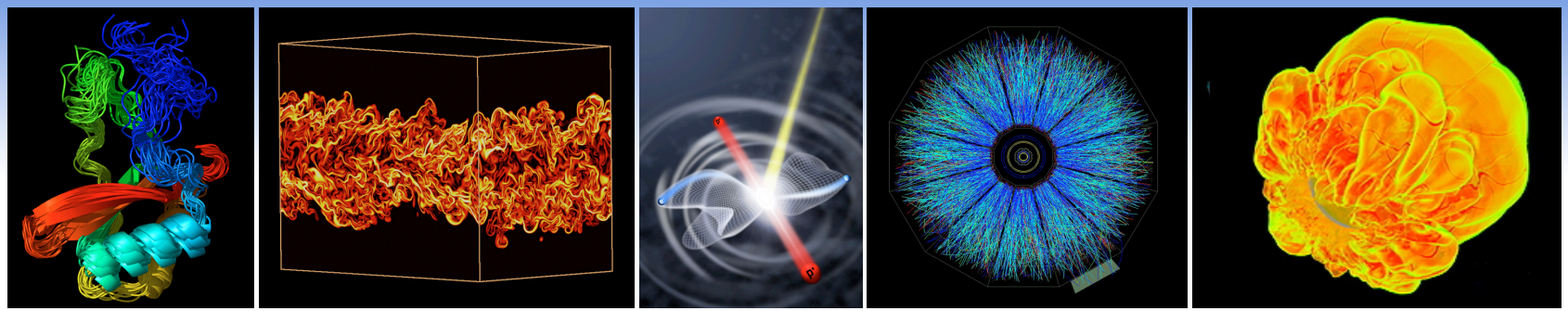


Breakthrough Science at NERSC

Harvey Wasserman

NERSC Science Driven System Architecture Group

Cray Technical Workshop, Europe

September 24-26, 2008

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC

Berkeley Lab

Outline

- Overview of NERSC
- Cray XT4 Usage
- Breakthrough Science
- Machine overload
- Current / Future plans

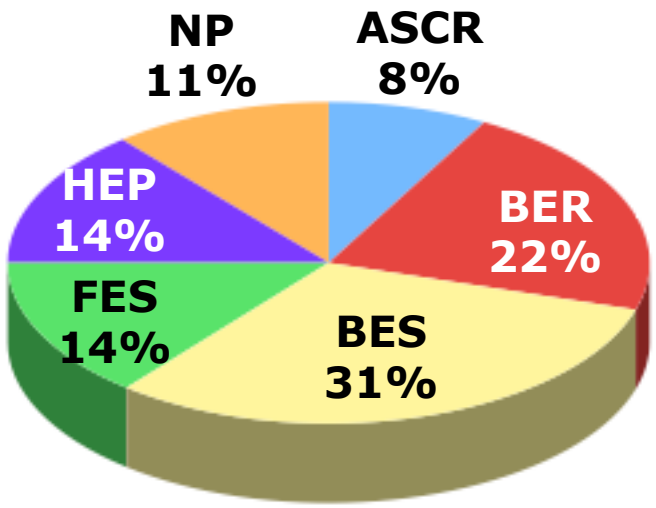
Intro to NERSC

- **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 facility for DOE SC.**

NERSC User Community

- **NERSC serves all areas**
  - **~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**
    - **Used throughout SC; not just DOE mission**
  - **70% Annual Production (ERCAP) awards:**
    - **From 10K hour (startup) to 5M hour**
    - **Via Call For Proposals; DOE chooses**
  - **10% NERSC and DOE/SC reserve, each**
- **Award mixture offers**
  - **High impact through large awards**
  - **Broad impact across science domains**

DOE View of Workload

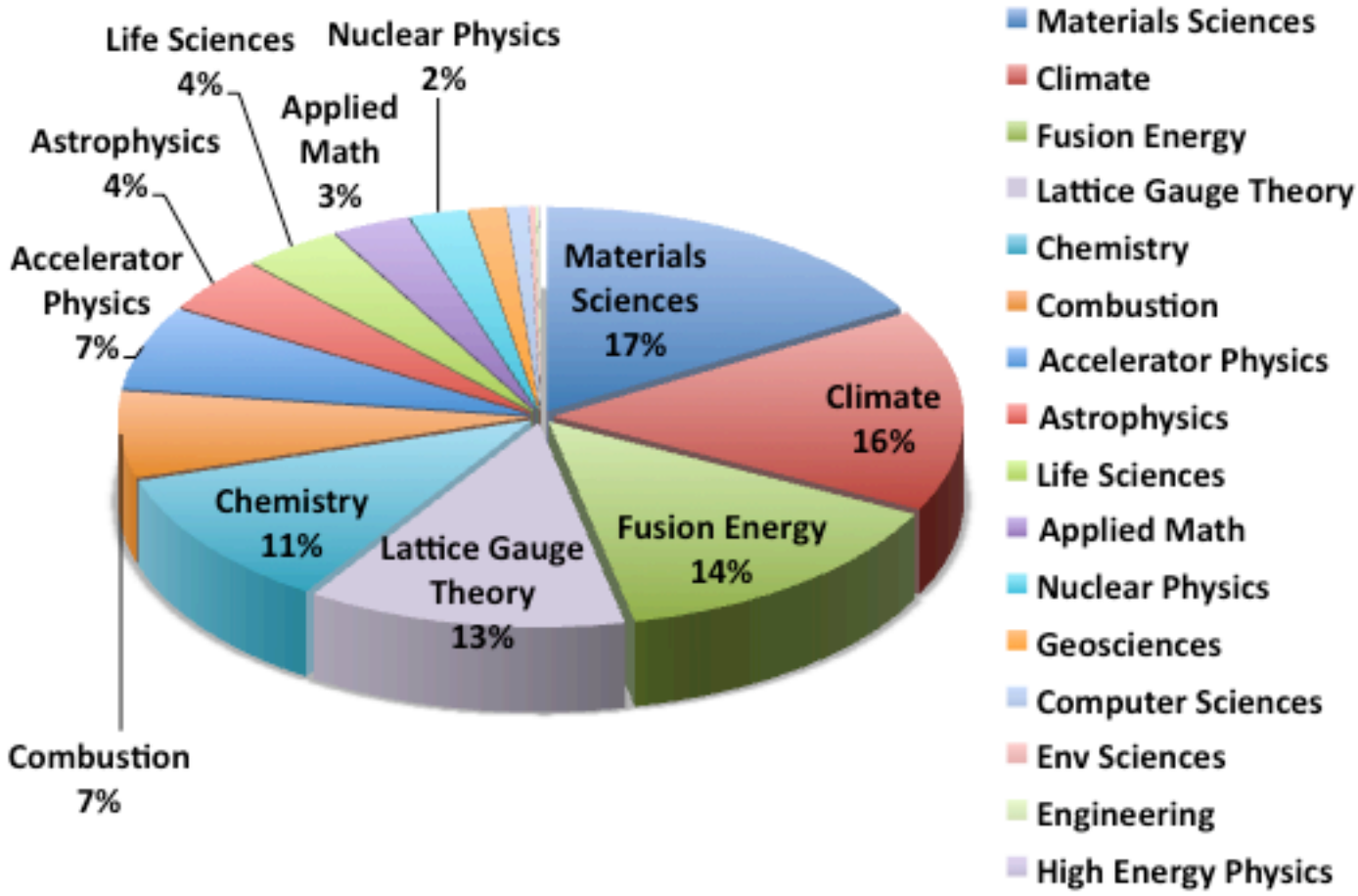


ASCR	Advanced Scientific Computing Research
BER	Biological & Environmental Research
BES	Basic Energy Sciences
FES	Fusion Energy Sciences
HEP	High Energy Physics
NP	Nuclear Physics

NERSC 2008 Allocations By DOE Office

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Science View of Workload



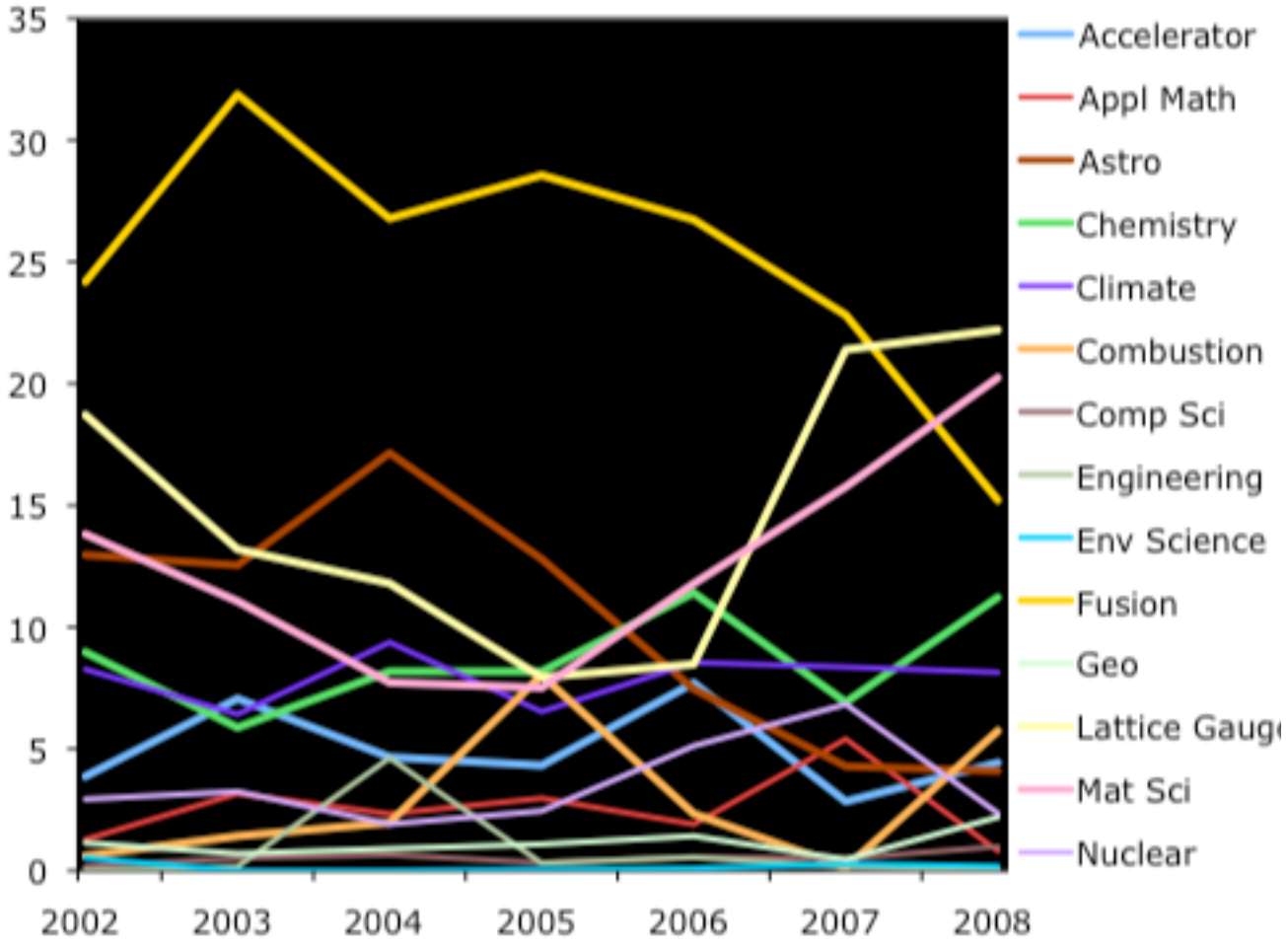
NERSC 2008 Allocations By Science Area

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Science Priorities are Variable

Usage by Science Area as a Percent of Total Usage



NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Franklin: NERSC's Cray XT4

NERSC's Cray XT4

- "Franklin" (NERSC-5)
  - 102 Cabinets in 17 rows
  - 9,660 nodes (19,320 cores)
  - 39.5 TBs Aggregate Memory (4 x 1GB DIMMs per node)
- Sustained performance: discussed later
- Interconnect: Cray SeaStar2, 3D Torus
  - >6 TB/s Bisection Bandwidth
  - >7 GB/s Link Bandwidth
- Shared Disk: 400+ TBs
- Network Connections
  - 24 x 10 Gbps + 16 x 1 Gbps
  - 60 x 4 Gbps Fibre Channel

Franklin Early User Program

- **Franklin in 2007**
  - Accepted in October, 2007
  - DOE allocations didn't start until January 2008
  - Before acceptance, full NERSC workload was running
  - All this usage was "free" (not charged against allocation)
- **Result:**
  - Franklin was 80%-95% utilized within a week of acceptance
  - Users consumed 5x more time than allocated in 2007 (14x for largest users)
  - Users produced important science results, experiment with new algorithms, and scaling to new levels (next slides…)

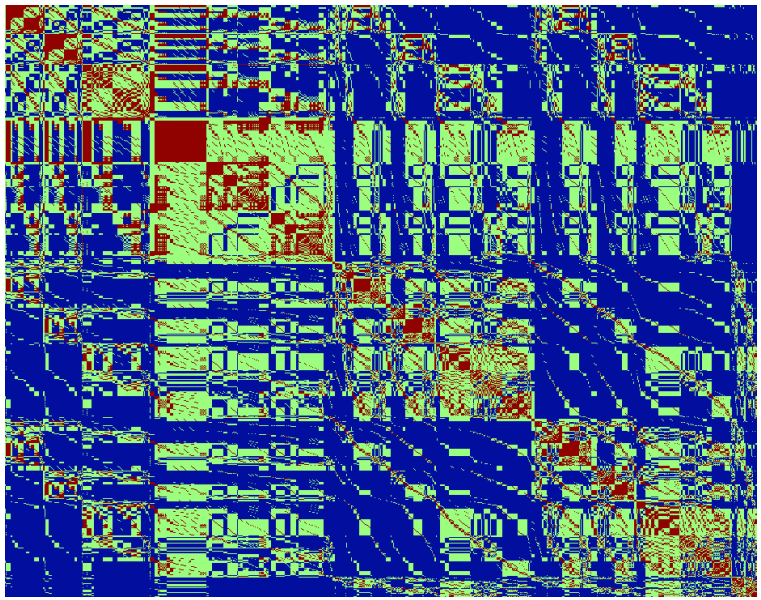
Six Breakthrough Science Stories

- Nuclear Physics
- Geochemistry
- Plasma Turbulence
- Combustion
- Nanoscience
- Climate

Nuclear Physics (1 of 2)

James Vary, P. Maris Iowa State

- High accuracy *ab initio* calculations on O^{16} using no-core shell model and no-core full configuration interaction model
- SciDAC Project: Universal Nuclear Energy Density Functional
- "Many Fermion Dynamics — nuclear" code (MFDn) evaluates many-body Hamiltonian to obtain low-lying eigenvalues and eigenvectors using the Lanczos algorithm
- I/O-dominated at high core counts.

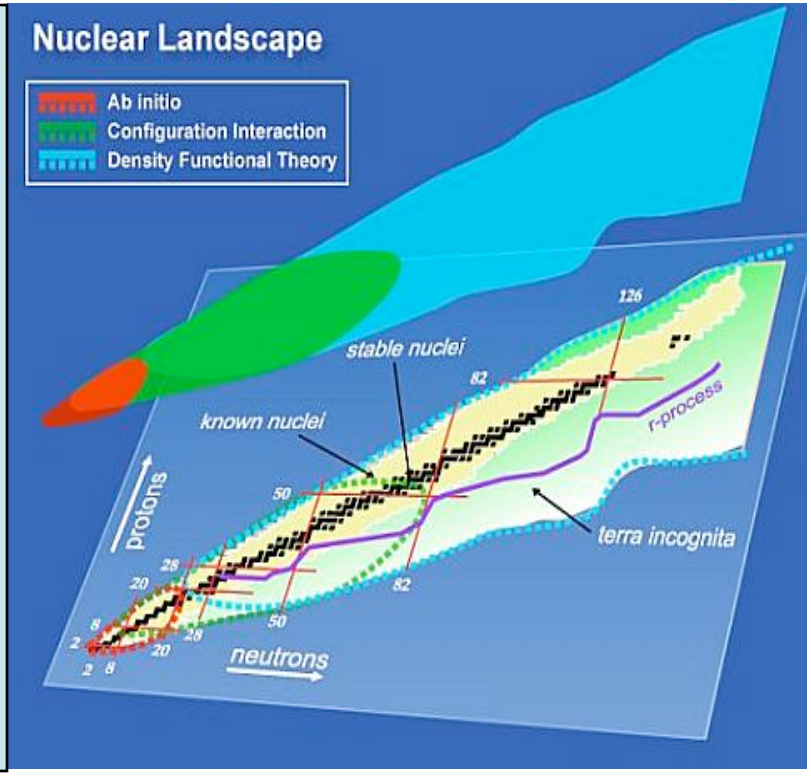


nonzeros
potentially nonzero blocks
zero blocks

Nuclear Physics (2 of 2)

- SC08 paper: "Accelerating Configuration Interaction Calculation for Nuclear Structure" Tuesday, 11/18

- **Science Results:**
  - First of a kind, most accurate calculations for this size nucleus
  - Can be used to parametrize new density functionals for nuclear structure simulations
- **Scaling Results:**
  - 4M hours used; 200K allocated
  - 12K cores; vs 2-4K before Franklin uncharged time
  - Diagonalize matrices of dimension up to 1 billion in 4.5 hrs.



Scaling Science

Inspired by P. Kent, "Computational Challenges in Nanoscience: an *ab initio* Perspective", Peta08 workshop, Hawaii (2008) and Jonathan Carter (NERSC).

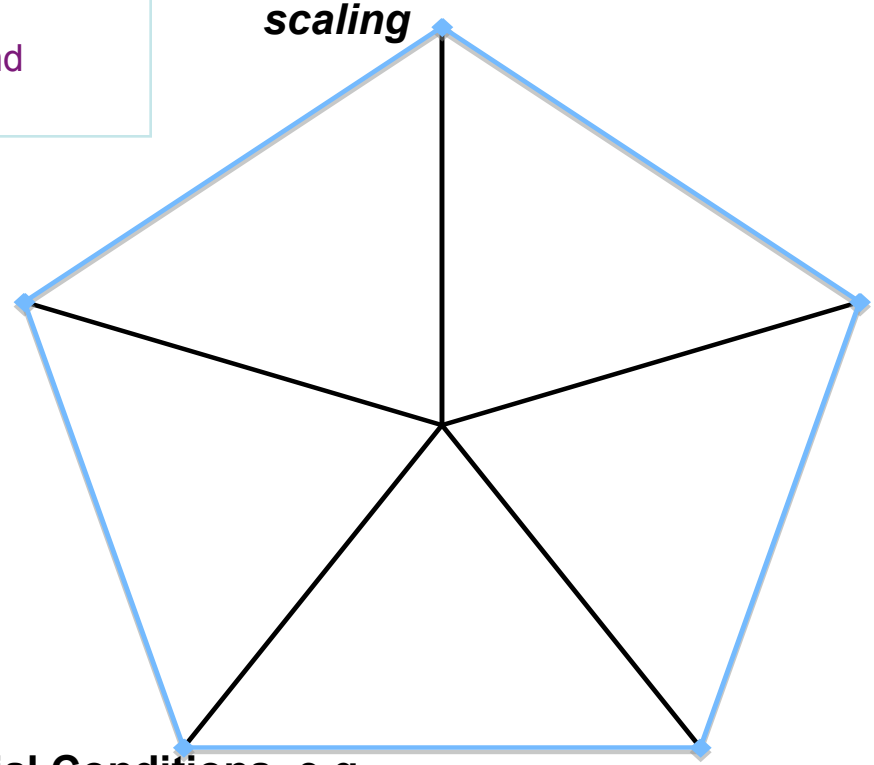
Length, Spatial extent, #Atoms, *Weak scaling*

Time scale Optimizations, *Strong scaling*

Simulation method, e.g. DFT, QMC or CC, LES or DNS

Initial Conditions, e.g. molecule, boundaries, *Ensembles*

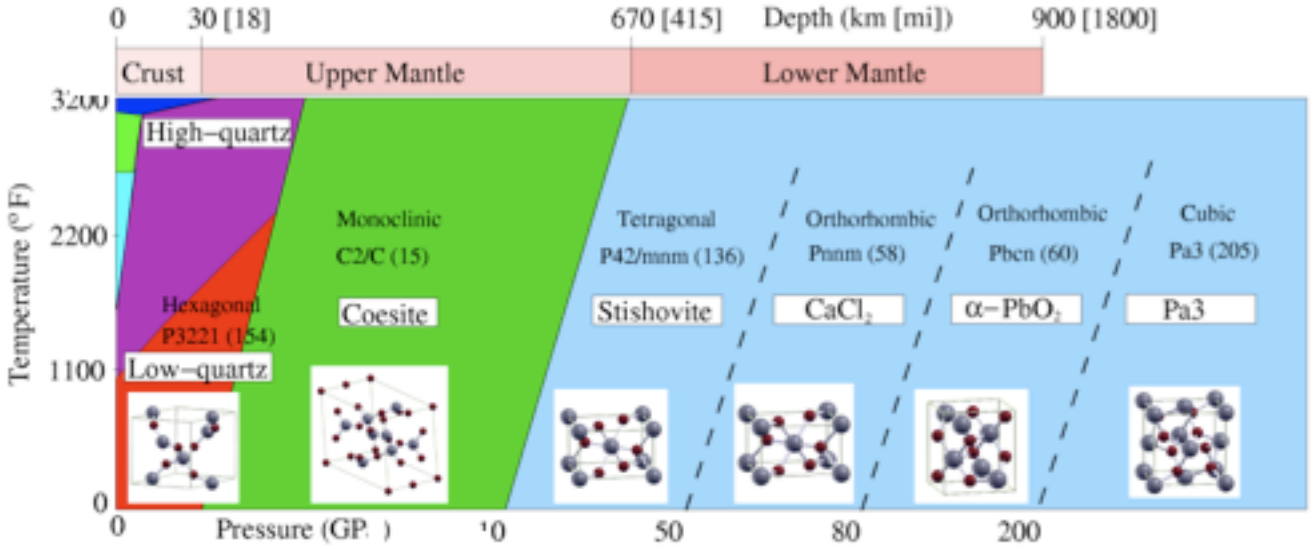
Convergence, systematic errors due to cutoffs, etc.



Quantum Monte Carlo Geophysics

Kevin Driver, John Wilkins (Ohio State)

- **Demonstrated success using CASINO on Franklin: first QMC elastic constants for a solid**
  - **~3 million CPU hours, 4000-8000-core jobs**
  - **--1 meV error bars on predicted QMC energies**
- **QMC maps Earth's mantle phase diagram!**
  - **Jump in seismic wave velocity due to structural changes in silicates under pressure.**
  - **QMC agrees with DFT for bulk modulus but not for transition pressure.**



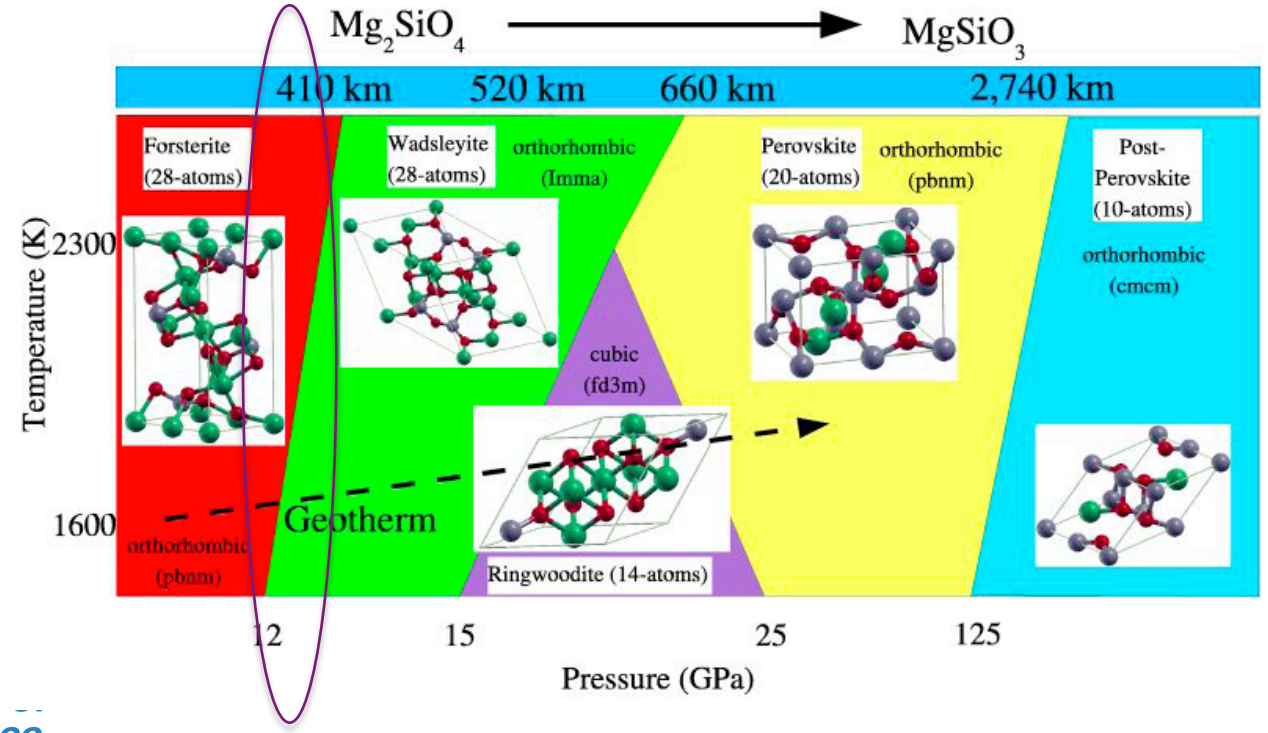
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

QMC Geophysics (2 of 2)

Kevin Driver, John Wilkins (Ohio State)

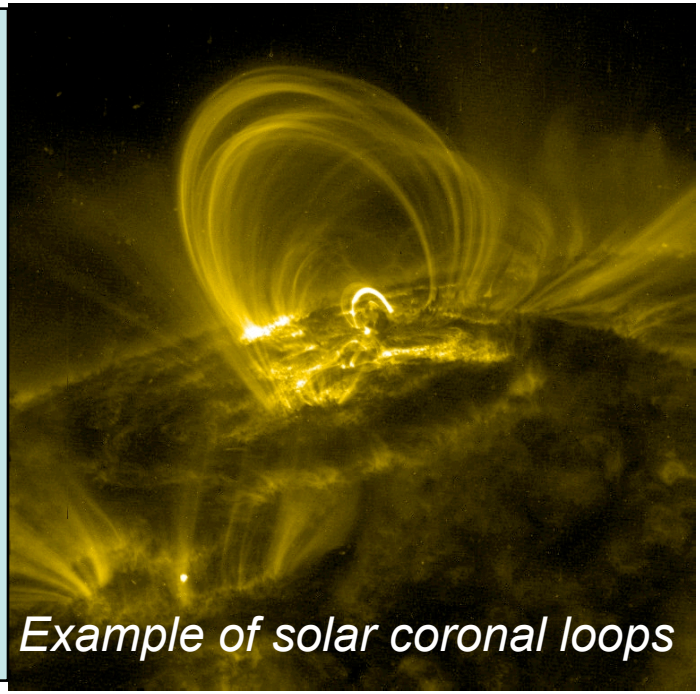
- **Current work: Use CASINO on QC Franklin to calculate pressure of Fosterite-Wadsleyite p-transition; DFT fails.**
- **QMC well suited for multi-core extreme parallelism; NERSC allows method to be applied to realistic materials.**



Kinetic Plasma Turbulence (1 of 2)

- **AstroGK, new gyrokinetic code for astrophysical plasmas**
- **Different from fusion plasmas because of vastly disparate scales; typically caused by violent events rather than gradients.**
- **PIs: William Dorland (U. of Maryland), Gregory Howes, T. Tatsuno**

- **Possible applications include**
  - **Solar wind,**
  - **Interstellar scintillation due to e^- density fluctuations in Milky Way's interstellar medium (ISM),**
  - **Magnetorotational instability (MRI) of black holes.**
- **Combination of spectral/finite -difference methods**



Example of solar coronal loops

http://astro.berkeley.edu/~ghowes/astrogk/

Kinetic Plasma Turbulence (2 of 2)

- (nx, ny, nz, nξ, nE, ns) = (32, 32, 64, 128, 32, 2)
  = 536,870,912 points; 19,118 cpu-hours
- Science Results
  - Shows how magnetic turbulence leads to particle heating
- Scaling Results
  - Runs on 16K cores
- Franklin early user program produced this publication; INCITE grant in 2008.

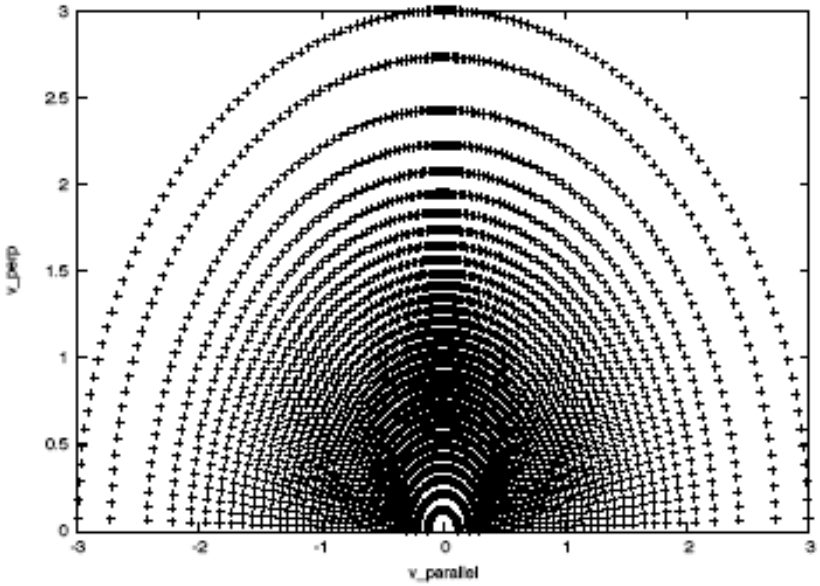


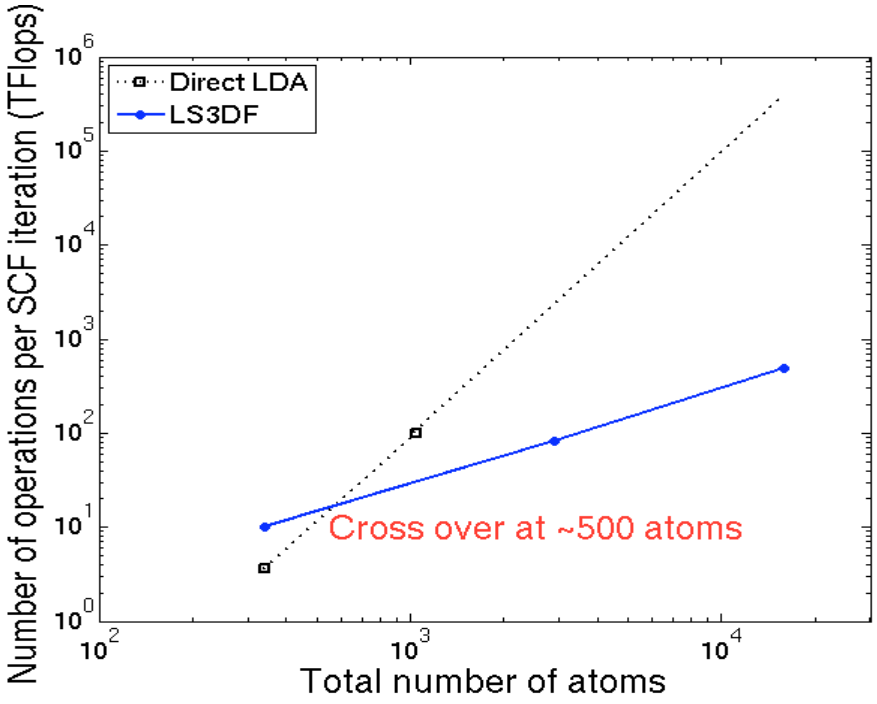
FIG. 1: Plot of grid used in velocity space with 128 pitch angles and 32 energies. Grid point locations are chosen spectrally in a Legendre polynomial basis.

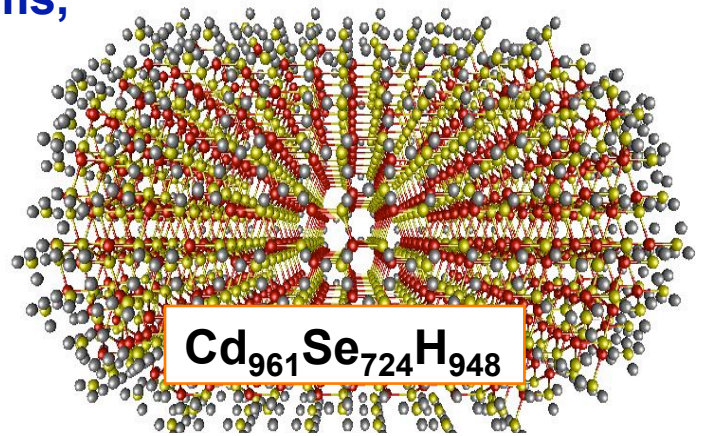
G Howes, Phys. of Plasmas, DPP Mtng. Spec. Iss.
http://arxiv.org/abs/0711.4358v1

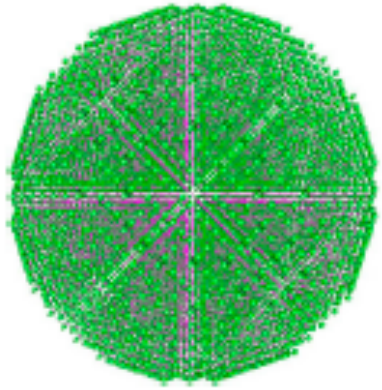
Scalable Nanoscience (1 of 2)

Lin-Wang Wang, LBNL

- **Linear Scaling 3D Fragment (LS3DF).**
  - **Novel divide & conquer approach to solve DFT but scales with O(n), number of atoms, rather than O(n^3).**



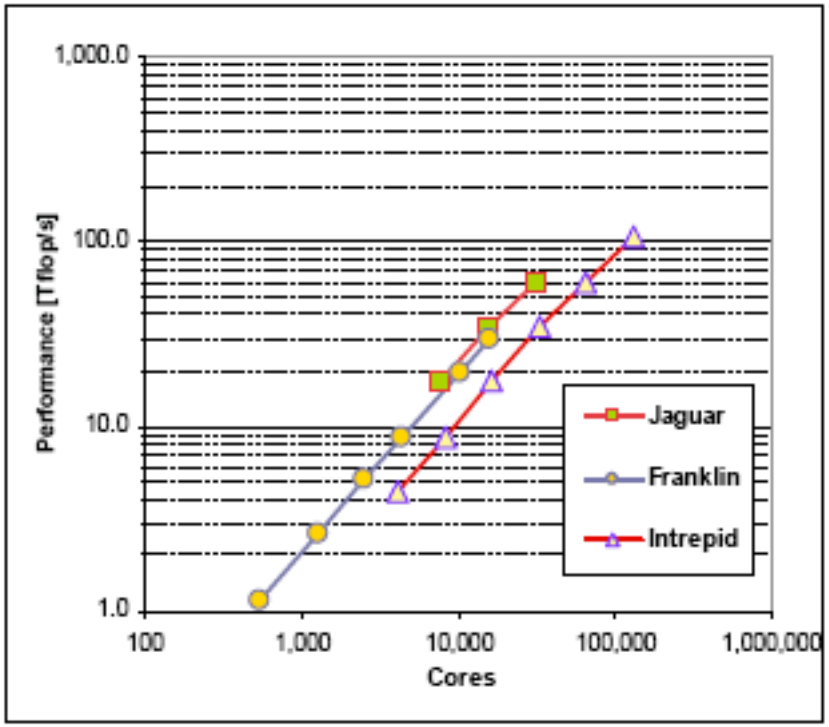




$Si_{13607}H_{2236}$

Scalable Nanoscience (2 of 2)

- **Gordon Bell Prize finalist, SC08** ***Thursday Nov, 20***



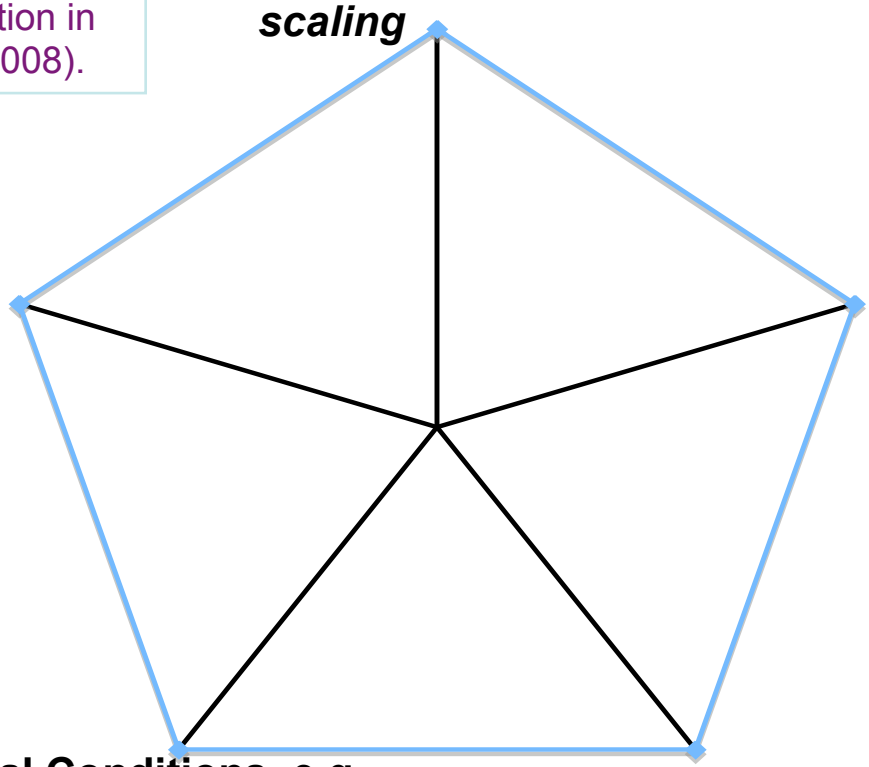
- **Science Results**
  - Calculated dipole moment on 2633 atom CdSe quantum rod, $Cd_{961}Se_{724}H_{948}$.
- **Scaling Results**
  - Ran on 2560 cores
  - Took 30 hours *vs.* many months for $O(n^3)$ algorithm
  - Good parallel efficiency (80% on 1024 relative to 64 procs)

Lin-Wang Wang, B. Lee, H. Shan, Z. Zhao, J. Meza, E. Strohmaier, D. Bailey, "Linear Scaling Divide-and-conquer Electronic Structure Calculations for Thousand Atom Nanostructures," SC08, to appear.

Scaling Science

Inspired by P. Kent, "Computational Challenges in Nanoscience: an *ab initio* Perspective", presentation in Peta08 workshop, Hawaii (2008).

Length, Spatial extent, #Atoms, *Weak scaling*



Time scale Optimizations, *Strong scaling*

Simulation method, e.g. DFT, QMC or CC, LES or DNS

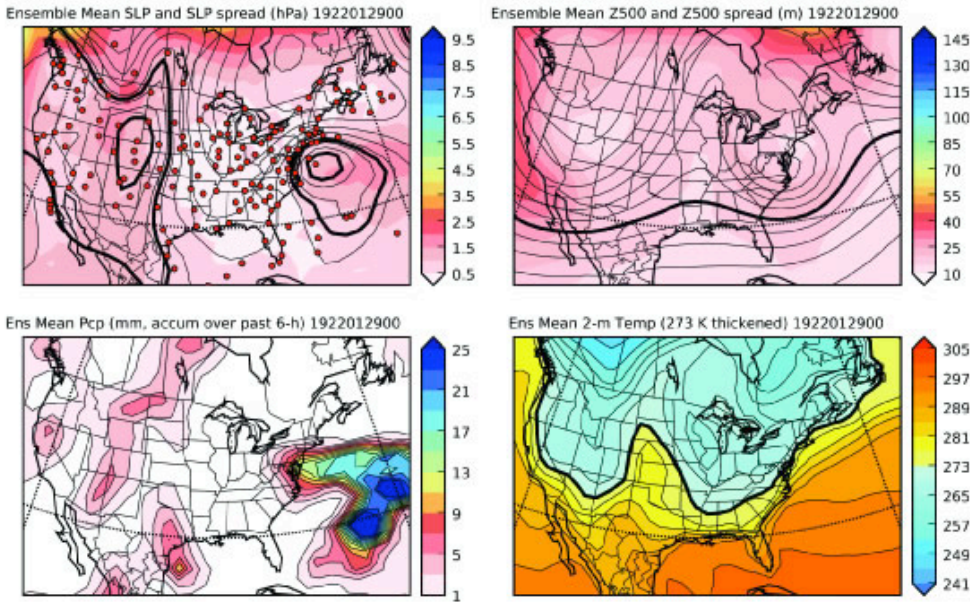
Initial Conditions, e.g. molecule, boundaries, *Ensembles*

Convergence, systematic errors due to cutoffs, etc.

Validating Climate Models

- **INCITE Award, "20th Century Reanalysis" PI: G. Compo, U. Colorado**
- **Generated 6-hourly global weather maps spanning 1918 to 1949**

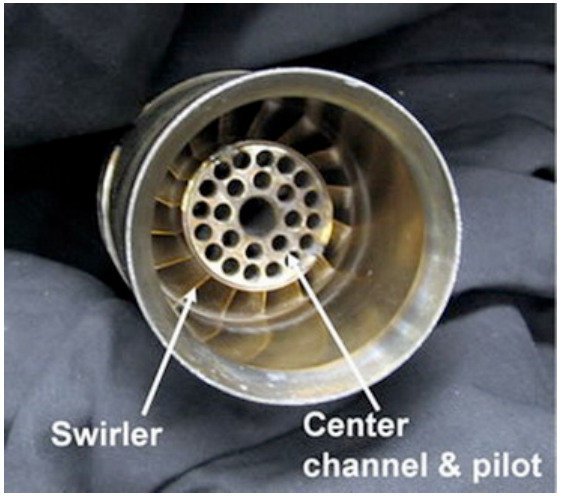
- **Science Results:**
  - **Reproduced 1922 Knickerbocker storm**
  - **Data can be used to validate climate and weather models**
- **NERSC Results:**
  - **3.1M CPU Hours in allocation**
  - **Scales to 2.4K cores**
  - **Switched to higher resolution algorithm with Franklin access**

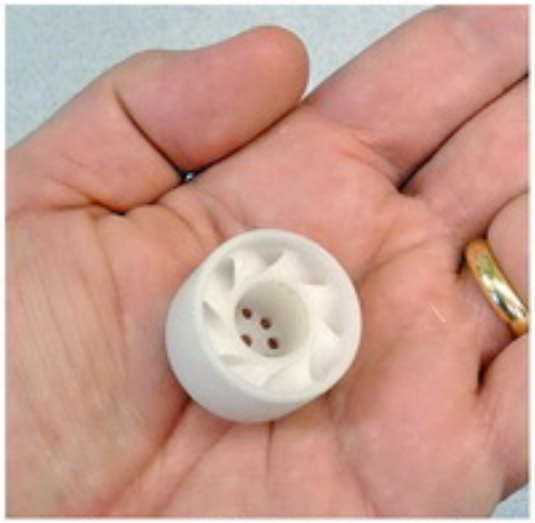


Sea level pressure data dating back to 1892 with color showing uncertainty (a&b); precipitation (c); temperature (d). Dots indicate measurements locations (a).

Low-Swirl Burner Simulation (1 of 2)

- **Discovered in 1991 at LBNL.**
- **Now being developed for fuel-flexible, near-zero-emission gas turbines (2007 R&D 100 Award)**





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

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

http://eetd.lbl.gov/aet/combustion/LSC-info/

Low-Swirl Burner Simulation (2 of 2)

PI: John Bell, LBNL

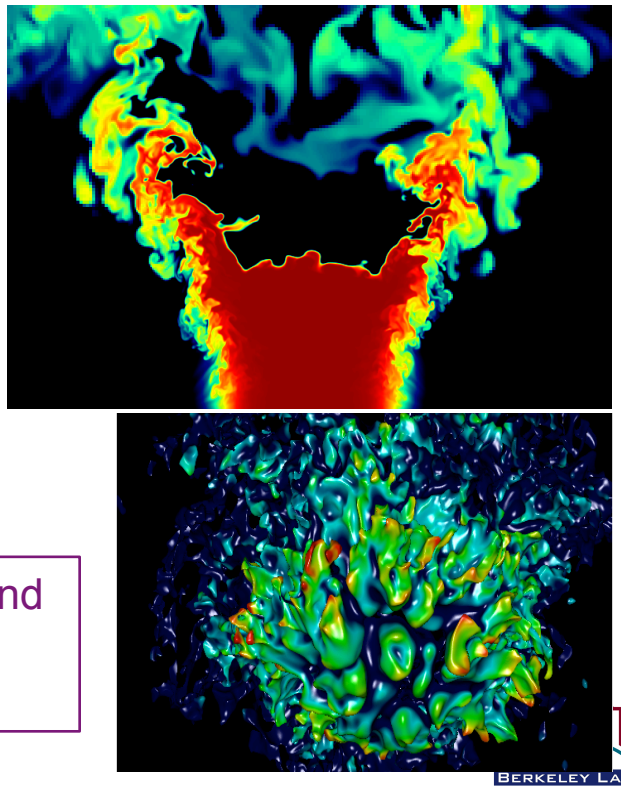
- Numerical simulation of an ultra-lean premixed hydrogen flame in a laboratory-scale low-swirl burner.
- Interaction of turbulence and chemistry.
- Method captures the hydrogen flame cell structures (lower right).

Science Result:

- **Adaptive low Mach number algorithm for reacting flow** instead of the traditional compressible equations with explicit DNS.

NERSC Results:

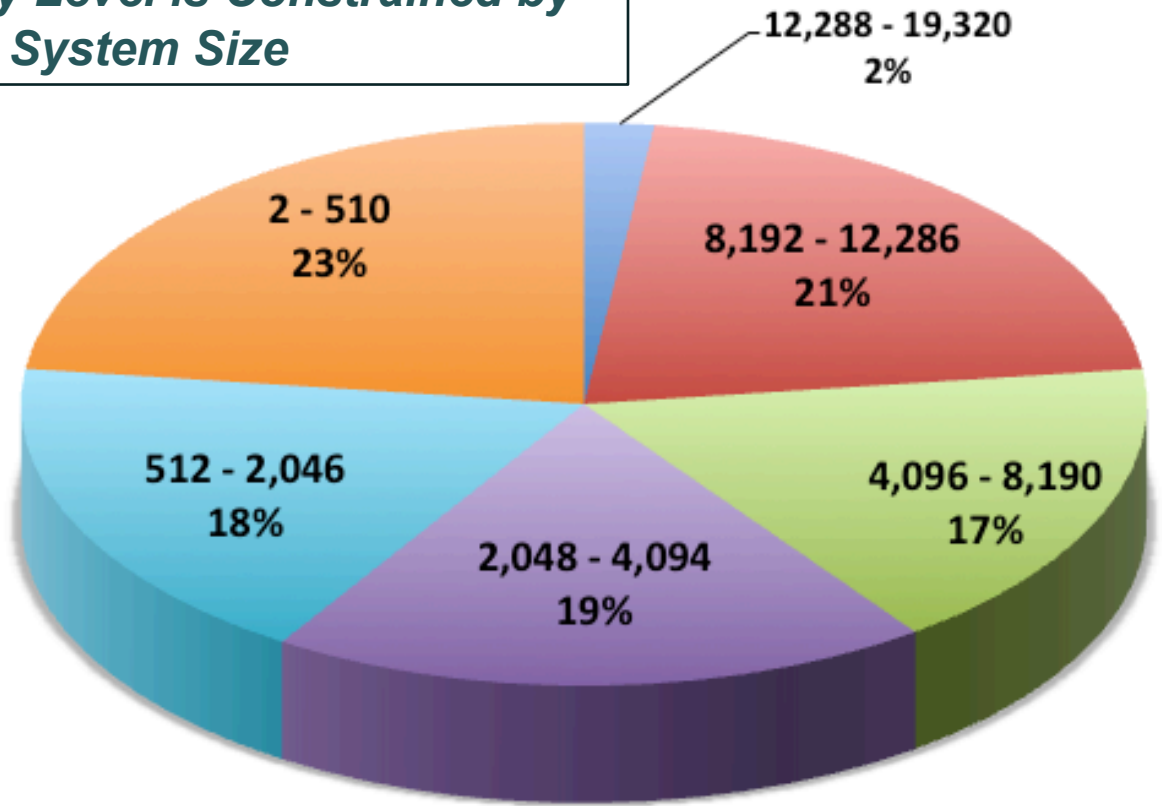
- **AMR saves memory and time.**
- **Scales to 6K cores, typically run at 2K**
- **Used 2.2M early science hours on Franklin**



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

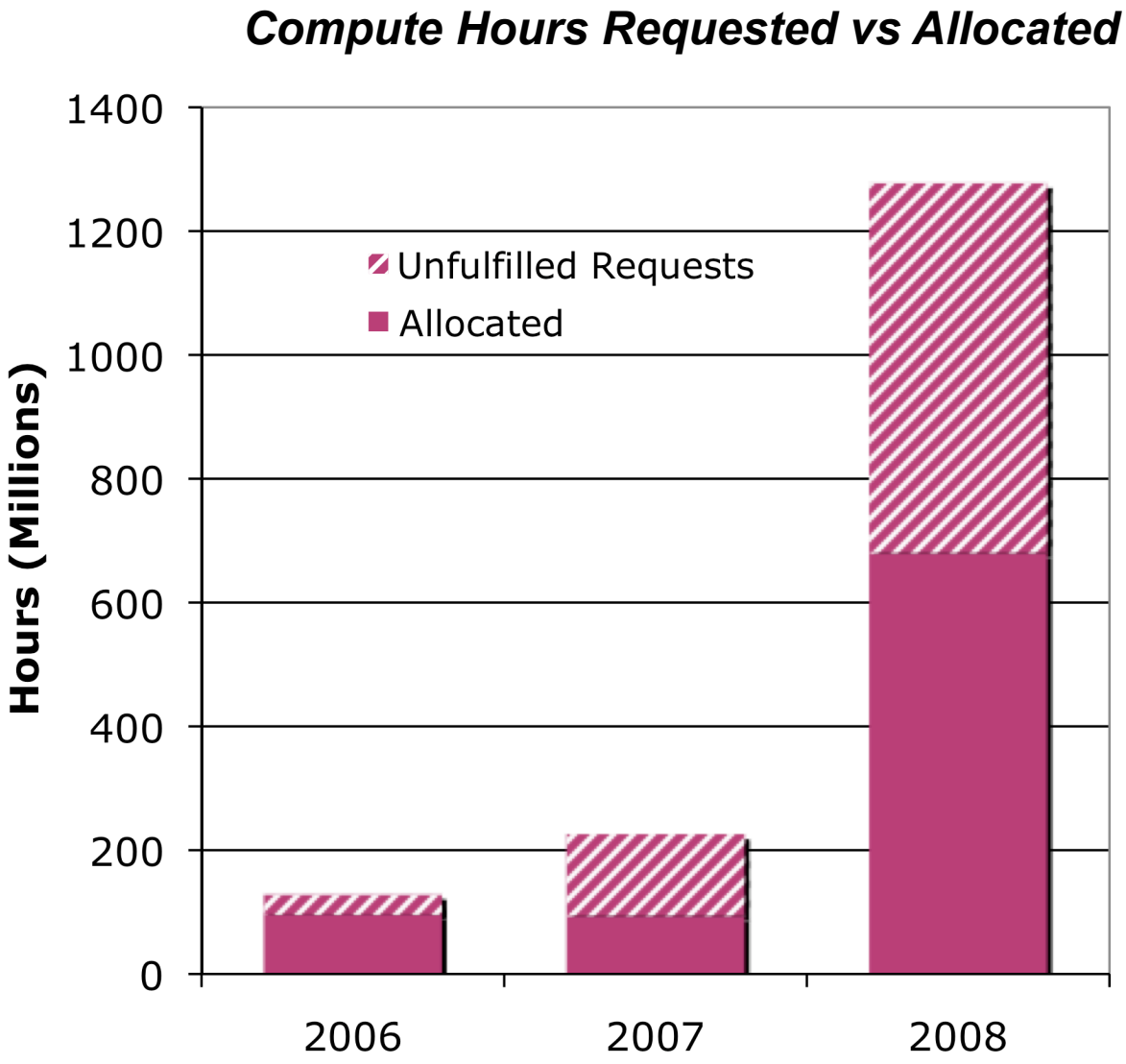
Parallelism on Franklin

Concurrency Level is Constrained by System Size



Raw Hours used on Franklin FY08 Q1-Q3 by # of cores (Raw Hours = wallclock hours * nodes * 2 CPUs/node)

Demand for More Computing

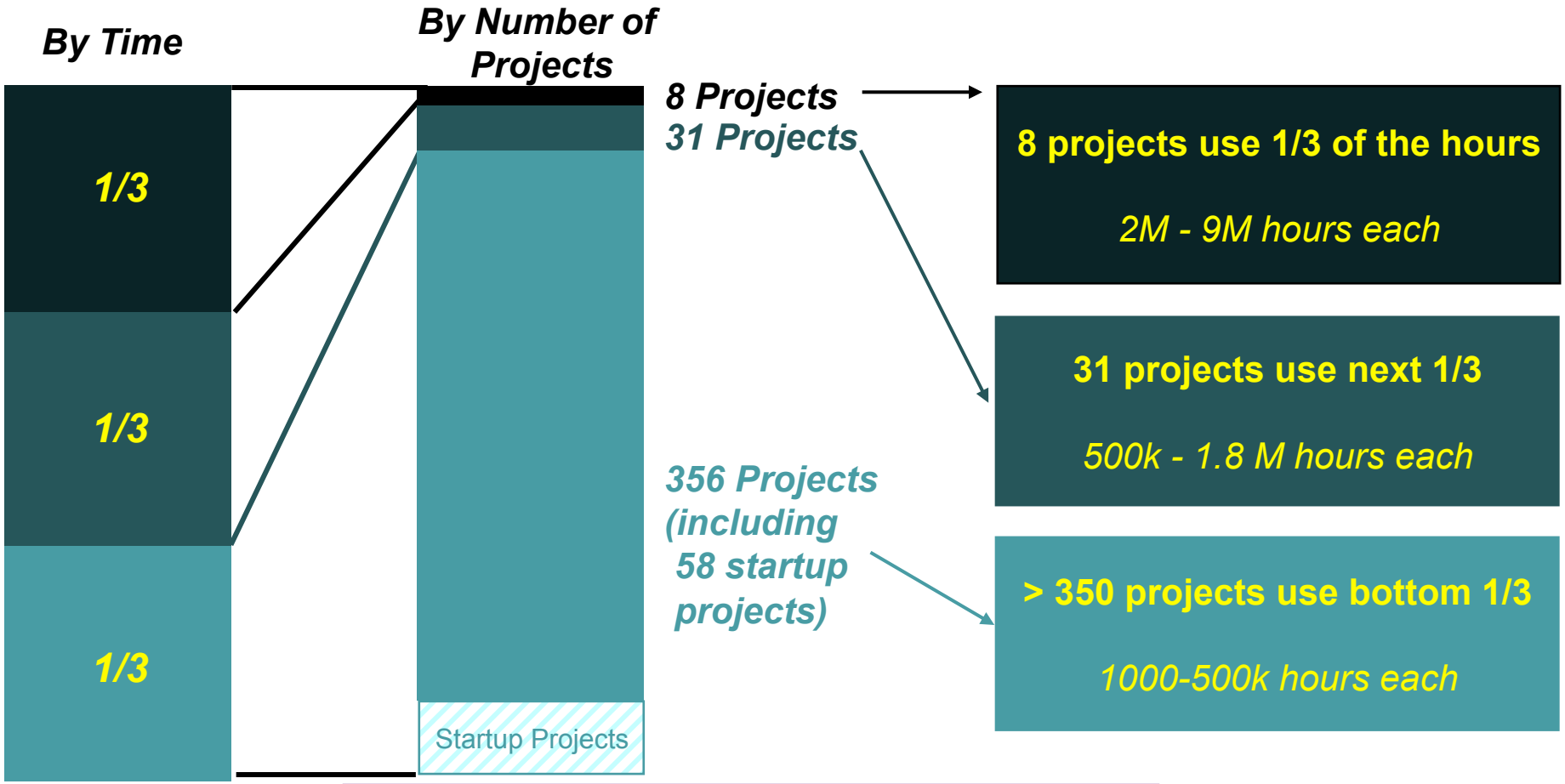


- *Each year DOE users request 2x as many hours as can be allocated*
- *This 2x is artificially constrained by perceived availability*
- *Backlog of meritorious requests amount to hundreds of millions of compute hours in 2008*

NERSC Allocation Breakdown

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Total NERSC allocation time divided into thirds



Middle projects have insufficient resources for large production runs

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NERSC Response

- **Further progress in these key science missions (and others) requires increased computational capability.**
- **NERSC Strategy: Increase user scientific productivity via timely introduction of the best new technologies designed to benefit the broadest subset of the NERSC workload**

  => Upgrade Franklin
  => Commence NERSC-6.

Franklin Quad Core Upgrade

- In-place, no-interruption upgrade taking place between July and October, 2008.
- All 9,672 nodes change from 2.6-GHz AMD64 to 2.3-GHz Barcelona-64.
- QC nodes have 8 GB memory, same average GB/core as on DC Franklin.
- Memory from 667 MHz to 800 MHz.

Franklin QC Upgrade

- **Parts changed:**
  - 9,680 Opterons
  - 38,720 DIMMs
- **Additional goals**
  - Minimize HSN & I/O degradation
  - Sufficient burn-in time to minimize production problems
  - Fallback / backout plan if needed

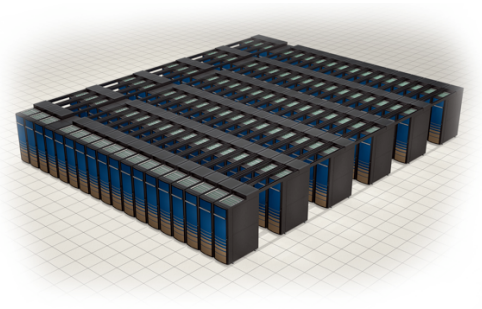
Franklin QC Upgrade

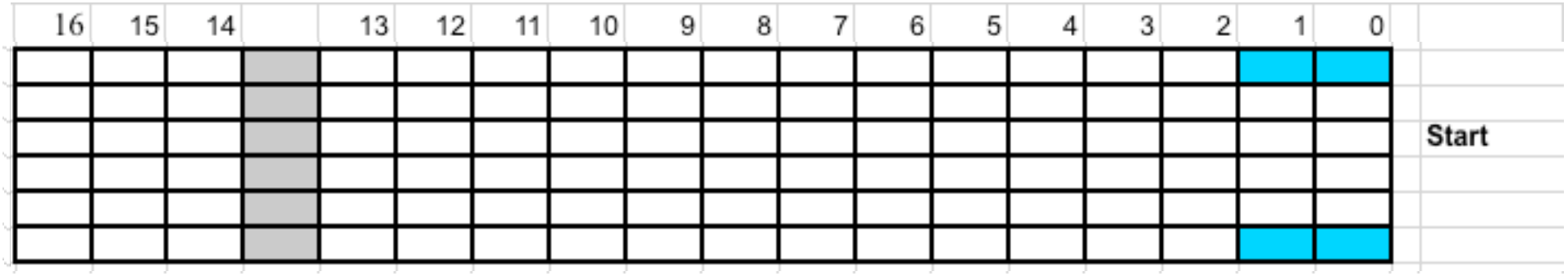
- **Four phases containing "*in-situ*" production and test environments.**
  - **Goal to deliver >= 75% production system during upgrade.**
  - **User testing began in mid-July, no-charge.**
  - **Charging (at DC rate) began Sept 10**
  - **Decision to do CLE 2.1 concurrently**
  - **No extended shutdown**
  - **Torus cleaved but minimal effect**

https://www.nersc.gov/nusers/systems/franklin/quadcore_upgrade.php

Quad Core Upgrade Phases

Legend
Production DC Non-Upgraded
Production SIO DC
Empty Column
Test DC Non-Upgraded
Test SIO Module Installed QC
Test QC Upgraded
Production QC Upgraded
Production SIO Module Installed QC





- **19,320 cores in production**

Thanks to D. Unger (Cray),
Nick Cardo (NERSC) for these figures.

How to Transform a System and Keep Users Happy

- **Phasing served multiple purposes:**
  - Continued science service; no extended downtime
  - Accommodated inventory availability
    - Accommodate availability of trained staff
  - Reduced Risk
  - 98 Days total with only about 2 days out of service
  - Early user access
  - Additional key performance benefit…

Opteron QC Changes

Core
- 128-bit wide SSE3 (4 FLOPs / CP) 9.2 GFLOP/s @ 2.3 GHz
- Increased L1 BW to 2x128bit loads/CP (instead of 2x64bit loads/CP)
- Increase physical addressing to 48bits / 256TB (was 40bits)

Cache
- L2 Cache: Reduced to 512k per core, still private
- L3 2-MB Shared among 4 cores (victim cache for private L2 caches)

TLB
- Adds 1-GB "huge" page support
- L1 TLB: 48 entries fully associative (any page size)
- L2 TLB
  - 512 small pages L2
  - 128 large pages (2M and 1Gig) L2

Memory
- Dual channel memory now "unganged," operates as independent channels.
- Improves chance of hitting an open memory channel and increases parallelism in memory accesses

Prefetch
- Dedicated prefetch buffers in memory controller (no cache evictions for speculative prefetch)
- HW prefetch brings data into L1 (not L2)
- HW prefetch now detects positive/negative/nonunit strides
- SW prefetch for write (prefetchw) treated differently than prefetch for read (similar to PowerPC DCBZ)

Quad Core Benchmarking

"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."

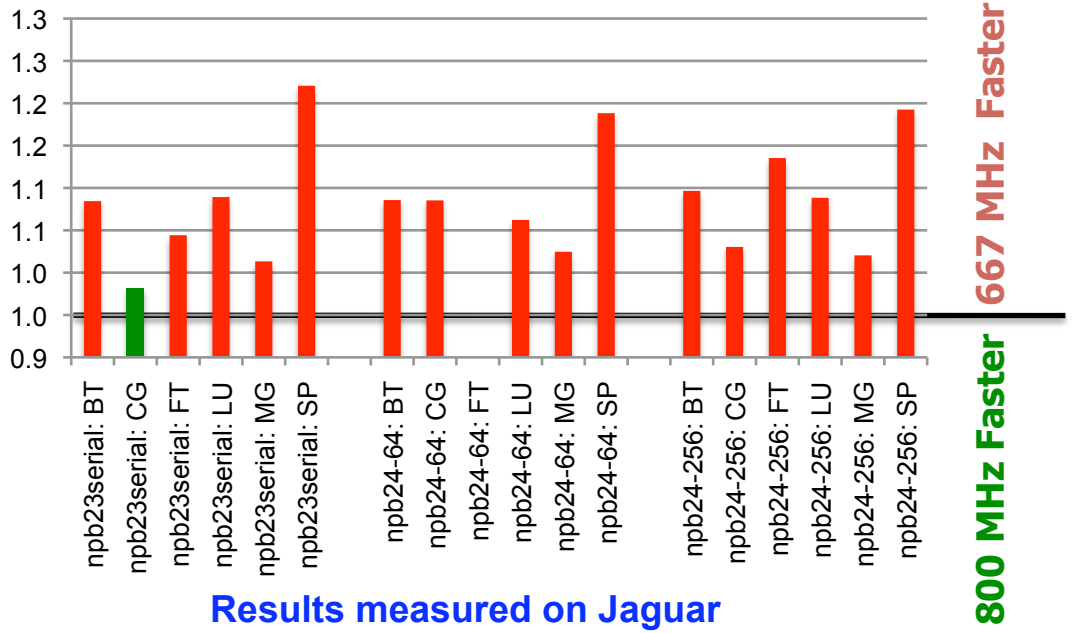
Ingrid Bucher & Joanne Martin, LANL, 1982

Interesting Quad-Core Memory Result

- **Some codes actually perform *better* with 667-MHz memory than with 800 MHz.**
  - **Mostly NPB Class-B serial (packed) and Class-D MPI/64/256**

- **Cannot use 800 MHz un-ganged mode on the XT4.**

- **Essentially no effect for NERSC-5 apps.**



Results measured on Jaguar

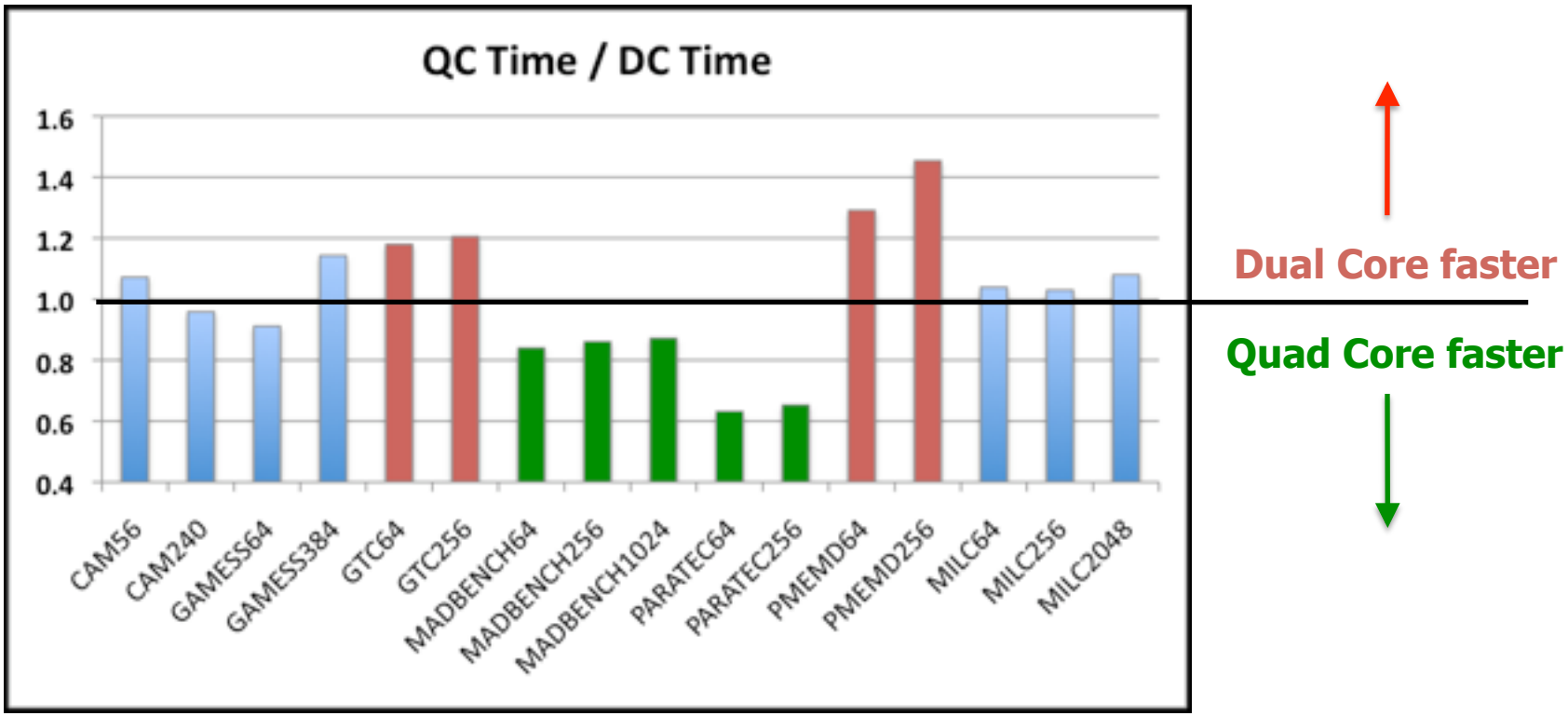
NERSC-5 Application Benchmarks

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

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	64, 384 (Same as Ti-06)	DFT gradient, MP2 gradient	F77	DDI, BLAS
GTC	Fusion (FES)	PIC, finite difference	64, 256 Weak scaling	10 particles per cell	F90	
PMEMD	Life Science (BER)	Particle Mesh Ewald	64, 256 Strong scaling		F90	
MadBench	Astrophysics (HEP & NP)	Power Spectrum Estimation	64,256, 1024 Weak scaling	Vary Npix; 730 MB per task, 200 GB disk	F90	Scalapack, LAPACK
MILC	Lattice Gauge Physics (NP)	Conjugate gradient, sparse matrix; FFT	64, 256, 2048 Weak scaling	16^4 Local Grid, ~4,000 iters	C, assem.	
PARATEC	Material Science (BES)	DFT; FFT, BLAS3	64, 256 Weak scaling	250-686 Atoms, 1372 bands, 10 iters	F90	Scalapack, FFTW

Initial QC / DC Comparison

NERSC-5 Benchmarks

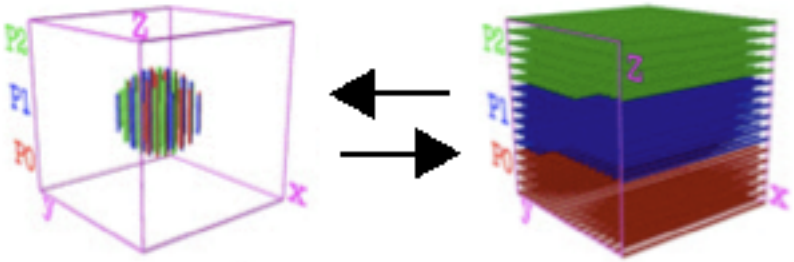


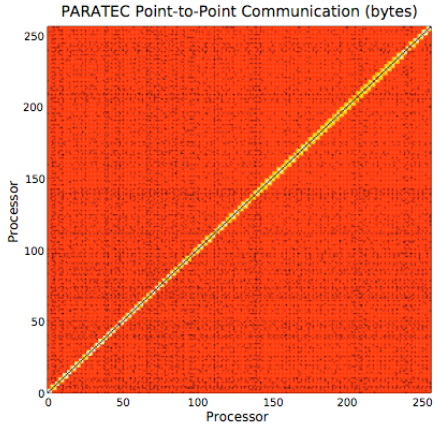
Compare time for *n* cores on DC socket to time for *n* cores on QC socket.

Data courtesy of Helen He, NERSC USG

PARATEC: Parallel Total Energy Code

- **Captures the performance of ~70% of NERSC material science computation.**
- **Planewave DFT; calculation in both Fourier and real space; custom 3-D FFT to transform between.**
- **Uses MPI / SCALAPACK / FFTW / BLAS3**
- **All-to-all data transpositions dominate communications.**





Communication Topology for PARATEC from IPM.

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
Franklin Dual-Core	198	4,524	671 (50%)
Franklin Quad-Core	309	7,517	1,076 (46%)
Jaguar Quad-Core	270	6,397	966 (45%)
BG/P	207	567	532 (61%)
HLRB-II	194	993	760 (46%)
BASSI	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

NERSC Sustained Performance

- 7 application benchmarks
- Two machines (DC & QC)

- How do we summarize performance?
- How do we express computing capability over time?

Sustained System Performance (SSP)

- **Aggregate, un-weighted measure of sustained 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 predetermined on a reference system by NERSC.**

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

NERSC Composite SSP Metric

The time for the largest concurrency run of each full application benchmark is used to calculate the SSP.

NERSC-5 SSP

CAM	GAMESS	GTC	PMEMD	MADBench	MILC	PARATEC
240	384	256	256	1024	2048	256
Climate Modeling	Quantum Chemistry	Fusion	MolDyn	Astro	Lattice QCD	MatSci DFT

For Franklin DualCore, N = 19,344
QuadCore, N = 38,640

Franklin QC Upgrade SSP

- **Performance of Franklin is expected to go from**

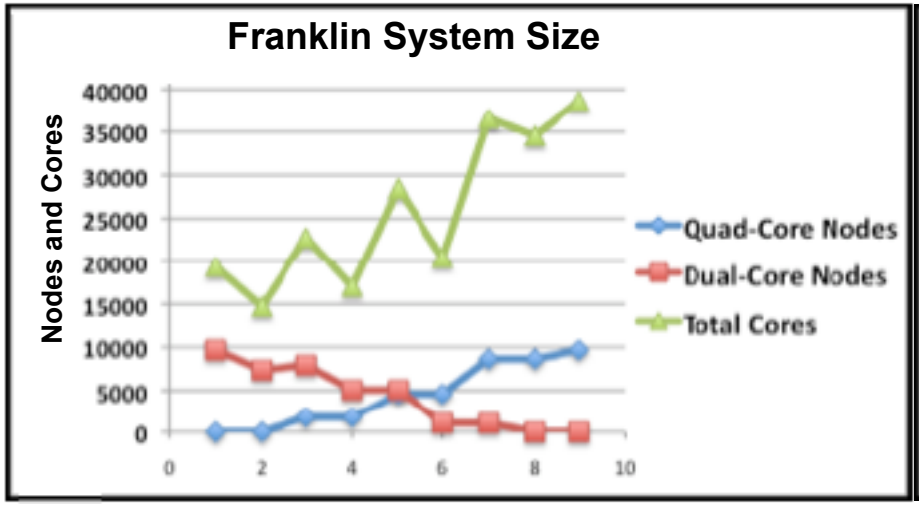
  SSP = 19.3 TF (Oct. 2007)

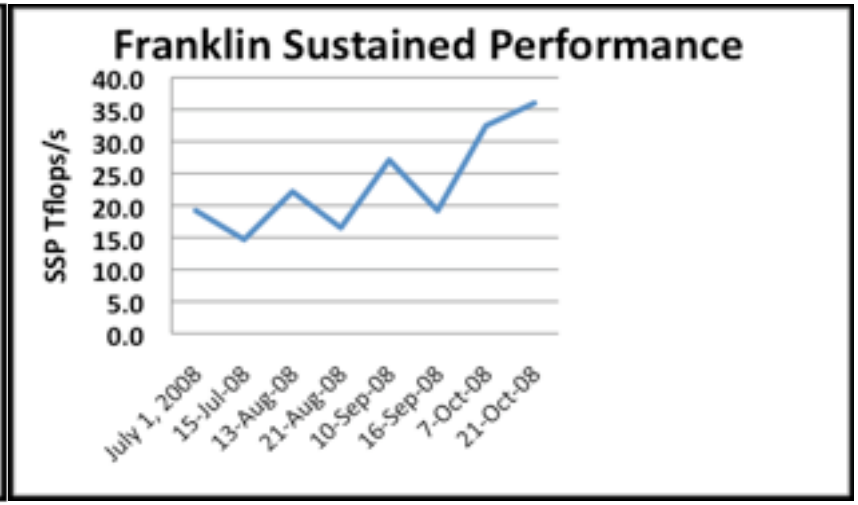
  to

  SSP ≅ 38 TF

- **Why does in-place QC upgrade matter?**

Maintaining Service While Improving Service

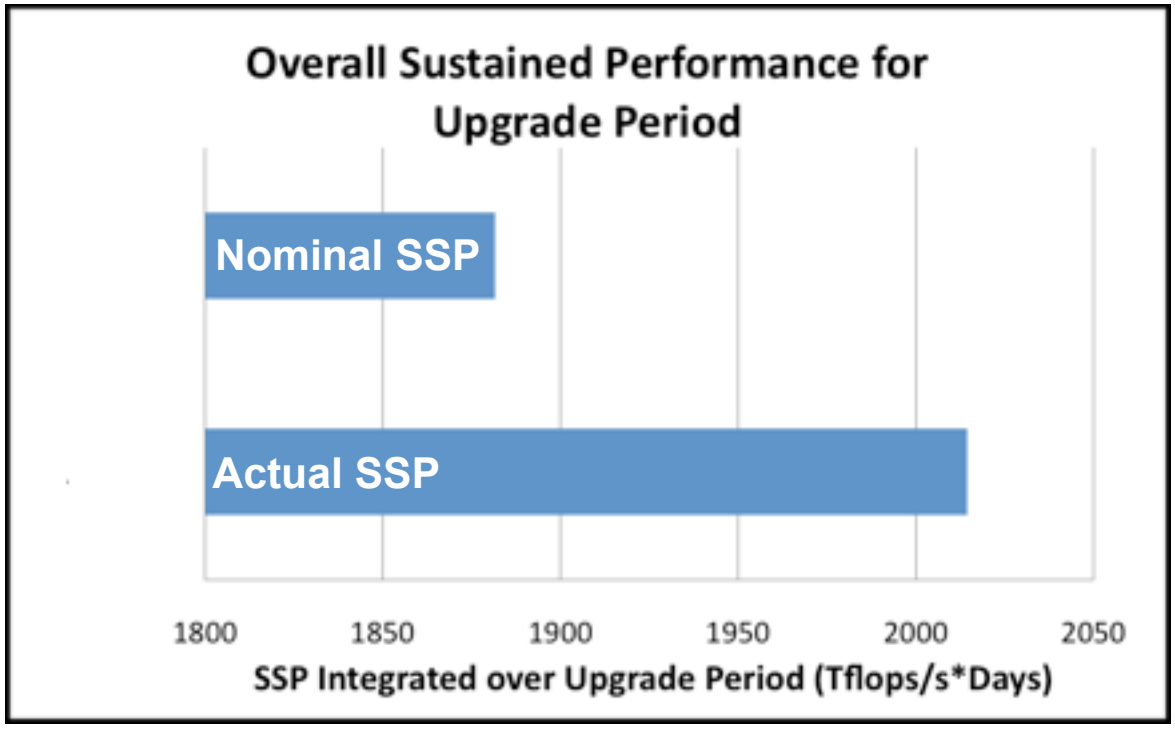




Phase	Start Date	Number of Dual Core Racks	Number of Quad Core Racks	Sustained Performance (SSP Tflops/s)	SSP Tflop/s-Days
Before	July 1, 2008	102	0	19.2	
1	15-Jul-08	78	0	14.7	425.8
2a	13-Aug-08	84	18	22.2	177.3
2b	21-Aug-08	54	18	16.5	330.4
3a	10-Sep-08	54	48	27.1	162.6
3b	16-Sep-08	12	48	19.2	403.2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

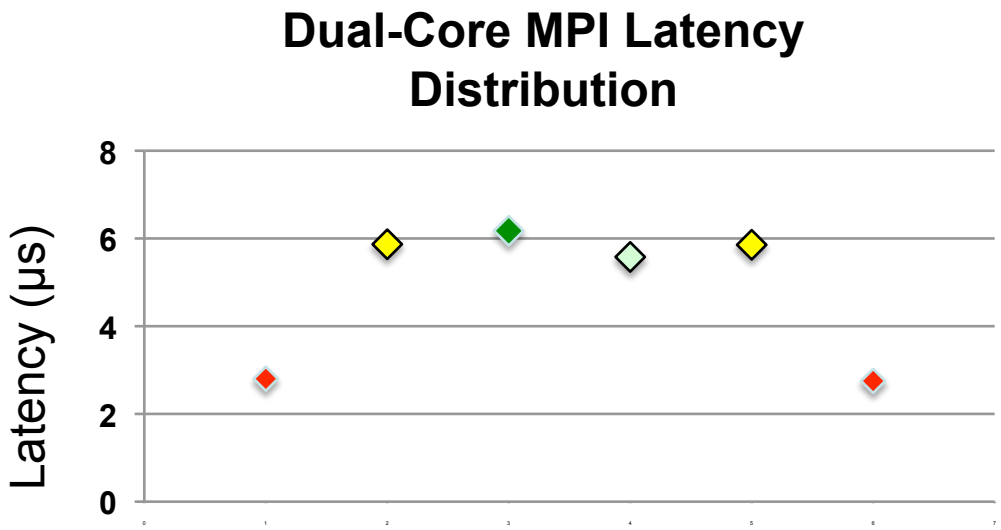
Key Phased Upgrade Benefit

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



MPI Latency

- **Core 0 on each node handles all SeaStar interrupts.**
  - **Results in one core being MPI-favored**



Unfavored-Unfavored (6.2 µs)

Unfavored-Favored (5.8 µs)

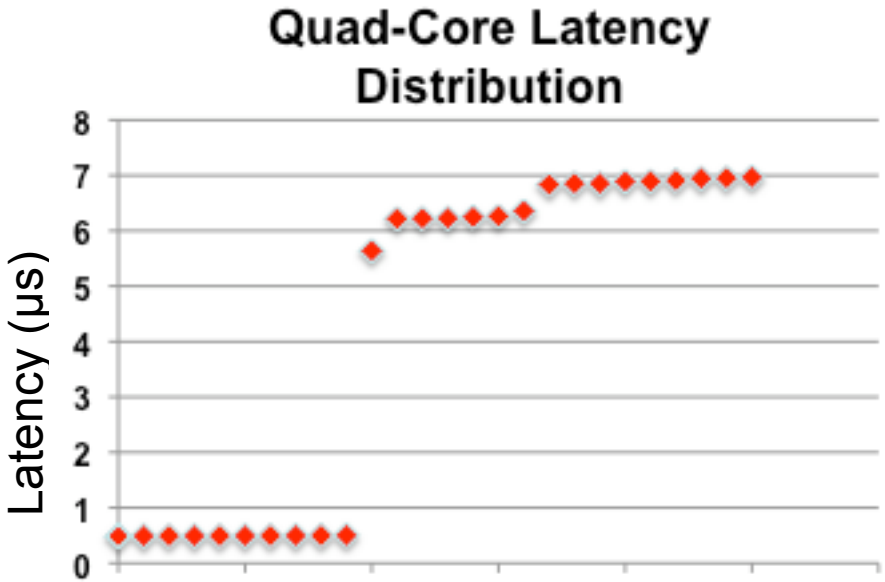
Favored-Favored (5.6 µs)

Intra-node (2.8 µs)

Thanks to Joe Glenski, et al. (Cray), for pointing this out.

MPI Latency

- **Quad-core XT4: Favored-favored less likely**
  - **Intra-node much lower; worst-case slightly higher**



12 Unfavored-Unfavored (6.9 µs)

6 Unfavored-Favored (6.2 µs)

1 Favored-Favored (5.6 µs)

Intra-node (0.48 µs)

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

NERSC Next-Generation System

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 and available for DOE allocation**
  - **RFP release September 4, 2008.**
  - **Approach designed to select the best machine for science with greatest flexibility for both NERSC and vendors.**

NERSC-6 Benchmarks

- **New codes/methods address evolution of the workload, emerging programming models, algorithms**
  - **New SSP applications: MAESTRO and IMPACT-T**
  - **UPC, AMR, implicit and sparse methods**
  - **Comprehensive workload study:**
  - http://www.nersc.gov/projects/procurements/NERSC6/NERSC6Workload.pdf
- **Largest concurrency increases from 2,048 to 8,196**
  - **Increased focus on strong scaling**
- **Two ways for vendors to run benchmarks…**

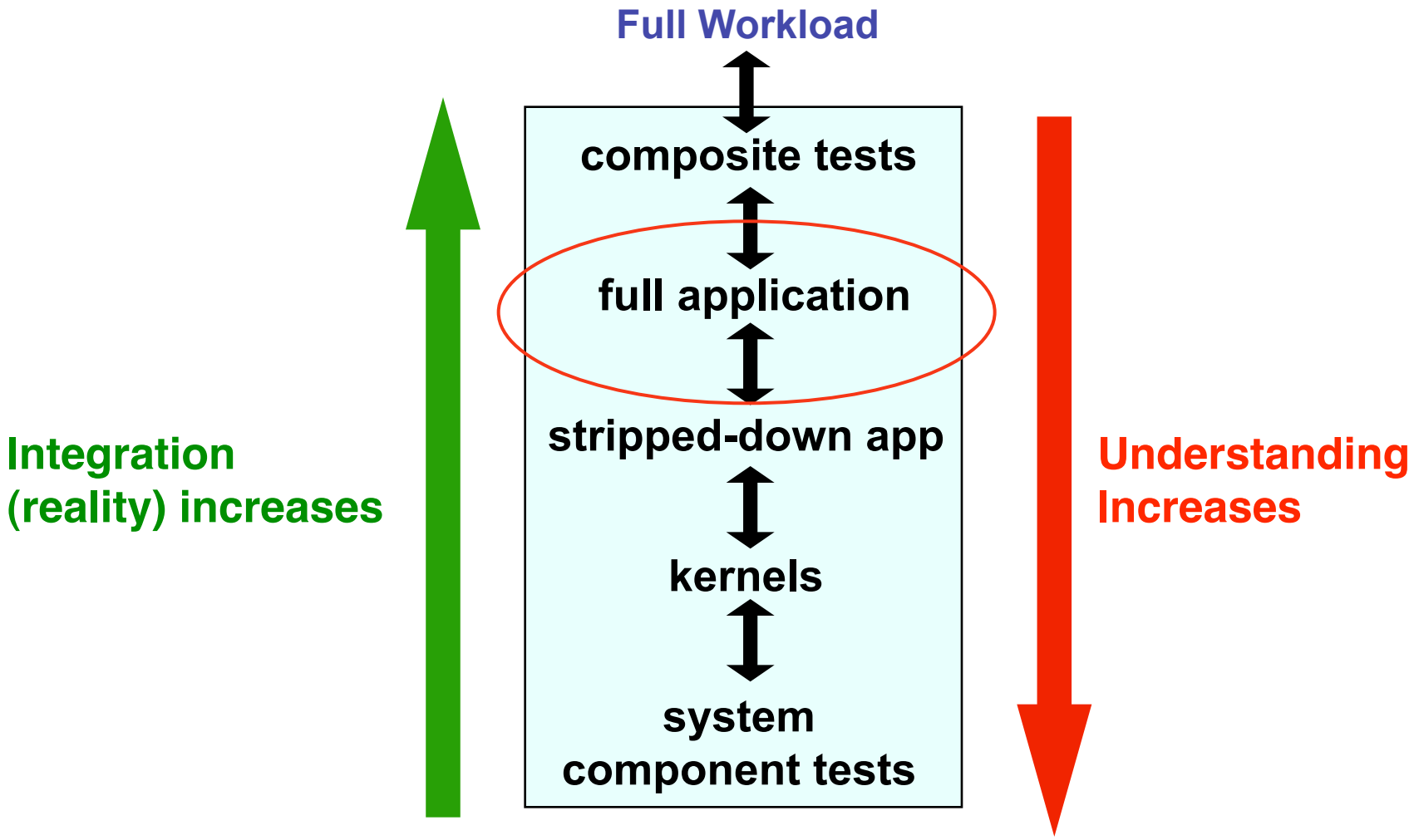
Base Case for Application Runs

- LCD for comparison among proposed systems.
- Limits the scope of optimization.
  - Modifications only to enable porting and correct execution.
- Limits allowable concurrency to prescribed values.
- MPI-only (even if OpenMP directives present).
- Fully packed nodes.
- Libraries okay (if generally supported).
- Hardware multithreading okay, too.
  - Expand MPI concurrency to occupy hardware threads.

Optimized Case for Application Runs

- Allow the Offeror to highlight features of the proposed system.
- Applies to seven SSP apps only, all test problems.
- Examples:
  - Unpack the nodes;
  - Higher (or lower) concurrency than reference base case;
  - Hybrid OpenMP / MPI;
  - Source code changes for data alignment / layout;
  - Any / all of above.
- Caveat: SSP based on total number of processors blocked from other use.

Use a Hierarchy of Benchmarks



NERSC-6 Application Benchmarks

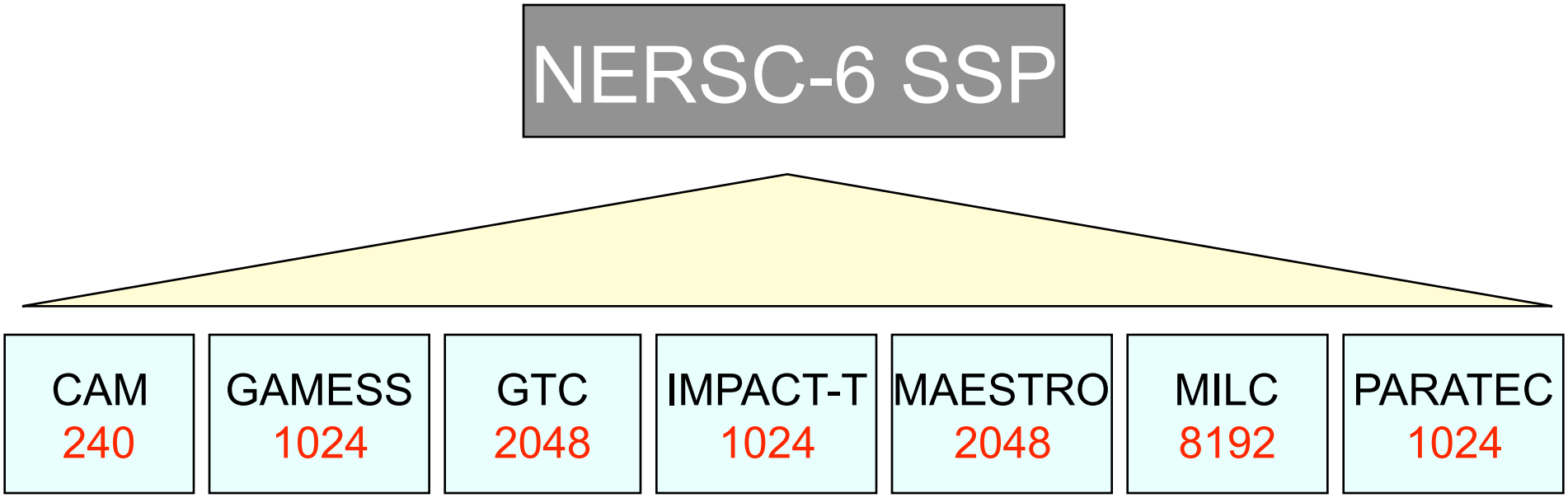
Benchmark	Science Area	Algorithm Space	Base Case Concurrency	Problem Description	Lang	Libraries
CAM	Climate (BER)	Navier Stokes CFD	56, 240 Strong scaling	D Grid, (~.5 deg resolution); 240 timesteps	F90	netCDF
GAMESS	Quantum Chem (BES)	Dense linear algebra	256, 1024 (Same as Ti-09)	DFT gradient, MP2 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

Lower-Level Benchmarks

CODE	PURPOSE / DESCRIPTION
STREAM	Single- and multi-core memory bandwidth.
FCT	Full-Configuration Test, run a single app over all cores; FFT mimics planewave DFT codes.
PSNAP	FWQ operating system noise test.
NAS PB serial & 256-way MPI	Serial application performance on a single **packed** node; measures memory BW/ computation rate balance and compiler capabilities. **Packed** means all cores run.
NAS PB UPC	Measure performance characteristics not visible from MPI for FT benchmark.
Multipong	NERSC MPI PingPong for “latency” and BW, nearest- and furthest nodes in topology; also intra-node.
AMR Elliptic	C++/F90 LBNL Chombo code; proxy for AMR Multigrid elliptic solvers; 2 refinement levels; weak scaling with geometry replication; very sensitive to OS noise;

NERSC-6 Composite SSP Metric

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

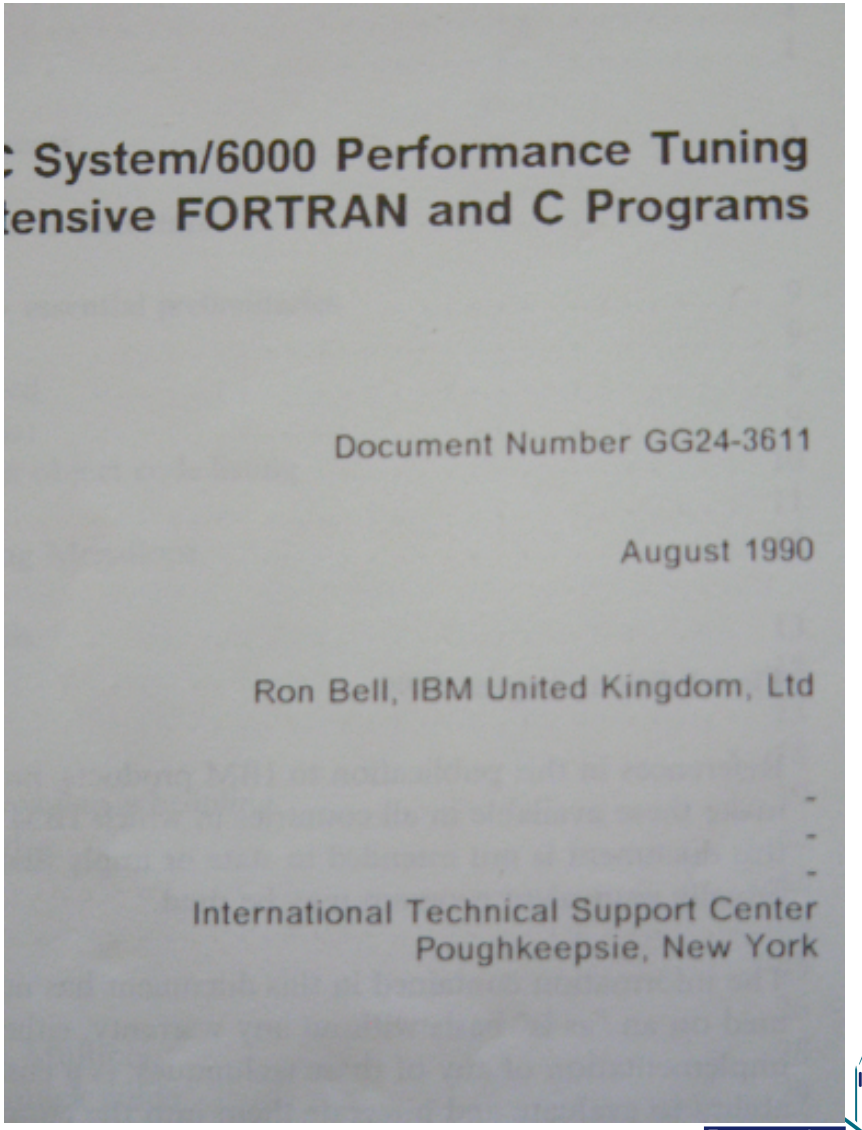


For each benchmark measure

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

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Acknowledgement



Acknowledgements

- Kathy Yelick, NERSC Director
- Bill Kramer, NERSC General Manager
- John Shalf, NERSC SDSA Lead
- Nick Cardo, NERSC Franklin Project Lead
- Helen He, Jonathan Carter, Katie Antypas, NERSC USG
- Joe Glenski, et al., Cray

Summary

- **Science continues to thrive at NERSC.**
  - **Still fun, too.**
- **System improvement continuing at (an appropriately) rapid pace.**
  - **HW & SW**
- **Users very satisfied.**

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

"Backup" Slides

Office of Science
U.S. DEPARTMENT OF ENERGY

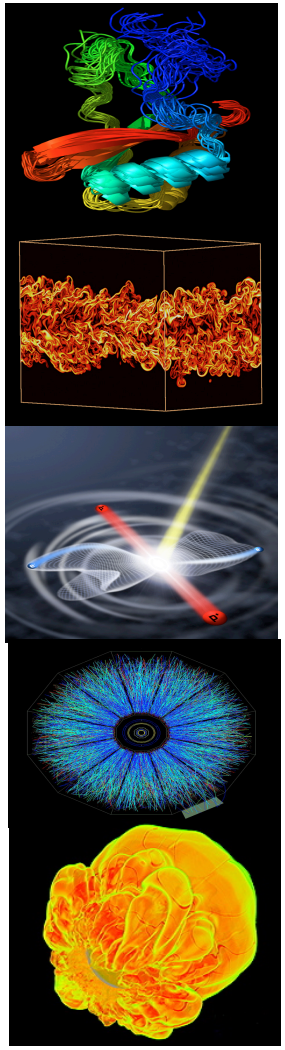
BERKELEY LAB

Science Driven System Architecture Group

- **Analyze requirements of broad scientific workload**
  - **Benchmarking**
  - **Algorithm tracking**
- **Track future trends in supercomputing architecture**
  - **Assess emerging system technologies**
- **Understand bottlenecks in current computing architecture**
  - **Use the NERSC workload to drive changes in computing architecture.**

http://www.nersc.gov/projects/SDSA

About the Cover



Schematic representation of 2° 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

PARATEC: Performance

Medium Problem (64 cores)

	Dual Core	Quad Core	Ratio
FFTs	425	537	1.3
Projectors	4,616	7,779	1.7
Matrix-Matrix	4,744	8,211	1.7
Overall	2,902 (56%)	4,594 (50%)	1.6

Large Problem (256 cores)

	Dual Core	Quad Core	Ratio
FFTs	198	309	1.6
Projectors	4,524	7,517	1.7
Matrix-Matrix	4,726	8,197	1.7
Overall	2,803 (56%)	3,971 (43%)	1.4

- **Rates in MFLOPS from PARATEC output.**
- **Projector rate is dominated by BLAS3 routines.**
  - **SciLIB takes advantage of wider SSE in Barcelona.**

Anatomy of an O(N) DFT method

(LS3DF as an example)

- **Total energy of a system can be decomposed into two parts**
  - **Quantum mechanical part:**
    - wavefunction kinetic energy and exchange correlation energy
    - Highly localized
    - Computationally expensive part to compute
  - **Classical electrostatic part:**
    - Coulomb energy
    - Involves long-range interactions
    - Solved efficiently using poisson equation even for million atom systems

- **LS3DF exploits localization of quantum mechanical part of calculation**
  - Divide computational domain into discrete tiles and solve quantum mechanical part
  - Solve global electrostatic part (no decomposition)
  - Very little interprocessor communication required! (almost embarrassingly parallel)
  - Result is O(Natoms) complexity algorithm: enables exploration of larger atomic systems as we move to petaflop and beyond.