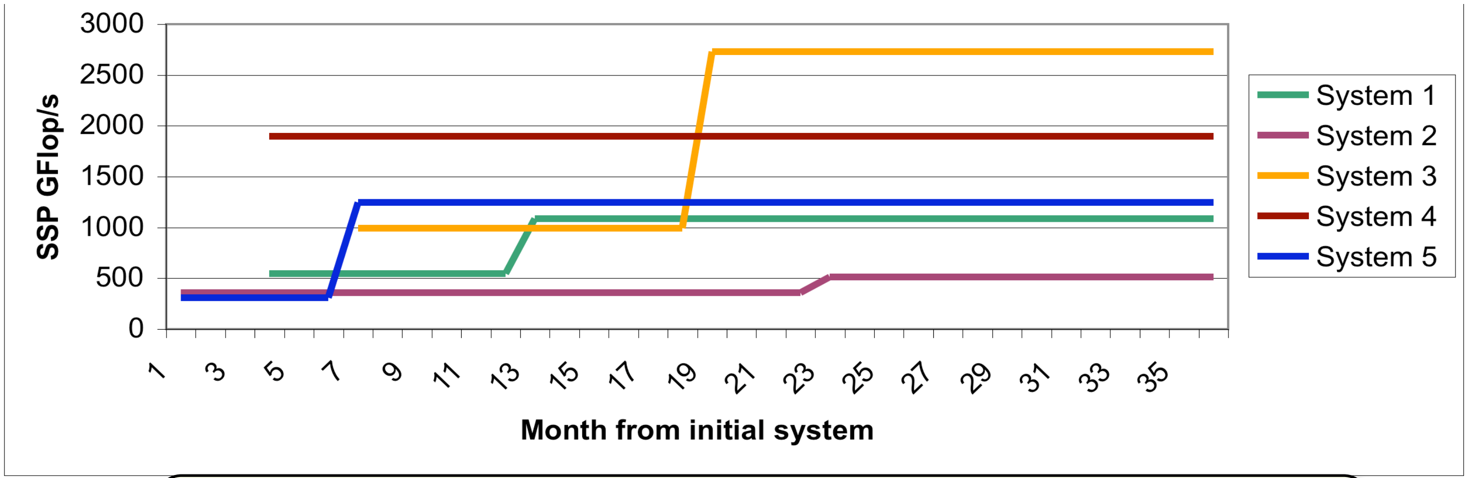
- *FLOP counts on a reference system*
- *Wall clock run time on various systems*

Key Point - Sustained System Performance (SSP) Over Time

- Integrate the SSP over a particular time period.
- SSP can change due to
  - System upgrades, Increasing # of cores, Software Improvements
- Allows evaluation of systems delivered in phases.
- Takes into account delivery date.
- Produces metrics such as SSP/Watt and SSP/$

$$Value_s = \frac{Potency_s}{Cost_s}$$

SSP Over 3 Year Period for 5 Hypothetical Systems



Area under SSP curve, when combined with cost, indicates system 'value'

SSP Example

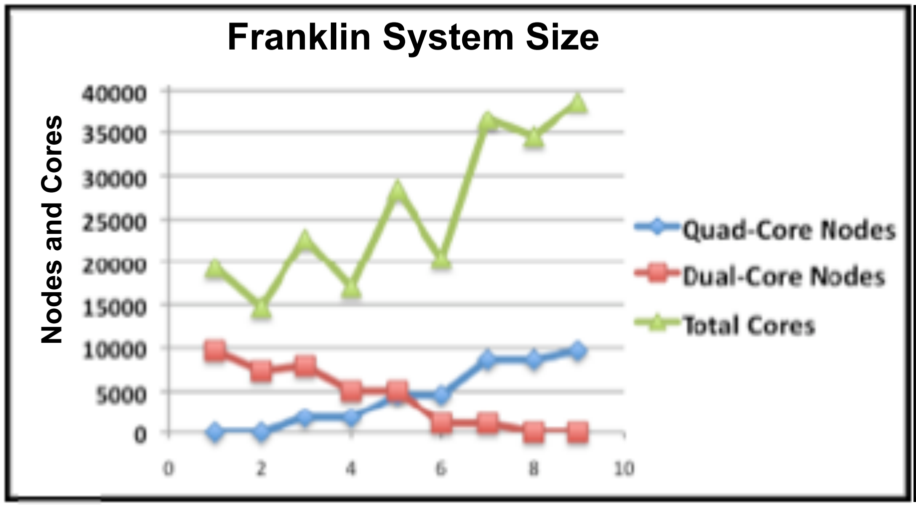
	Reference		Results	
Code	Tasks	GFLOP Count	Time	Rate per Core
cam	**240**	**57,669**	**408**	0.59
gamess	**1024**	**1,183,900**	**2478**	0.47
gtc	**2048**	**3,639,479**	**1493**	1.19
ImpactT	**1024**	**399,414**	**627**	0.62
maestro	**2048**	**1,122,394**	**2570**	0.21
milc	**8192**	**7,337,756**	**1269**	0.71
paratec	**1024**	**1,206,376**	**540**	2.18
SSP for 19,344 cores				**13.1**

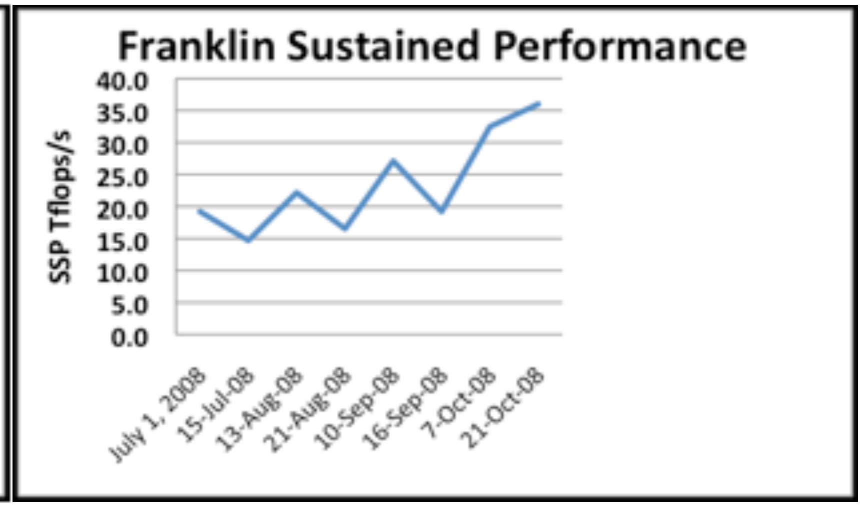
Rate Per Core = GFLOP count / (Tasks * Time)

Flop count measured on reference system

Measured wall clock time on system of interest

Maintaining Service While Improving Service

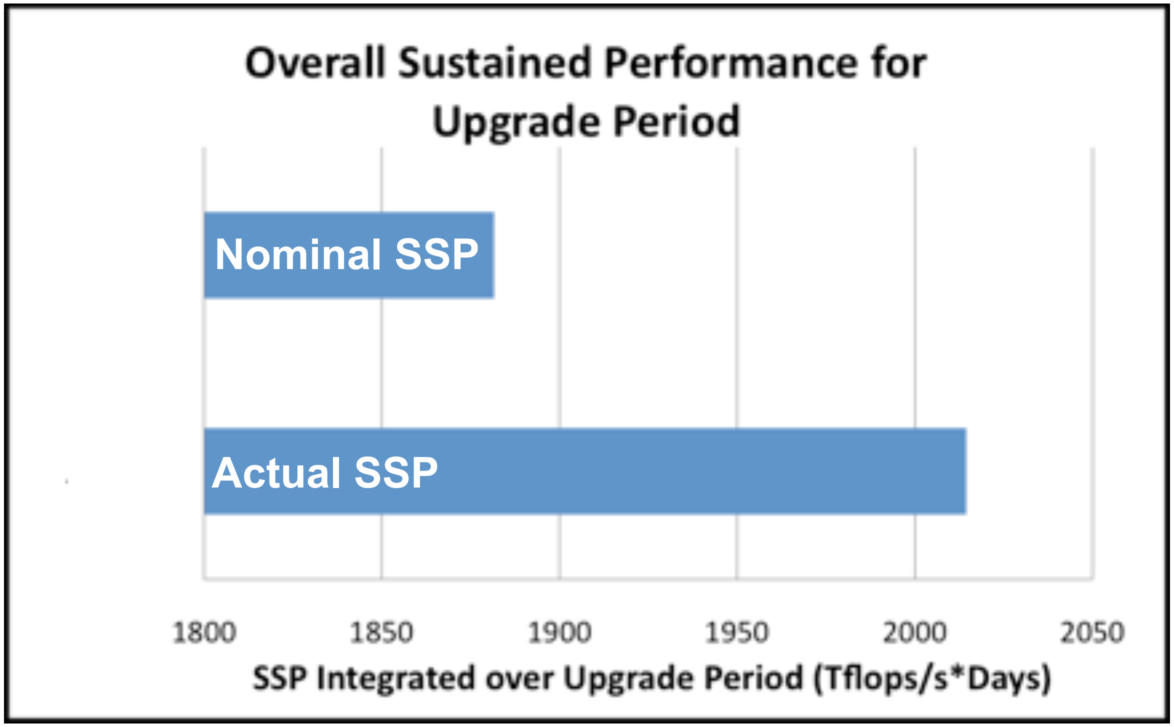




Phase	Start Date	Number of Dual Core Racks	Number of Quad Core Racks	Sustained Performance (SSP Tflops/s)	SSP Tflop/s-Days
Before	July 1, 2008	102	0	19.2	
1	15-Jul-08	78	0	14.7	425.8
2a	13-Aug-08	84	18	22.2	177.3
2b	21-Aug-08	54	18	16.5	330.4
3a	10-Sep-08	54	48	27.1	162.6
3b	16-Sep-08	12	48	19.2	403.2
4a	7-Oct-08	0	92	32.5	454.6
4b	21-Oct-08	0	102	36.0	

Key Phased Upgrade Benefit

- **Overall implementation provided 7% more science computing than waiting for all parts**



Some Common Science Trends

- **Increase support of engineering design studies**
  - **Eg., ITER and laser/plasma wakefield accelerators**
- **V&V increasingly important**
  - **Only scant experimental data available; often large uncertainties**
- **Hundreds of 2-D runs required to optimize beam properties for 3-D runs.**
  - **Parameter studies to reproduce experimental beam charge / energy**
- **Multiple length and time scales:**
  - **Requires resolution of the laser wavelength (microns, in 3-D) over the acceleration length (mm-cm, in 2-D), order 10^5 steps, 10^8 cells, and 10^9 particles**

Summary

- Workload-based evaluation.
- Workload characterization at different levels
- Main challenge: Living benchmarks, Good science
- Need to abstract the methods rather than the code.
- Appropriate aggregate metrics.
- Formal methodology for tests.
- Wide range of tests from all levels of the benchmark hierarchy.
- Metrics for system effectiveness.

Scientists Need More Than Flop/s

- Performance — How fast will a system process a code in isolation?
- Effectiveness — How fast will a system process an entire workload?
- Reliability — How often is the system available and operating correctly?
- Consistency — How often will the system process user work as fast as it can?
- Usability — How easy is it for users to get the system to go as fast as possible?

PERCU: NERSC's method for ensuring HPC system usability.

www.nersc.gov

THANK YOU.

“Backup” Slides

"Related Work"

- ***Workload Characterization Analysis (WCA):***
  - **Simple: list of programs known to be important, and a sample run-time for each.**
  - **Thorough:**
    - **distributions of program run-times,**
    - **frequencies of execution,**
    - **fraction of total time consumed,**
    - **plus historical trends used to estimate likely changes.**
- **Also *Workload Analysis with Weights (WAW)***

John R. Mashey, "*War of the Benchmark Means: Time for a Truce,*" ACM SIGARCH Computer Architecture News, Vol. 32, No. 4, September 2004

“Related Work”

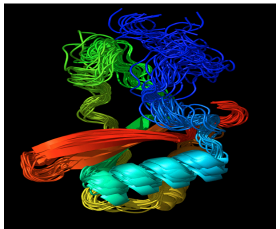
- ***Sample Estimation of Relative Performance Of Programs (SERPOP):***
  - **constructs a multi-element benchmark suite as a sample of some population of programs**
  - **Examples: LFK, NPB, SPEC**

John R. Mashey, “*War of the Benchmark Means: Time for a Truce,*” ACM SIGARCH Computer Architecture News, Vol. 32, No. 4, September 2004

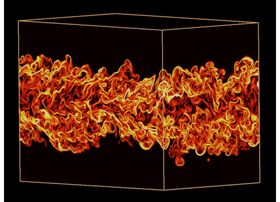
Chemistry Workload

- Some overlap with Material Science
- Multi-functional codes: GAMESS /Gaussian/NWChem
- Codes are proxies for exposing communication performance characteristics not visible from MPI
- Inflection point in terms of methods due to machine scale?

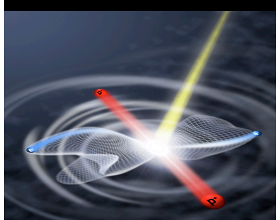
About the Cover



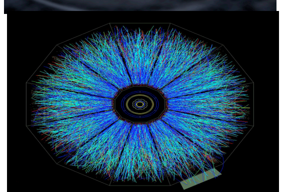
Schematic representation of 2^o secondary structure of native state simulation of the enzyme RuBisCO, the most abundant protein in leaves and possibly the most abundant protein on Earth. http://www.nersc.gov/news/annual_reports/annrep05/research-news/11-proteins.html



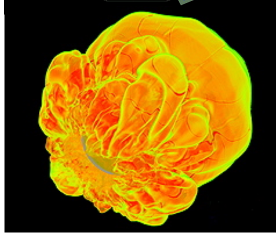
Direct Numerical Simulation of Turbulent Nonpremixed Combustion. Instantaneous isocontours of the total scalar dissipation rate field. (From E. R. Hawkes, R. Sankaran, J. C. Sutherland, and J. H. Chen, "Direct Numerical Simulation of Temporally-Evolving Plane Jet Flames with Detailed CO/H_2 Kinetics," submitted to the 31st International Symposium on Combustion, 2006.)



A hydrogen molecule hit by an energetic photon breaks apart. First-ever complete quantum mechanical solution of a system with four charged particles. W. Vanroose, F.Martín, T.N. Rescigno, and C. W. McCurdy, "Complete photo-induced breakup of the H_2 molecule as a probe of molecular electron correlation," Science **310, 1787 (2005)**



Display of a single Au + Au ion collision at an energy of 200 A-GeV, shown as an end view of the STAR detector. K. H. Ackermann et al., "Elliptic flow in Au + Au collisions at = 130 GeV," Phys. Rev. Lett. **86, 402 (2001).**



Gravitationally confined detonation mechanism from a Type 1a Supernovae Simulation by D. Lamb et al, U. Chicago, done at NERSC and LLNL