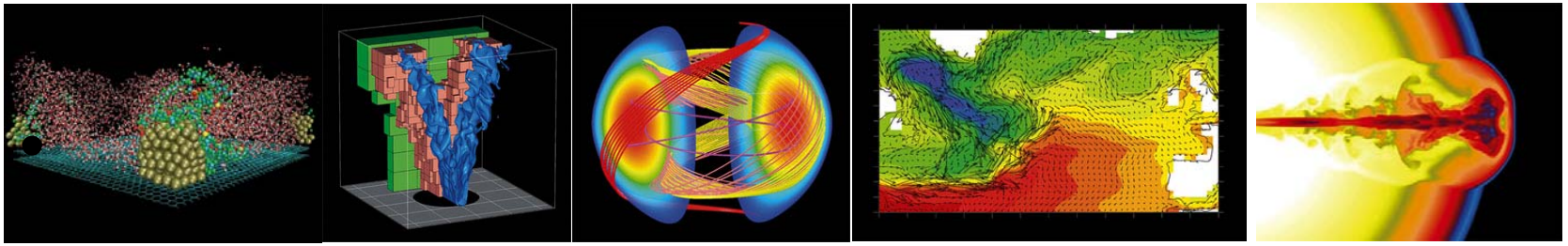


PERCU Results in a A Reawakened Relationship for NERSC and Cray

William T.C Kramer
NERSC General Manager

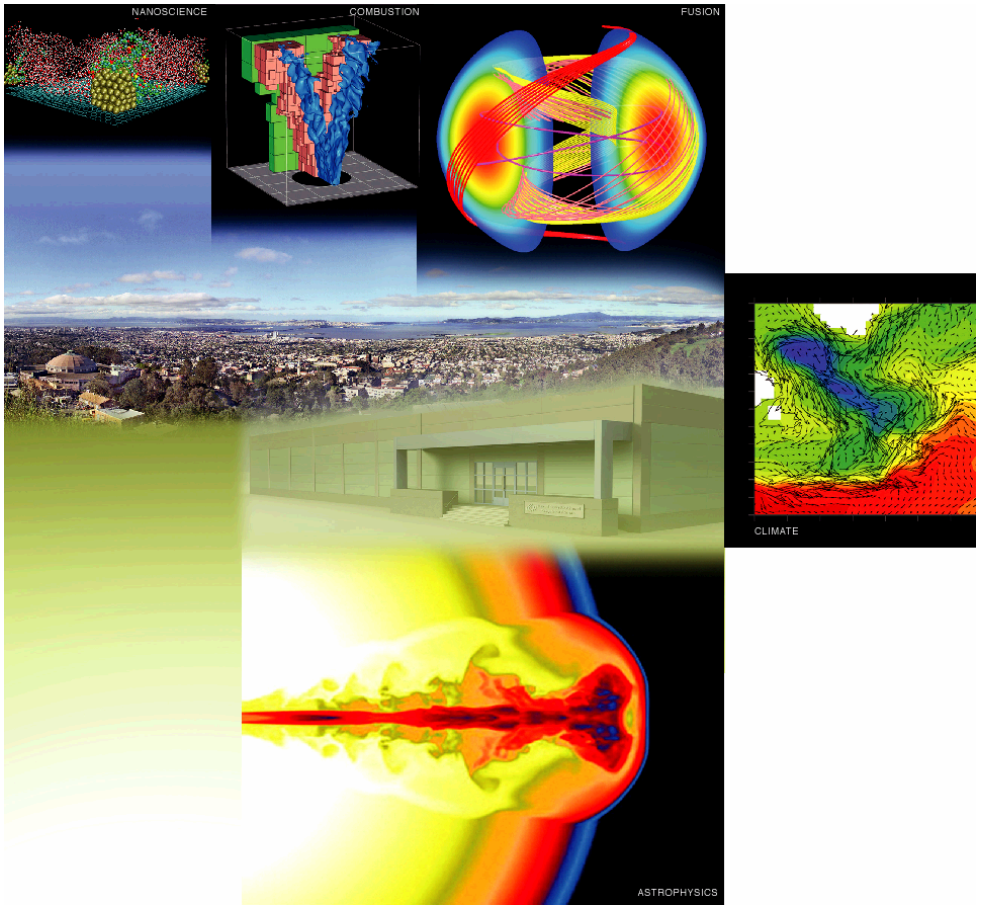
kramer@nersc.gov

510-486-7577

Ernest Orlando Lawrence
Berkeley National Laboratory

This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

Outline



- Background about NERSC
- NERSC-5
  - How we decide
  - Details about NERSC-5
- Current Status
- Future Plans

NERSC Mission

NERSC is the DOE Office of Science *Flagship* HPC Facility as well as a Leadership facility.

The mission of the National Energy Research Scientific Computing (NERSC) Facility is to accelerate the pace of scientific discovery by providing high performance computing, information, data, and communications services for all research sponsored by the DOE Office of Science (SC).

NERSC is also the senior computational facility in the Office of Science – being founded in 1974

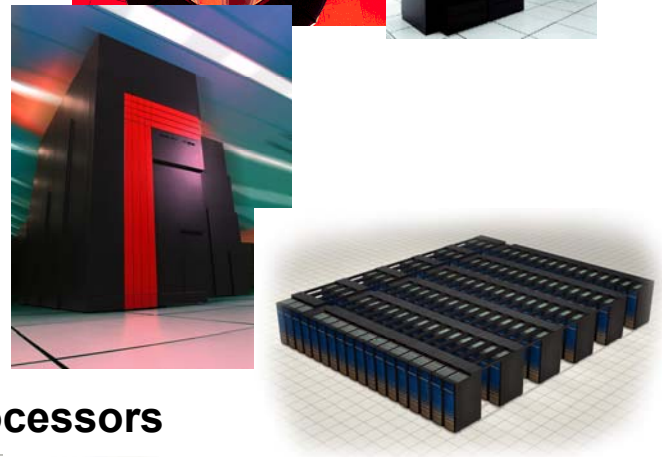
NERSC Facility Overview

- Funded by DOE, FY06-07 annual budget $38M, about 60 staff
  - Expected to increase in FY 08-12
- Supports open, unclassified, basic and applied research
- Delivers a complete, balanced HPC environment (computing, storage, visualization, networking, grid services, cyber security)
- Focuses on intellectual services to enable computational science on the most capable HPC equipment
- Provides close collaborations between universities and other research groups in computer science and computational science

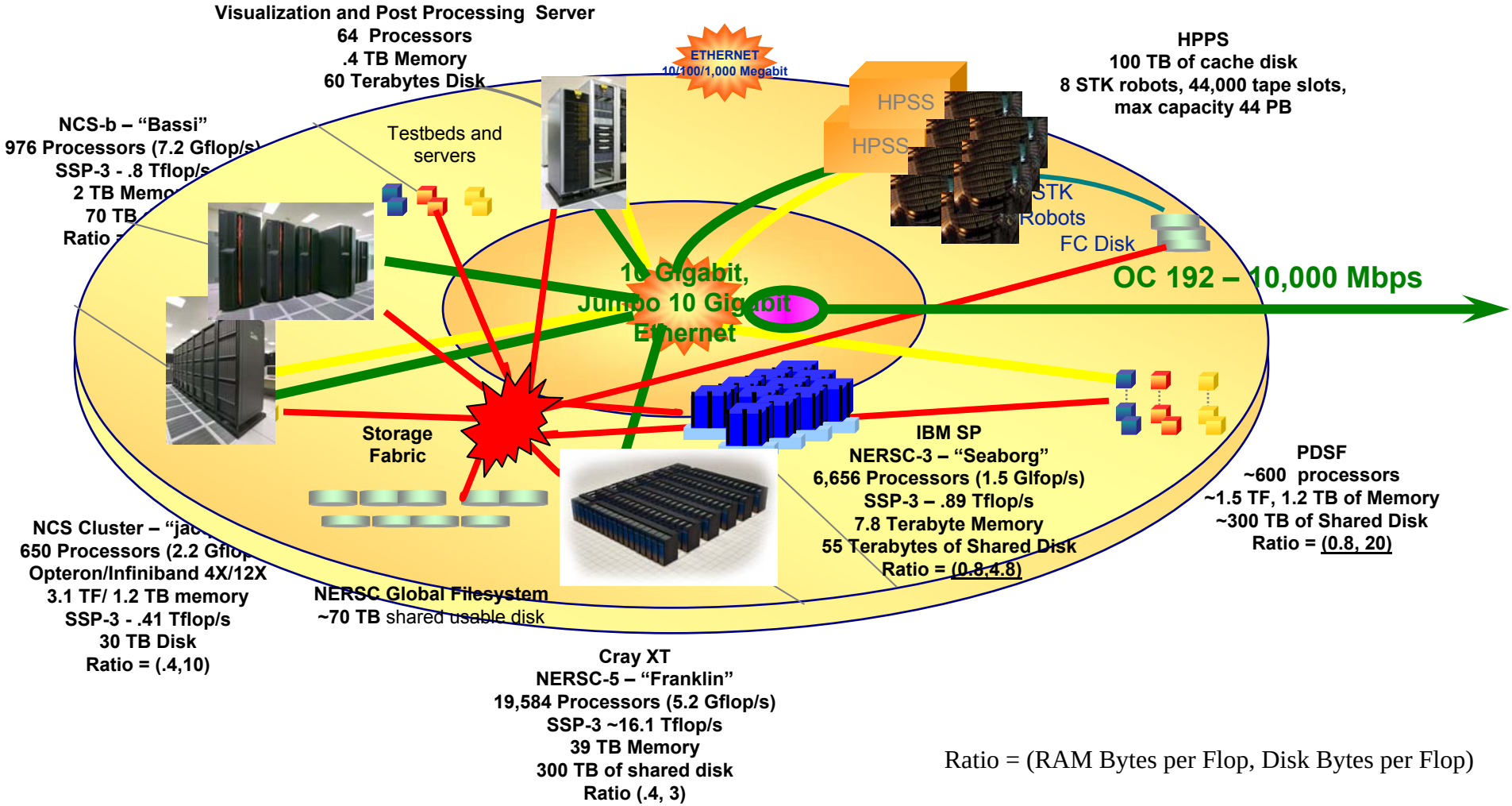
Science-Driven Computing Strategy 2006 -2010

NERSC and Cray have a Rich History

- 1974 - NERSC began with a CDC 6600
- 1975 – Used LBNL CDC 7600 remotely
- 1978 – Cray 1 (SN 6)
  - CTSS first used
  - NERSC joins CUG
- 1981 – Second Cray 1
- 1984 – Cray XMP
- 1985 – First Cray-2 (SN 1)
  - Demonstrated UNICOS
- 1990 – Only 8 processor Cray-2
- 1992 – 8 processor XMP
- 1993 – 16 processor C-90 (SN 4005)
- 1994 – Installed early T3D
- 1996 – NERSC moves to LBNL
- 1996 – 128 processor T3E-600 (SN 6306) and J-90 (SN 8192)
- 1997 – Added 512 processor T3E-900 (SN 6711)
  - Unicos/mk
  - First C/R on an MPP
- 1998 – Increase T3E-900 to 696 processors
- 1998 - Installed first SV1s (SNs 9601, 02, 05)
- 2007 – Installed largest XT4 (SN 4501) – 19,584 processors



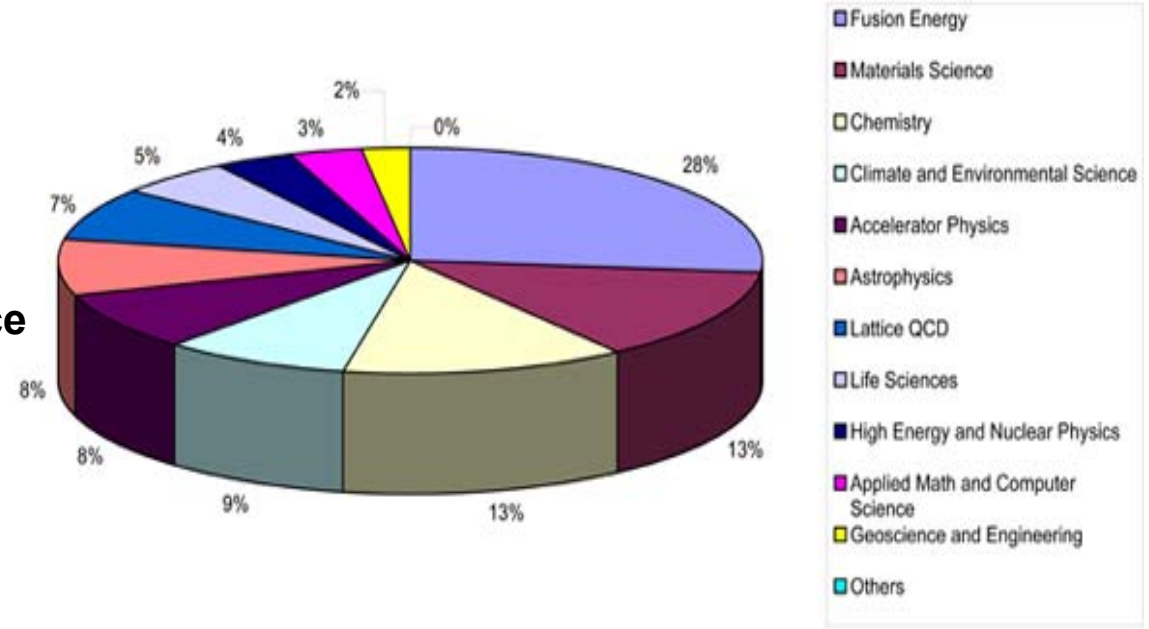
2007

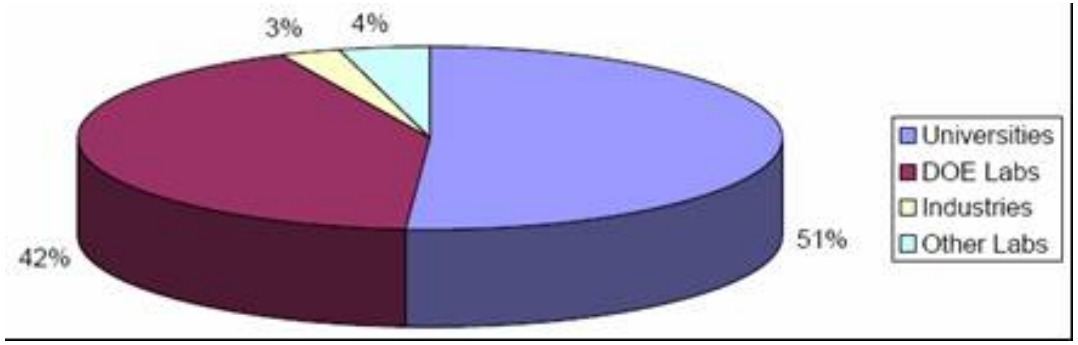


Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

Support Different Types of Usage

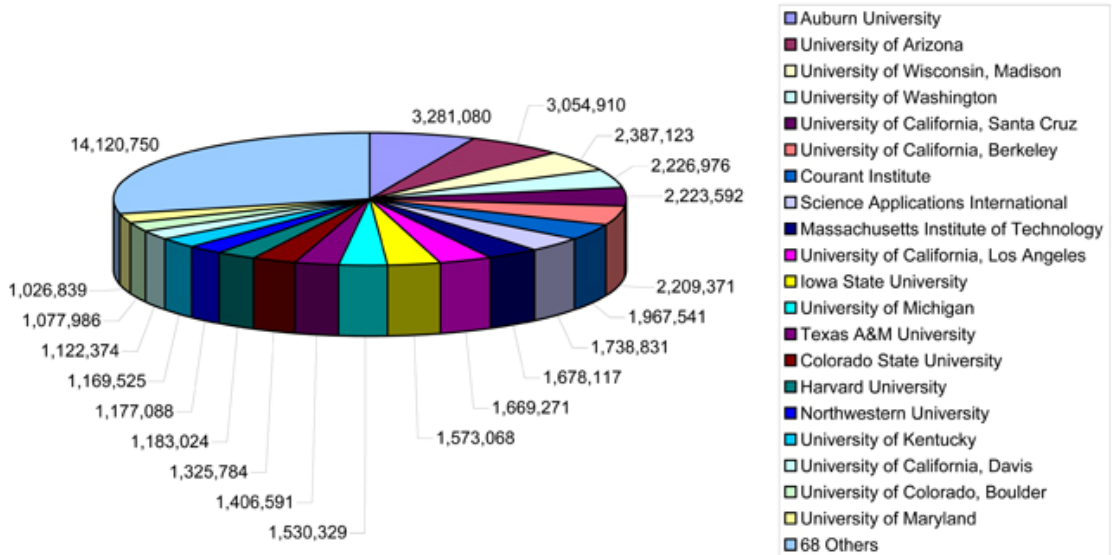
- National/International User Community
- Different types of projects
  - Single PI projects
  - Large computational science collaborations
  - Special National Projects
    - INCITE
    - SciDAC-II
    - National Need
- Large variety of applications
  - All scientific applications in DOE SC
- Range of Systems
  - Computational, storage, networking, analytics

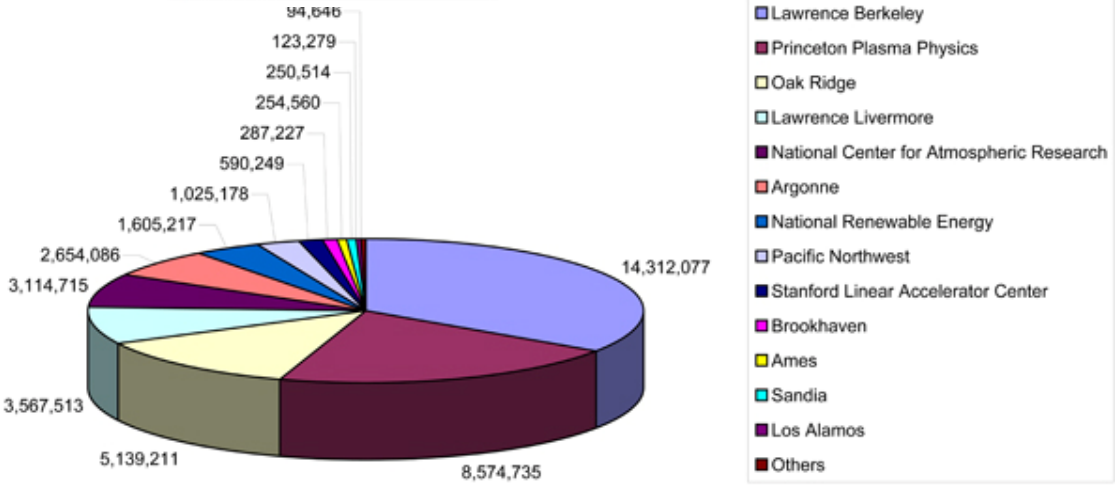




Office of Science
U.S. DEPARTMENT OF ENERGY

Institutional Usage





Number of Awarded Projects

Allocation Year	Production	INCITE & Big Splash	SciDAC	Startup
2007 (as of February)	291	7	45	44
2006	286	3	36	70
2005	277	3	31	60
2004	257	3	29	83
2003	235	3	21	76

New Applications and Algorithms Matrix

Science areas	Multi-physics, Multi-scale	Dense linear algebra	Sparse linear algebra	Spectral Methods (FFT)s	N-Body Methods	Structured Grids	Unstructured Grids	Data Intensive
Nanoscience	X	X	X	X	X	X		
Climate	X			X		X	X	X
Chemistry	X	X	X	X	X			
Fusion	X	X	X			X	X	X
Combustion	X		X			X	X	X
Astrophysics	X	X	X	X	X	X	X	X
Biology	X	X					X	X
Nuclear		X	X		X			X

Phil Colella's Seven Dwarfs analogy

New Applications and Algorithms Matrix

Science areas	Multi-physics, Multi-scale	Dense linear algebra	Sparse linear algebra	Spectral Methods (FFT)s	N-Body Methods	Structured Grids	Unstructured Grids	Data Intensive
Nanoscience	General purpose balanced system	High speed CPU, high Flop/s rate	High performance memory system	Bisection interconnect bandwidth	High performance memory system	High speed CPU, high Flop/s rate	Irregular data and control flow	Storage, Network Infrastructure
Climate								
Chemistry								
Fusion								
Combustion								
Astrophysics								
Biology								
Nuclear								

Phil Colella's Seven Dwarfs analogy

Changing Science of INCITE

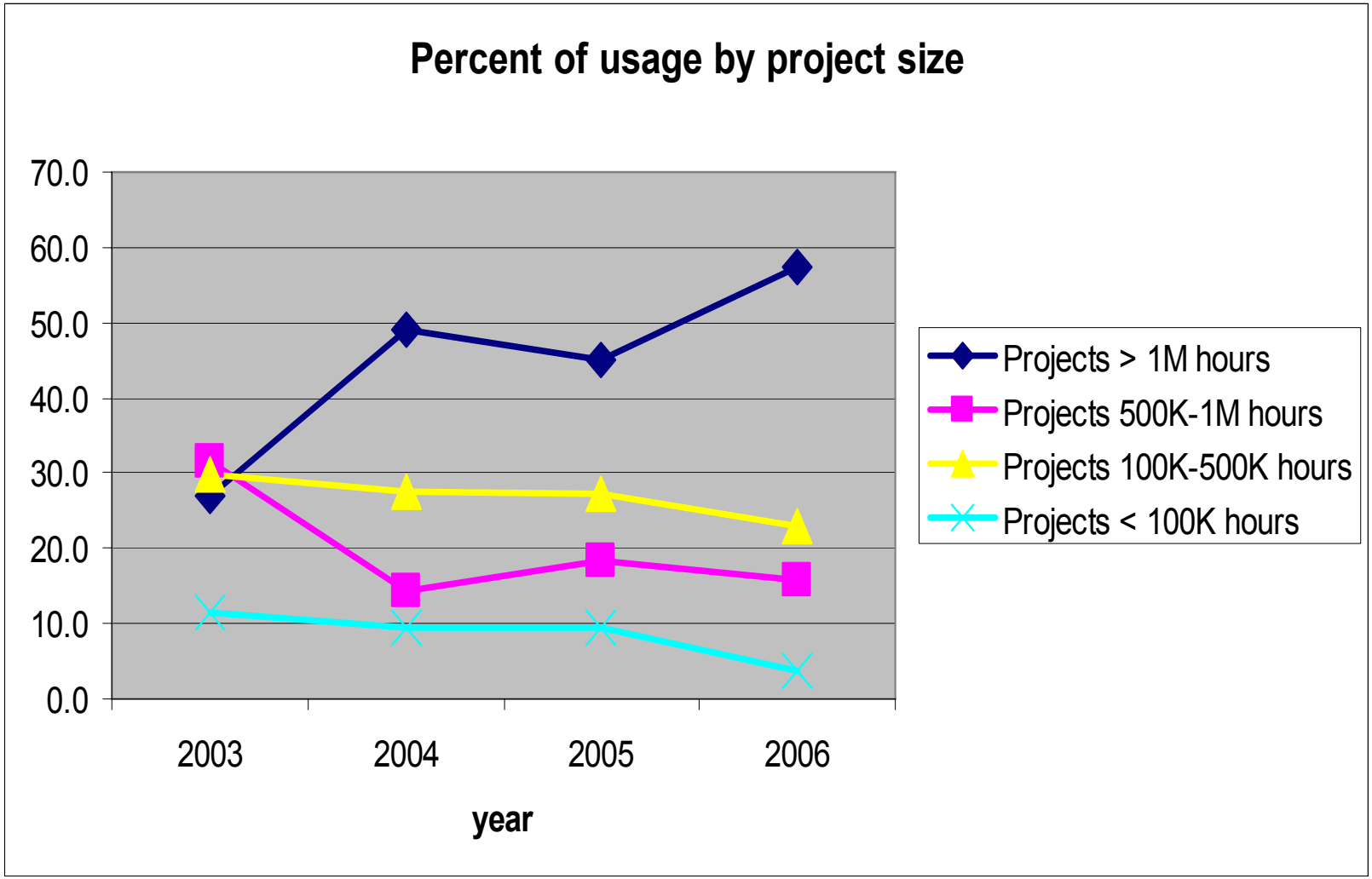
Year	Chemistry	Astrophysics	CFD	Biology	Accelerator Physics	Combustion	Climate	Fusion Energy
2004	X	X	X					
2005		X		X		X		
2006	X	X			X			
2007	X	X	X				X	X

New Changing Algorithms of INCITE

Year	Multi Physics/Multi Scale	Dense LA	Sparse LA	Spectral Methods	N-Body Methods	Structured Grids	Unstructured Grids	Map Reduce	Data Intensive
2004		X				X		X	X
2005	X	X		X	X	X			X
2006	X	X			X	X			X
2007	X	X				X	X		

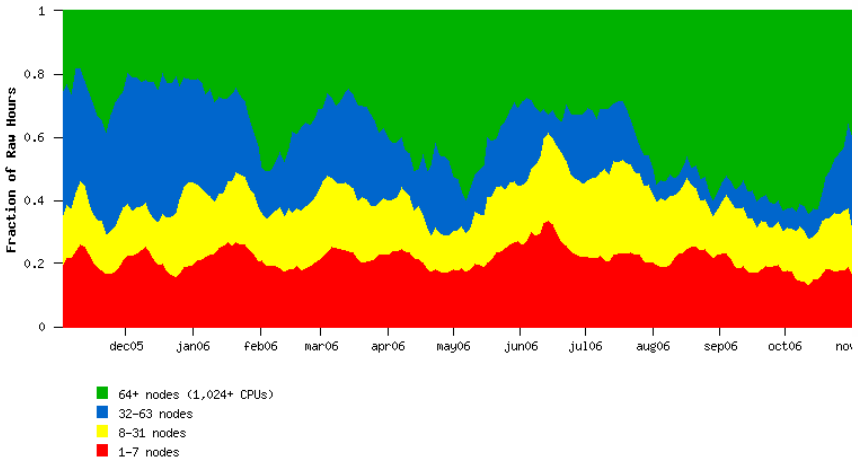
Phil Colella's Seven Dwarfs analogy

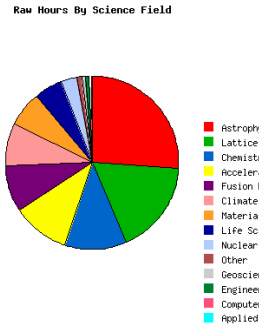
Large Scale Science

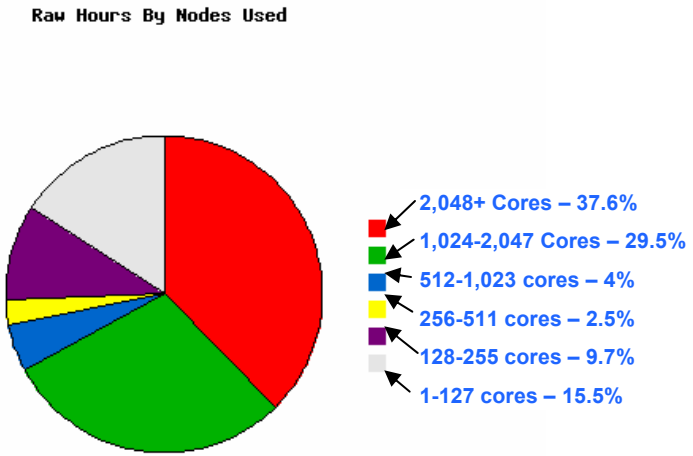


Large Scale Is Key

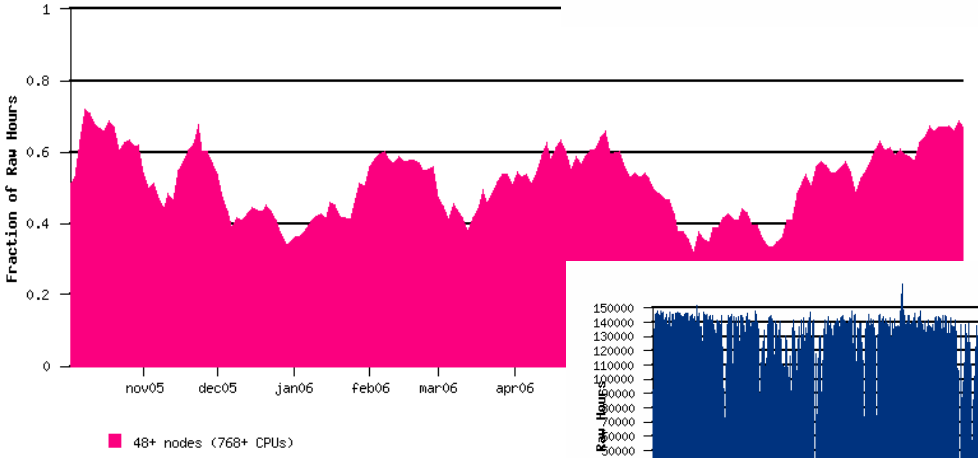
Discipline usage and Job Size since January 2007

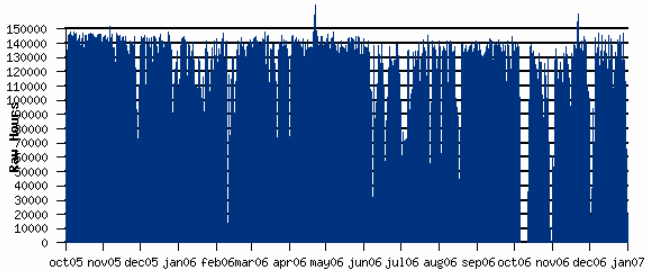






Percent of overall time used by science users	
AY 2004	**90.0%**
AY 2005	**93.5%**
AY 2006	**87.5%**
AY 2007 To date	**88.5%**





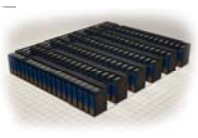
Berkeley Lab

Office of Science
U.S. Department of Energy

Some Example Science

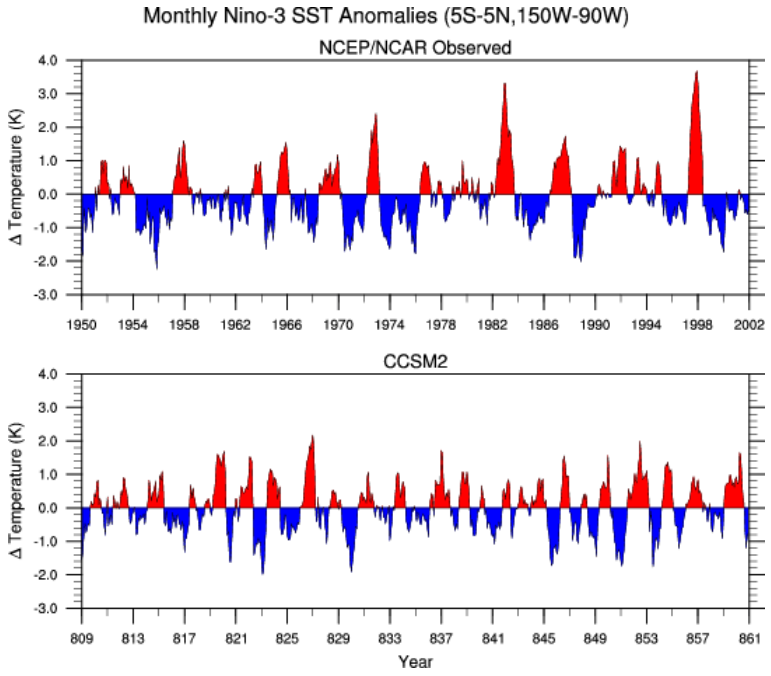




1000-Year Climate Simulation

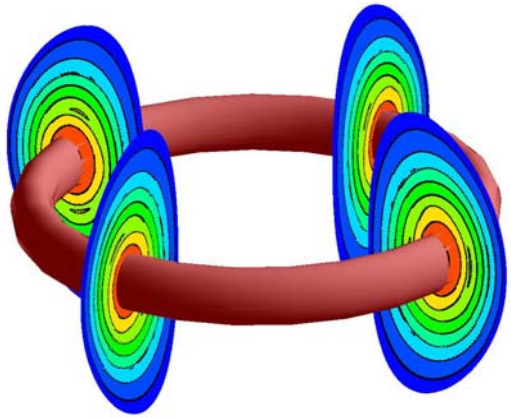
- ***Warren Washington and Jerry Meehl, National Center for Atmospheric Research; Bert Semtner, Naval Postgraduate School; John Weatherly, U.S. Army Cold Regions Research and Engineering Lab Laboratory.***

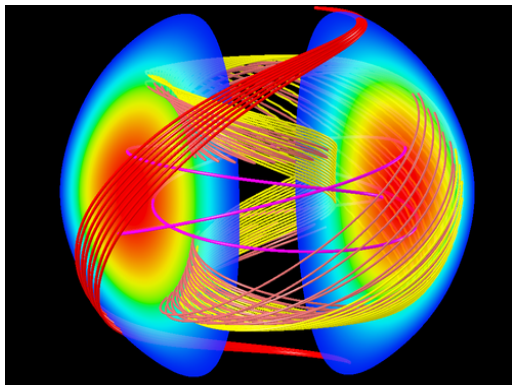


- **1000-year simulation demonstrates the ability of the new Community Climate System Model (CCSM2) to produce a long-term, stable representation of the Earth's climate.**
- **NERSC:**
  - **service and stability**
  - **special queue support**
  - **daily runs without impacting the rest of the workload**

Enabling Algorithms Tech. Transfer

- **NIMROD is a parallel fusion plasma modeling code using fluid-based nonlinear macroscopic electromagnetic dynamics.**
- **Joint work between CEMM and TOPS led to an improvement in NIMROD execution time by a factor of 5-10 on the NERSC IBM SP.**
- **This would be the equivalent of 3-5 years progress in computing hardware.**





- **Parallel SuperLU, developed at LBNL, has been incorporated into NIMROD as an alternative linear solver.**
  - **Physical fields are updated separately in all but the last time advances, allowing the use of direct solvers.**
    SuperLU is >100x and 64x faster on 1 and 9 processors, respectively.
  - **A much larger linear system must be solved using the conjugate gradient method in the last time-advance. SuperLU is used to factor a preconditioning matrix resulting in a 10-fold improvement in speed.**

http://w3.pppl.gov/CEMM

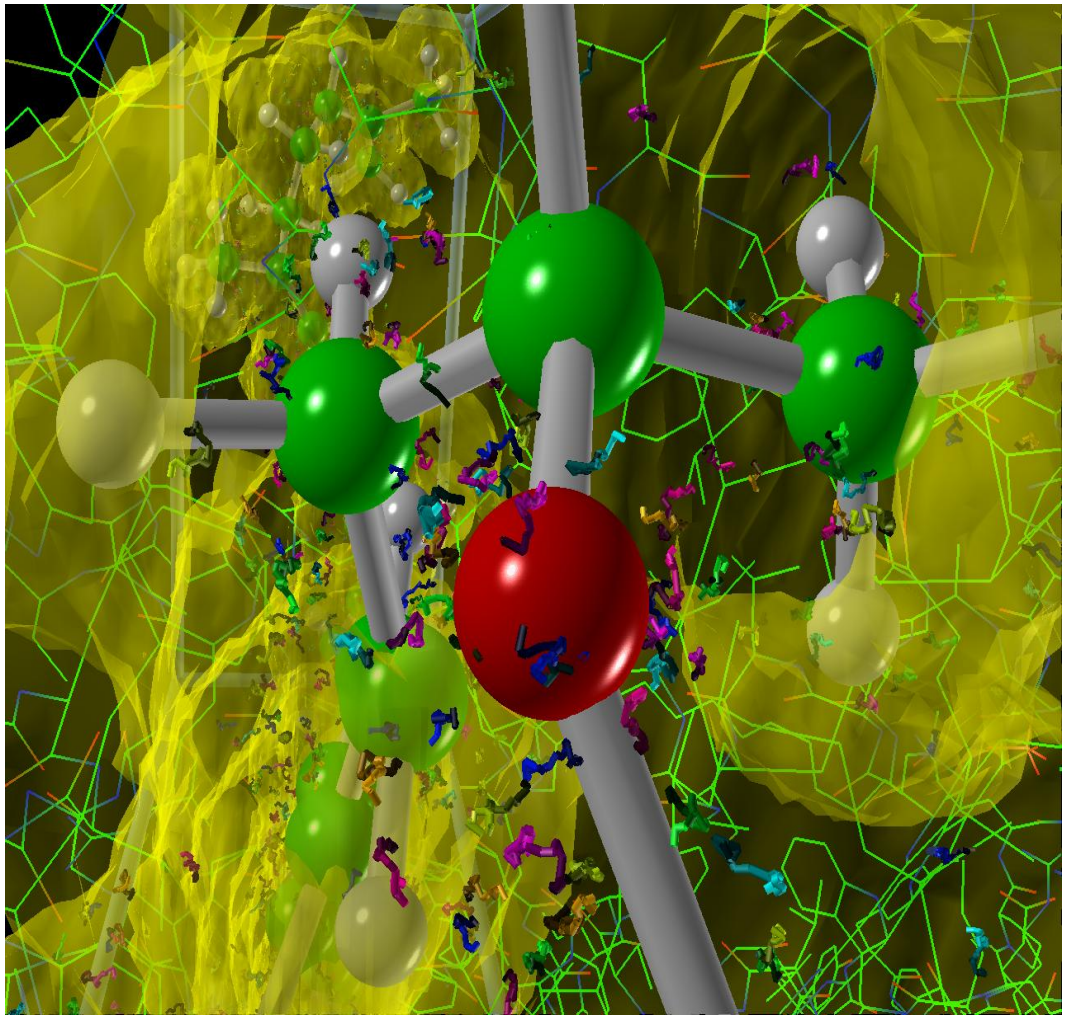
http://www.tops-scidac.org

Photosynthesis INCITE Project

- MPI tuning: 15-40% less MPI time
- Quantum Monte Carlo scaling: 256 to 4,096 procs
- More efficient random walk procedure
- Wrote parallel HDF layer
- Used AVS/Express to visualize molecules and electron trajectories
- Animations of the trajectories showed 3D behavior of walkers for the first time

"Visualization has provided us with modes of presenting our work beyond our wildest imagination"

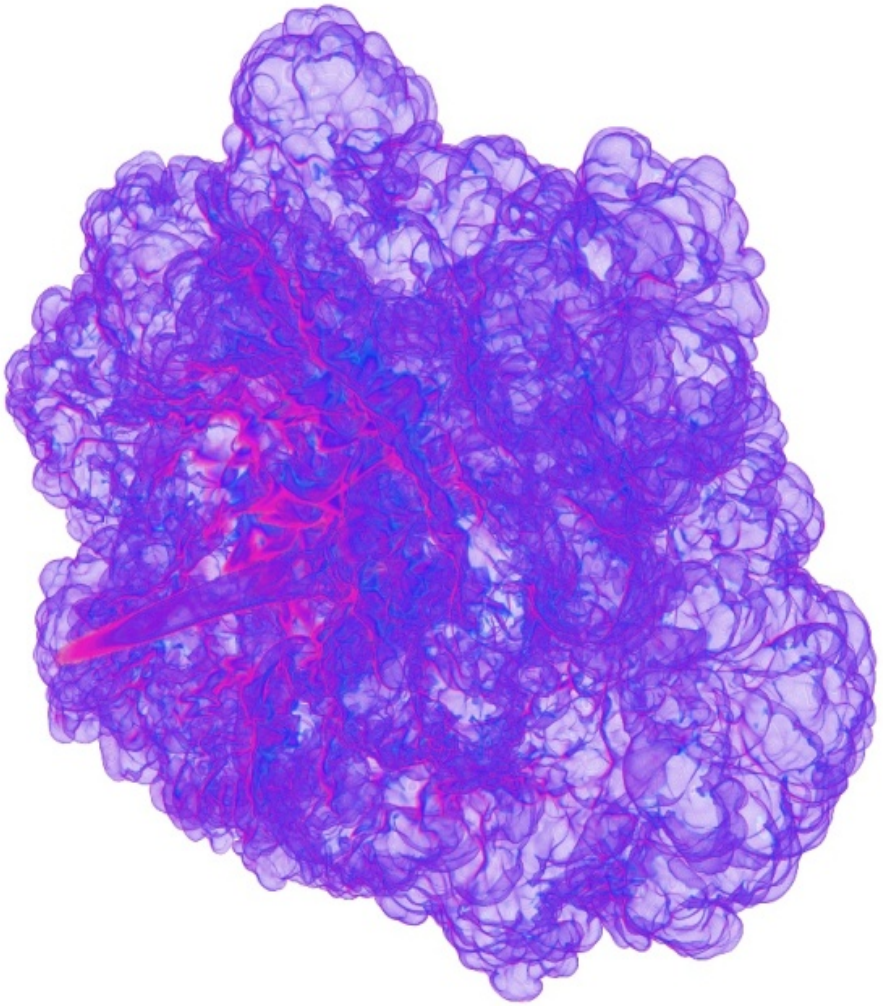
"We have benefited enormously from the support of NERSC staff"



Thermonuclear Supernovae INCITE Project

- Resolved problems with large I/O by switching to a 64-bit environment
- Tuned network connections and replaced scp with hsi : transfer rate went from 0.5 to 70 MB/sec
- Created automatic procedure for code checkpointing

"We have found NERSC staff extremely helpful in setting up the computational environment, conducting calculations, and also improving our software"

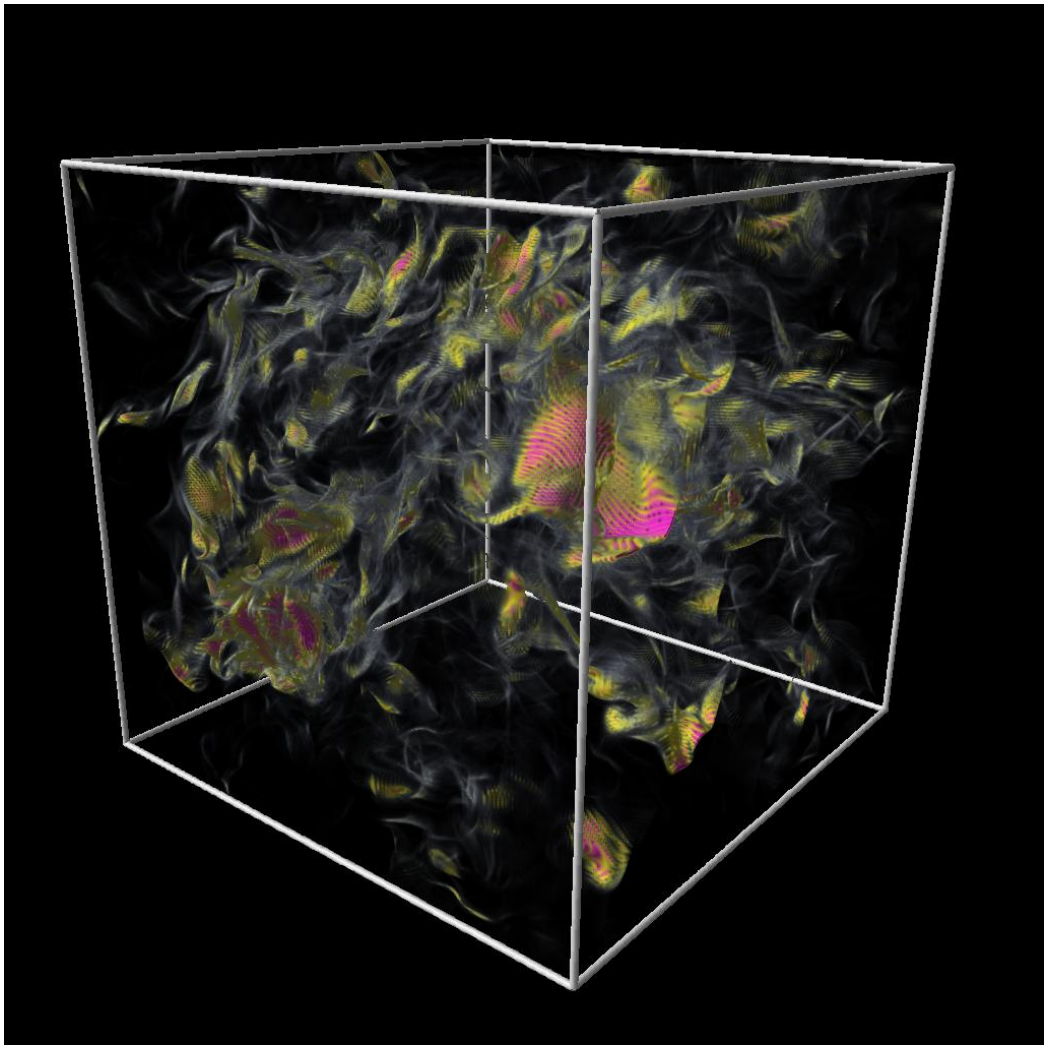


Fluid Turbulence INCITE Project

- **Reduced memory requirements and added threaded FFT: allowed group to solve larger and more interesting problems**
- **Visualization challenge: simulations produce large and feature-rich time-varying 3D data**
- **Vis solution: use Ensight parallel backend and Ensight client locally - collaboration resulted in deployment of Remote Vis License server**

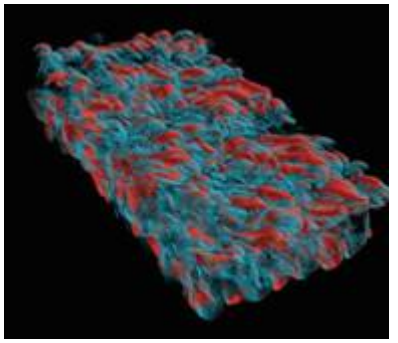
"We really appreciate the priority privilege that has been granted to us in job scheduling"

"The consultant services are wonderful. We have benefited from consultants' comments on code performance, innovative ideas for improvement, and diagnostic assistance"

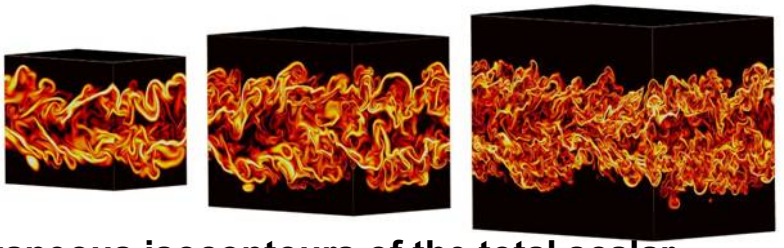


INCITE: Direct Numerical Simulation of Turbulent Non-premixed Combustion

- First direct 3D simulations of a turbulent nonpremixed H2/CO–air flame with detailed chemistry. The simulations, included 11 chemical species and 33 reactions.
- Project used 11.5M MPP hours
- Generated 10TB of raw DNS data that then was analyzed.
- Investigators - Jacqueline Chen, Evatt Hawkes, and Ramanan Sankaran of Sandia National Laboratories
- This project is now a primary user of the ORNL LCF



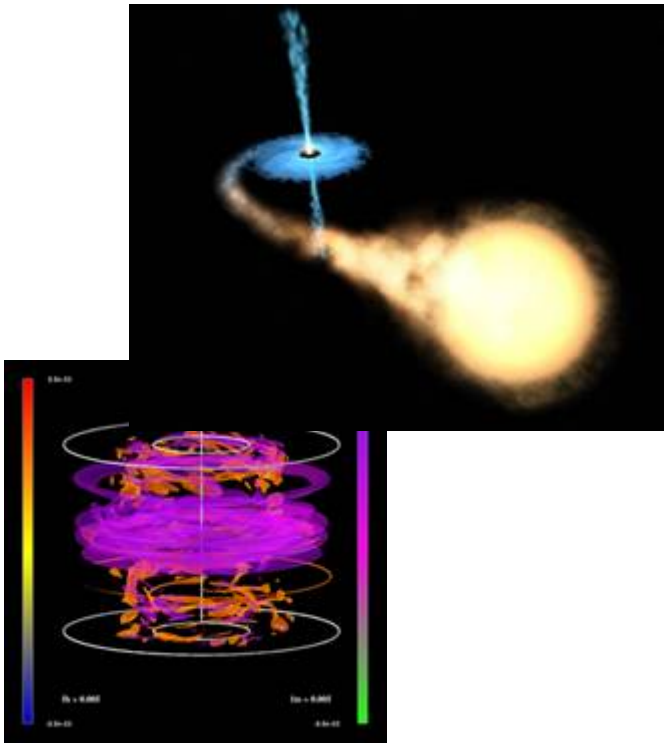
A simulated planar jet flame, colored by the rate of molecular mixing (scalar dissipation rate), which is critical for determining the interaction between reaction and diffusion in a flame.



Instantaneous isocontours of the total scalar dissipation rate field for successively higher Reynolds numbers at a time when re-ignition following extinction in the domain is significant.

INCITE: Magneto-rotational instability and turbulent angular momentum transport

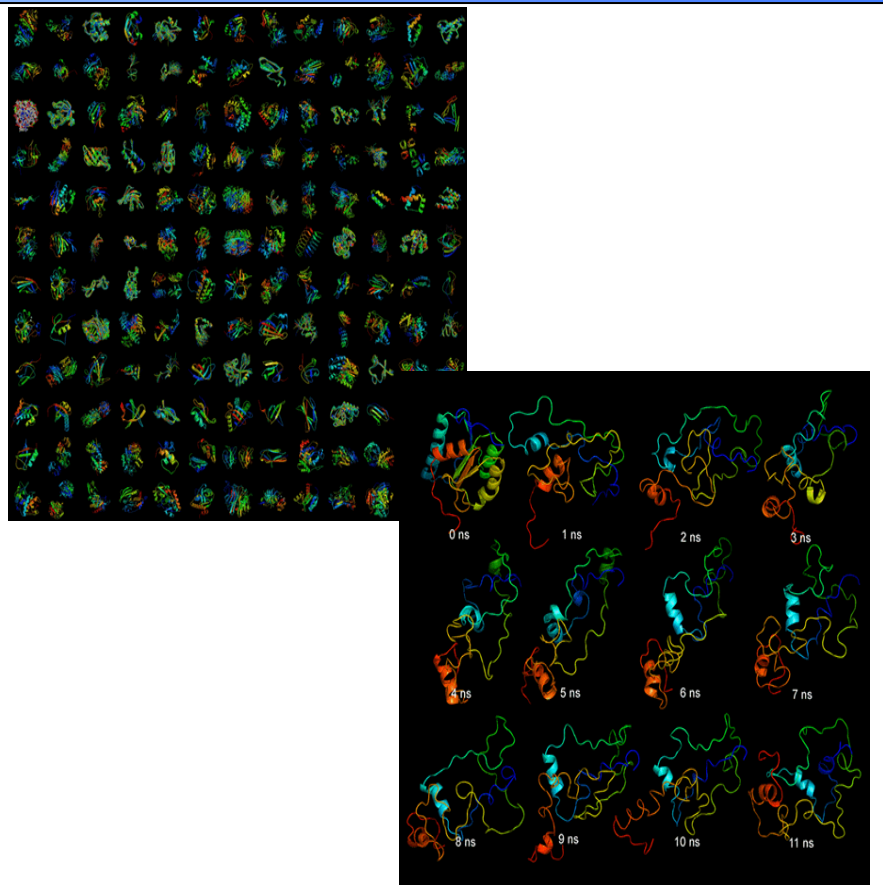
- Turbulent eddies provide a much more efficient mechanism for transporting angular momentum.
- Models of accretion disks that assume a reasonable amount of turbulence have produced credible accretion rates.
- Investigators - F. Cattaneo, P. Fischer, and A. Obabko



Visualization of the time evolution of the outward transport of angular momentum in a magnetic fluid bounded by rotating cylinders. The two colors correspond to the transport by hydrodynamic (orange) and hydromagnetic (purple) fluctuations.

INCITE: Molecular Dynameomics

- Awarded 2 million processor-hours.
- Combined molecular dynamics and proteomics to create an extensive repository of the molecular dynamics structures for protein folds, including the unfolding pathways.
- Approximately 1,130 known, non-redundant protein folds, of which her group has simulated about 30. predicting protein structure.
- Investigators – Valerie Daggart



Schematic representation of secondary structures taken at 1 ns intervals from a thermal unfolding simulation of inositol monophosphatase, an enzyme that may be the target for lithium therapy in the treatment of bipolar disorder.

Levee Analysis Project

- In 2006, of 800,000 MPP hours special allocations to the Army Corps of Engineers for studying ways to improve hurricane defenses along the Gulf Coast.
- As hurricanes move from the ocean toward land, the force of the storm causes the seawater to rise as it

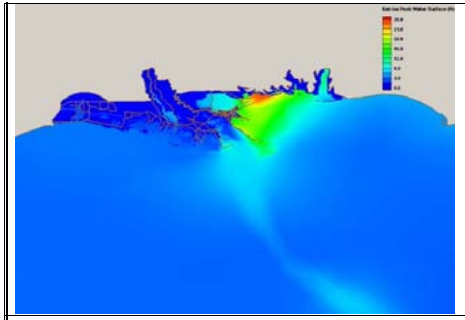


Figure 5. Overview simulation showing elevated storm surges along the Gulf Coast.

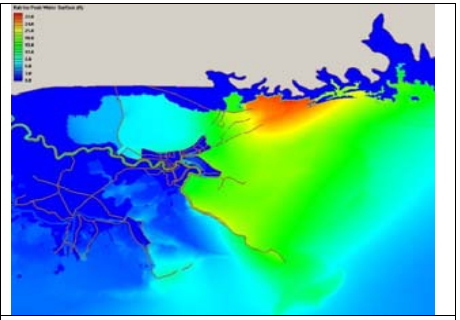


Figure 6. Simulation detail showing highest surge elevation (in red) striking Biloxi, Miss. New Orleans is the dark blue crescent to the lower left of Biloxi.

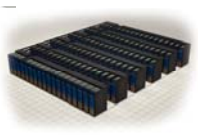
"Because these simulations could literally affect the lives of millions of Americans, we want to ensure that our colleagues in the Corps of Engineers have access to supercomputers which are up to the task,"

- Secretary Bodman, giving NERSC credit for its proven record of delivering highly reliable production supercomputing services.

How NERSC selected NERSC-5





PERCU - What Scientists Want from an HPC System

- **Performance**
  - How fast will a system process their work if everything is perfect
- **Effectiveness**
  - What is the likelihood they can get the system to do their work at the perfromance they expect
- **Reliability**
  - The system is available to do work and operates correctly all the time
- **Consistency/Variability**
  - How often will the system process their work as fast as it can
- **Usability**
  - How easy is it for scientists to get the system to go as fast as possible

PERCU

Best Value Source Selection (BVSS) – What Is It?

- Process developed at LLNL
  - Used and refined at LBNL on NERSC 3, NERSC 4, NCS, NCS-b and NERSC 5
  - Process adopted by other labs
- Intent is to reduce procurement time, reduce costs for technical evaluations, and provide efficient and cost effective way to conduct complex procurements
  - Used in competitive, negotiated contracting to select most advantageous offer
- Benefits
  - Flexible
    - Don't specify architecture
      - Can consider clusters, vector systems, others
    - Allows offerors to propose (and us to consider) different solutions from what we may have envisioned at the outset
  - Lets us evaluate and compare features in addition to price
    - Un-weighted and un-scored
    - Focuses on strengths and weaknesses of proposals
  - Provides more open communication with vendors
  - An art, not a science
    - Decision based on a rational analysis of competing proposals
- Requirements
  - ~53 total – all at high level
  - Minimum requirements
  - Performance features
  - Other items

Performance – Life cycle Purposes of Benchmarks

Benchmarks have four purposes

1. Evaluate systems (before selection or for general understanding)
2. Make sure the delivered system is what is expected
3. Make sure the system continues to operate as expected
4. Influence future systems by giving insight into architectural bottlenecks and into evolution of algorithms

1. Applications, limited kernels
2. Applications
3. Applications
4. Kernels, limited applications

Sustained System Performance, Potency and Value

$$SSP_s = \sum_{k=1}^{K_s} SSP_{s,k} = \sum_{k=1}^{K_s} \sum_{\alpha=1}^{A_{s,k}} \left(\Phi \left(W, P_{s,k,\alpha} \right) * N_{s,k,\alpha} \right)$$

$$Potency_s = \sum_{k=1}^{K_s} SSP_{s,k} * \left(\tau_{s,k+1} - \tau_{s,k} \right); \forall \tau_{s,k} \leq \tau_{\max}$$

$$Cost_s = \sum_{k=1}^{K_s} \sum_{l=1}^{L_s} c_{s,k,l}$$

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

Full description of this will be available soon in my dissertation from UC Berkeley

Composite Function Φ

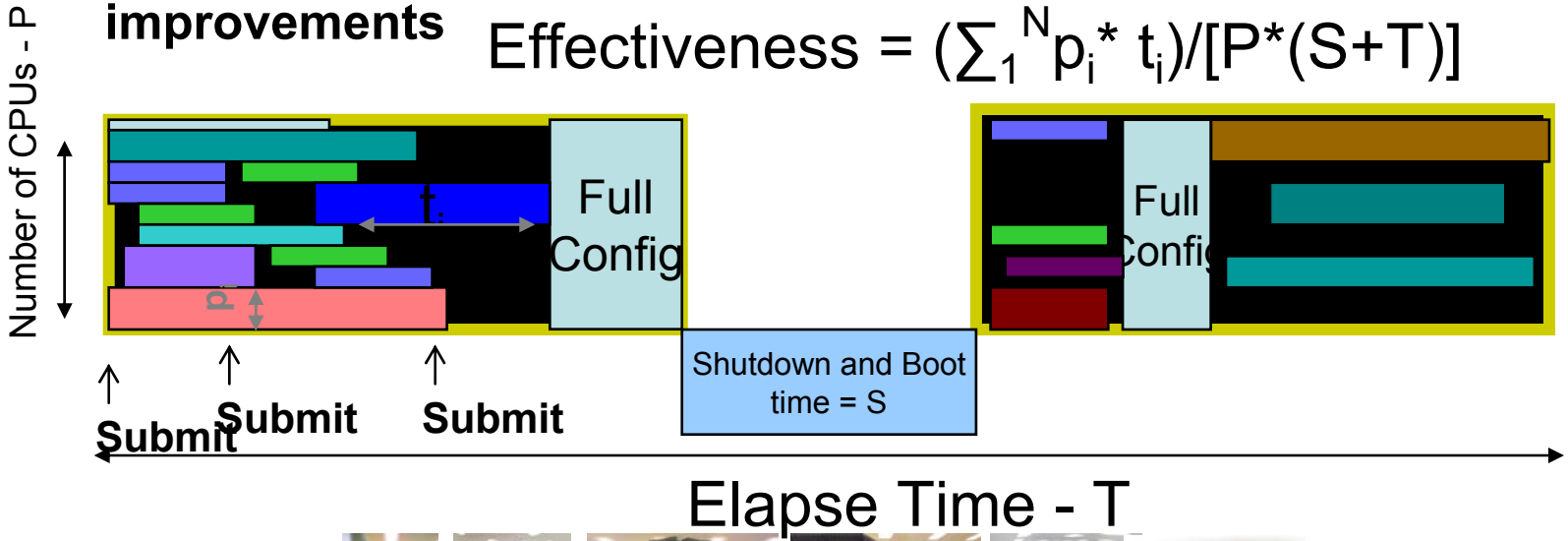
- **Examples of different composite functions on different systems using the NERSC-5 SSP**

	Seaborg (LBNL)	Bassi (LBNL)	Jacquard (LBNL)	Thunder Cluster (LLNL)
Computational Processors	6224	888	4096	640
Arithmetic SSP-4 (GFlops/s)	1,445	1,374	689	2,270
Geometric SSP-4 (GFlops/s)	902	835	471	1,637
Harmonic SSP-4 (GFlops/s)	579	570	318	1,183

Effective System Performance (ESP) Test

- Traditional methods of a throughput test do not address required features
- The ESP test measures
  - Both how much and how often the system can do scientific work
  - How well does a system get the right job to run at the right time
    - Needed for a Service Oriented Infrastructure
  - How easy can the system be managed
- Independent of hardware and compiler optimization improvements

$$\text{Effectiveness} = (\sum_1^N p_i * t_i)/[P*(S+T)]$$



Reliability

- Almost all metrics/requirements are reactive and after the decision
  - E.g. 99.999% availability
- The most common semi-proactive test is some type of availability test of a short time period
  - Run this code without interruption for 96 hours
  - Run this workload for 30 days with 96% availability
- Most people understand discrete hardware MTBF and MTTR and use that to decide hardware configurations

but

- Most major failures are software based – at least at NERSC
- Almost no wide ranging data on software reliability estimates or performance

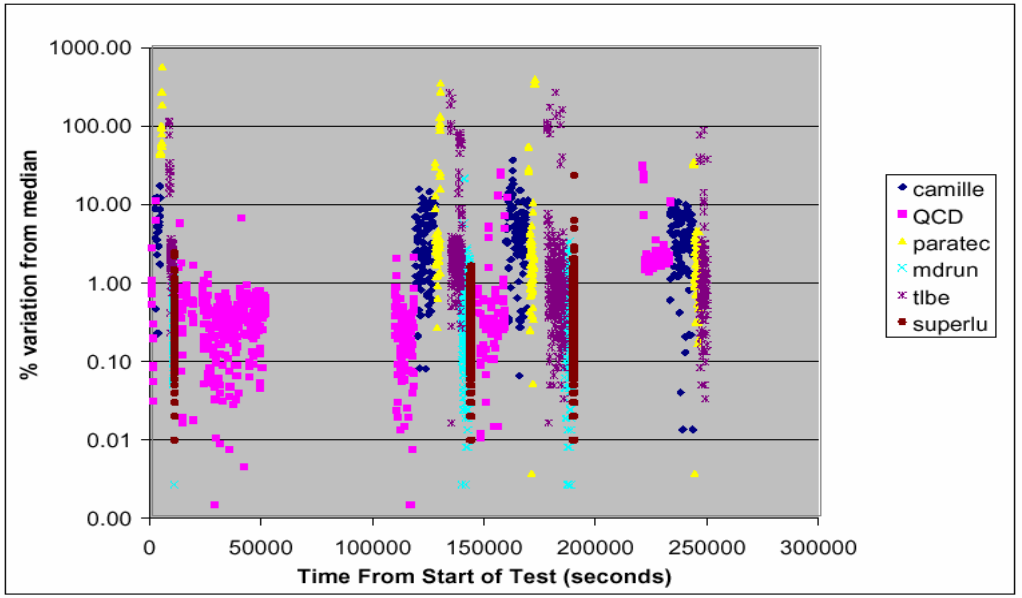
There should be as precise and complete understanding of software as there is for hardware

Reliability

- So the question is how to assess reliability proactively
- With the new world of horizontal integration, many reliability issues stem from component interaction and are not visible to any individual component provider.
- One modest attempt is to see how well providers understand the reliability of their components and then of the integration of the components
- There has been some work in reliability assessment for systems that have not been used for HPC
  - Injecting failure modes and assessing corrective reaction of systems
  - Probing for weak areas
  - Applying statistical learning theory/control theory to observe and then improve response
  - Most research is in discrete systems or Web oriented farms
    - E.g. Work at Rutgers [Richard Martin, Thu D. Ngyen, Kiran Nagaraja, et al] assess systems at a relatively high level, with the assumption that many low level faults are masked or handled by hardware or software before they impact applications.

Consistency

- Many examples of variability
- At NERSC, we have seen 10-20% more work coming from systems after consistency issues are address!
  - Loss of Cycles can be avoided
- Explicit variability metrics makes a difference
  - Coefficient of Variation on multiple benchmark runs, throughput tests, etc.
- Need large amounts of information to prove cause
  - One investigation took 9 months to determine the cause of a 10% performance difference between ½ the nodes in our system.
    - Solving it immediately generated the equivalent of a ½ TFlop/s more computing for users!



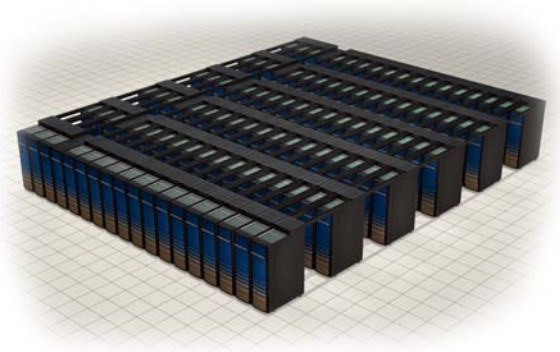
Usability

- What scientists really want to know is how much harder is it to use this system than they standard platform/tools
  - Most now use Linux desktops as their standard
- So, for HPC, we could conceive of a relative measure rather than an absolute measure
  - Relative to a scientist desktop – how much more effort is required to get X amount more work done on HPC systems than on their desktop?
  - Alternatively – is it worth learning how to use a much more sophisticated and efficient tool?
- How does "Productivity" relate to "Usability"?
- How to amortize the effort
  - First HPC conversion is high – others less so?
- How to craft a relative measures that are meaningful and discriminating

How to Use PERCU Measures

- Assess systems holistically
- Note I have not specified how a system is acquired.
  - PERCU simply points out what a system should do for it to be effective for users
- PERCU is a good way to address risk, particularly if there is a commitment to certain levels of performance by a provider
- PERCU also is relevant and explainable to the science community, and traceable to their requirements

NERSC-5








Original NERSC-5 Goals

- **Sustained System Performance over 3 years**
  - 7.5 to 10 Sustained Teraflop/s averaged over 3 years
- **System Balance**
  - Aggregate memory
    - Users have to be able to use at least 80% of the available memory for user code and data.
  - Global usable disk storage
    - At least 300 TB with an option for 150 TB more a year later
  - Can Integrate with the NERSC Global Filesystem (NGF)
- **Expected to significantly increase computational time for NERSC users in the 2007 Allocation Year**
  - January 9, 2007 – January 8, 2008
  - Have full impact for AY 2008

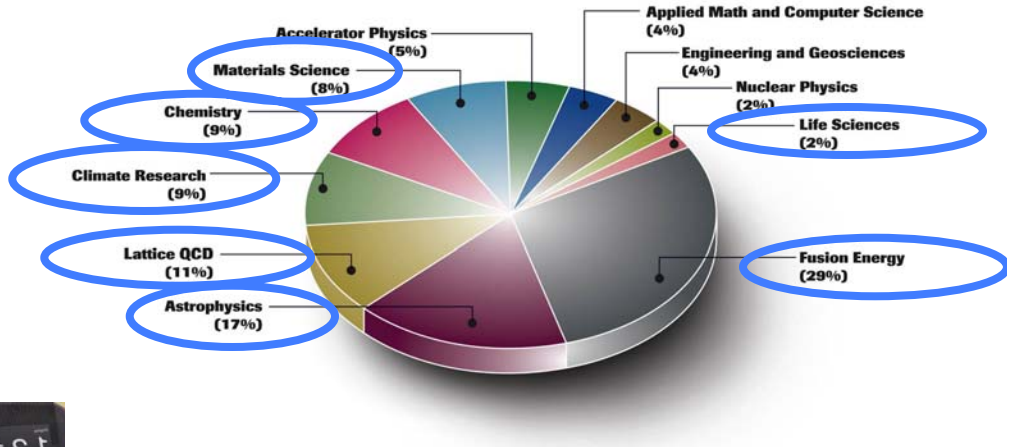
Application Benchmarks represent 85% of the Workload

Application	Science Area	Basic Algorithm	Language	Library Use
CAM3	Climate (BER)	CFD, FFT	FORTRAN 90	netCDF
GAMESS	Chemistry (BES)	DFT	FORTRAN 90	DDI, BLAS
GTC	Fusion (FES)	Particle-in-cell	FORTRAN 90	FFT(opt)
MADbench	Astrophysics (HEP & NP)	Power Spectrum Estimation	C	Scalapack
MILC	QCD (NP)	Conjugate gradient	C	none
PARATEC	Materials (BES)	3D FFT	FORTRAN 90	Scalapack
PMEMD	Life Science (BER)	Particle Mesh Ewald	FORTRAN 90	none

Micro benchmarks test specific system features - Processor, Memory, Interconnect, I/O, Networking

Composite Benchmarks

Sustained System Performance Test (SSP), Effective System Performance Test (ESP), Full Configuration Test, Throughput Test and Variability Tests



"Franklin"

Benjamin Franklin, America's First Scientist, performed ground breaking work in energy efficiency, electricity, materials, climate, ocean currents, transportation, health, medicine, acoustics and heat transfer.

Largest XT-4

9,740 nodes with 19,480 CPUs (cores)

102 Node Cabinets, 16 KWs per cabinet

39.5 TBs Aggregate Memory

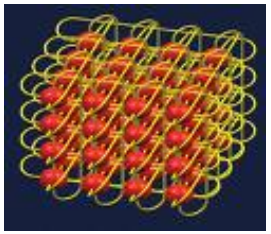
16.1+ Tflop/s Sustained System Performance

Seaborg - .9/Bassi - .8

Cray SeaStar2/3D Torus Interconnect (17x24x24)

6.3 TB/s Bi-Section Bandwidth

7.6 GB/s peak bi-directional bandwidth per link



345 TBs of Usable Shared Disk

Sixty 4 Gbps Fibre Channel Data Connections

Four 10 Gbps Ethernet Network Connections

Sixteen 1 Gbps Ethernet Network Connections

NERSC/Cray Center of Excellence for System Management and Storage

- **Cray Center of Excellence**
  - Joint Cray and NERSC managed activity
- **Initial Projects**
  - Integrate Berkeley Laboratory Checkpoint Restart (BLCR) with Portal and Computer Node Linux
    - BLCR is a research product of SciDAC activities
  - Petascale I/O Interface for compute nodes
    - IO Forwarding to increase integration potential for XT systems
- **Future projects will be jointly defined**
- **COE also involved with NERSC's SDSA efforts to perform a detailed analysis of dual and quad core systems.**
  - Helen He will talk about this study on Tuesday

Probably Software Configuration

- SuSE SLES 9.0 or 10.0 Linux on Service Nodes
- Compute Node Linux O/S for all compute nodes
  - Cray's light weight Linux kernel
- Portals communication layer
  - MPI, Shmem
- Compute node integration with the NERSC Global Filesystem
  - Global file systems (e.g. GPFS, Lustre, others) directly accessible from compute nodes with a "Petascale I/O Interface"
- Torque with Moab
  - Most expected functions including Backfill, Fairshare, advanced reservation
- Checkpoint Restart
  - Based on Berkeley Lab Checkpoint/Restart (Hargrove)
- Application Development Environment
  - PGI compilers - assembler, Fortran, C, UPC, and C++
  - Parallel programming models include MPI, and SHMEM.
  - Libraries include SCALAPACK, SuperLU, ACML, Portals, MPICH2/ROMIO.
  - Languages and parallel programming models shall be extended to include OpenMP, and Posix threads but are dependent on compute node Linux
  - Totalview or equivalent to 1,024 tasks
  - Craypat and Cray Apprentice
  - PAPI and Modules

NERSC Expectations for Franklin

Area	Final
SSP	16.09 TF
ESP	78.8%
Variation	Dedicated - 3% CVN / 4% CNL Production - 5% CNL or CVN
Streams	7,824 MB/s – 60% memory/node 3,552 MB/s- Full Node
Ping Pong	5 – 6.9 µs (best and worst case)
Full Configuration Test	22-30 seconds
I/O	> 12 GB/s aggregate
CPU Resources Used by O/S	< 1%
Memory Used by O/S	225 MB CVN/ 400 MB CNL
Availability	98%
System-wide MBTF	14 days
Job Completion	95% for jobs < 100,000 node hours (about 4 days for a 1,024 way job)
Cray Center of Excellence at NERSC for Storage and Resource Management at NERSC	2 FTEs

The Phasing of NERSC-5

- **Small Test System**
  - **Summer 2006 – small 52 (44 compute) node XT3**
  - **Fall 2006 – upgrade to XT4**
- **January 2007 - Phase 1**
  - **36 racks**
  - **All I/O and Service Nodes**
  - **Most of the disk – 330 TB**
  - **6 x 24 x 24 Torus**
- **February 2007 – Phase 2**
  - **66 more compute rack**
  - **More disks and controller – 402 TB total**
    - **71 TB and one controller move to NGF after Phase 2 acceptance**
  - **17 x 24 x 24 Torus**
  - **See Nick Cardo's Presentation later in the conference**
- **Winter 2007/2008 – option to upgrade to quad core opteron – 4 x peak performance increase**
  - **Likely only a 2x measured performance increase**
  - **Double memory per node to keep the constant B/F ratio**
  - **See Helen He's Presentation**
- **Spring to Summer 2008 – Major software upgrade**
- **Winter/Spring 2009 – *option for a 1 Petaflop/s system***

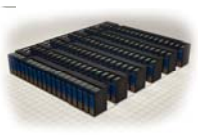
Current Status of NERSC-5

- Fielding very large, early systems is very challenging
  - Example - "Petascale Systems Integration Workshop"
    - May 15-16 in San Francisco
- Problems have been identified, diagnosed and corrected
  - Hardware
  - Software
- Testing is progressing about as expected
  - Most things are working as we expected
  - Issues identified when workload scales
  - Most are complex and subtle interactions
- Application Benchmark performance is encouraging
- Cray doing an excellent job providing the expertise and resources needed to make timely progress
- Currently expanding the workload diversity and scale in an organized manner

NERSC Futures



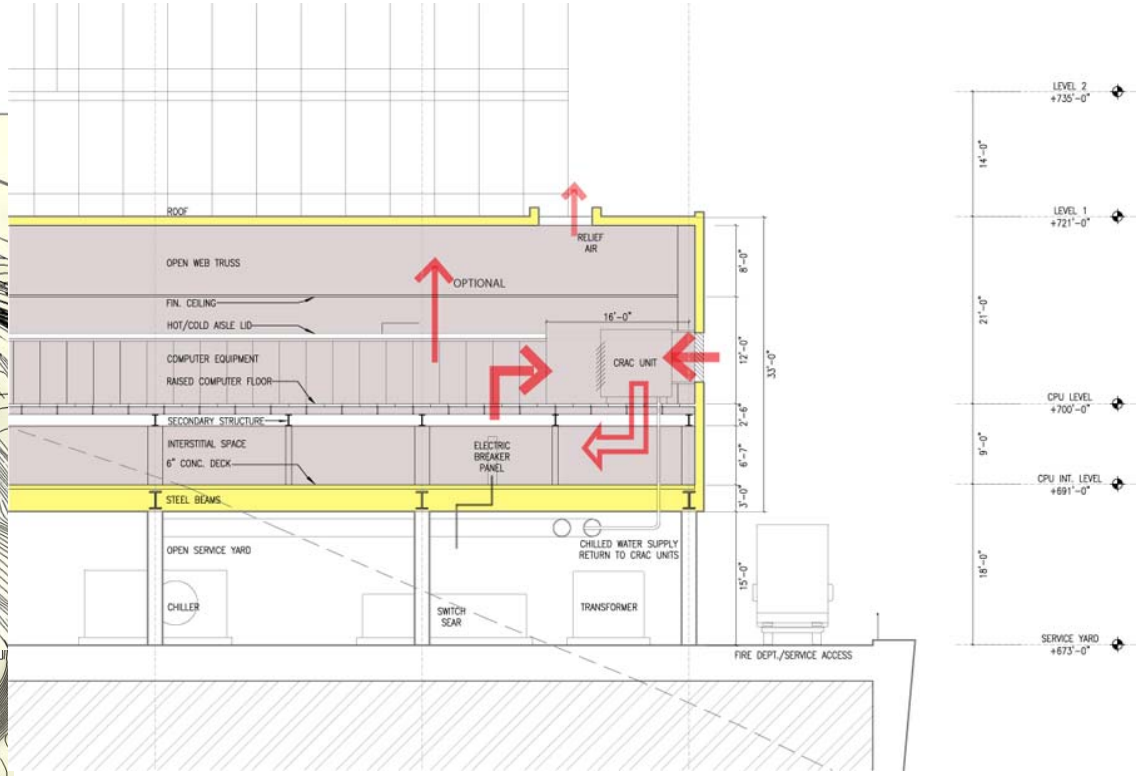


NERSC Futures

- **Continue to expand impact on DOE Science**
  - **Assist increasing scaling – particularly for those harder to scale area**
  - **Expand support for Data Oriented and Analytical computing**
- **NERSC – 6 – 2009-2010**
  - **Significant Increase in Computation**
- **New Computing Facility – 2010-2012**

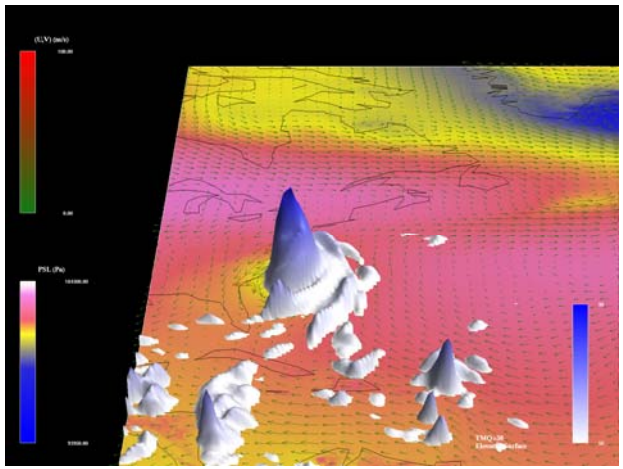
LBNL CRT Building

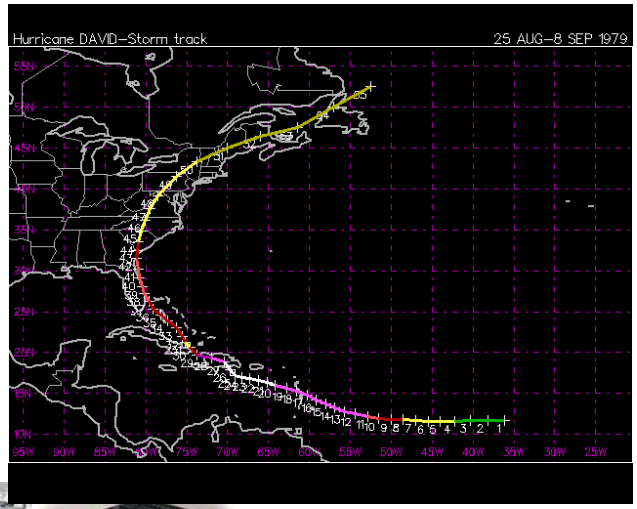


Office of Science
U.S. DEPARTMENT OF ENERGY

Comparing Real and Simulated Storm Data

- Michael Wehner (LBNL)
- The effect of climate change on the intensity and frequency of hurricanes in area is of utmost importance to policymakers.
- A workflow enabling fast qualitative comparisons between simulated storm data and real observations

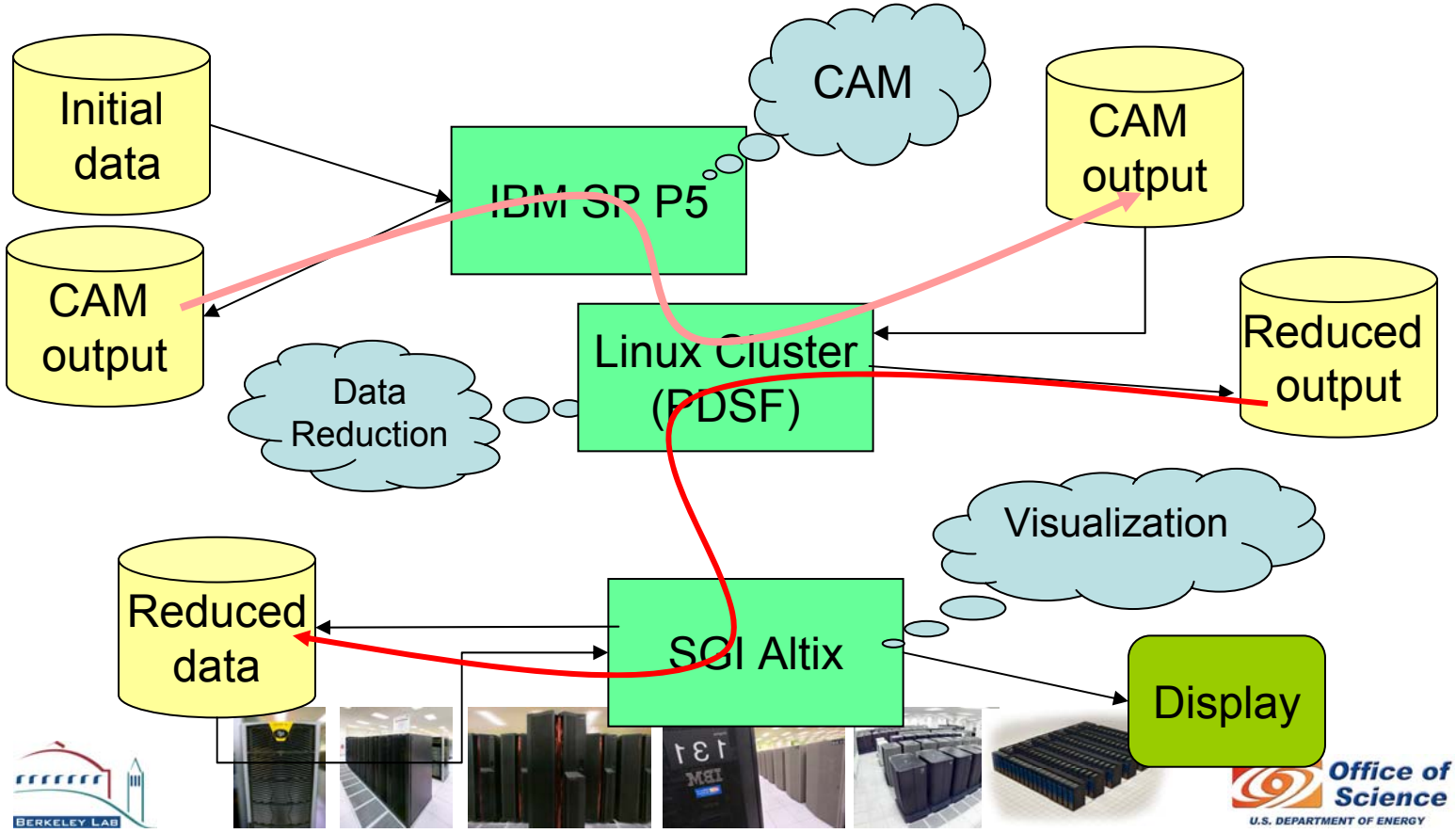




Office of Science
U.S. DEPARTMENT OF ENERGY

Comparing Real and Simulated Storm Data – The Old Way

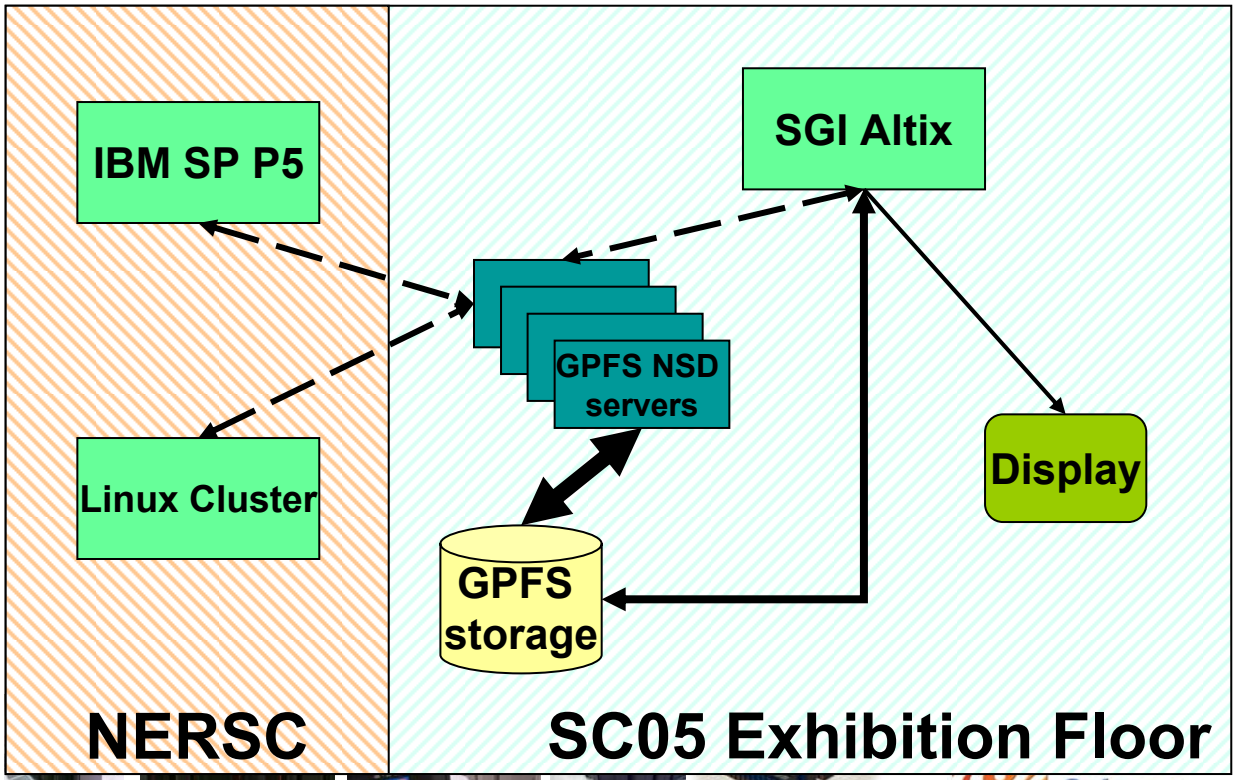
Data is reduced/pre-processed on commodity cluster and transferred to and processed by SMP system for visualization



The New Way

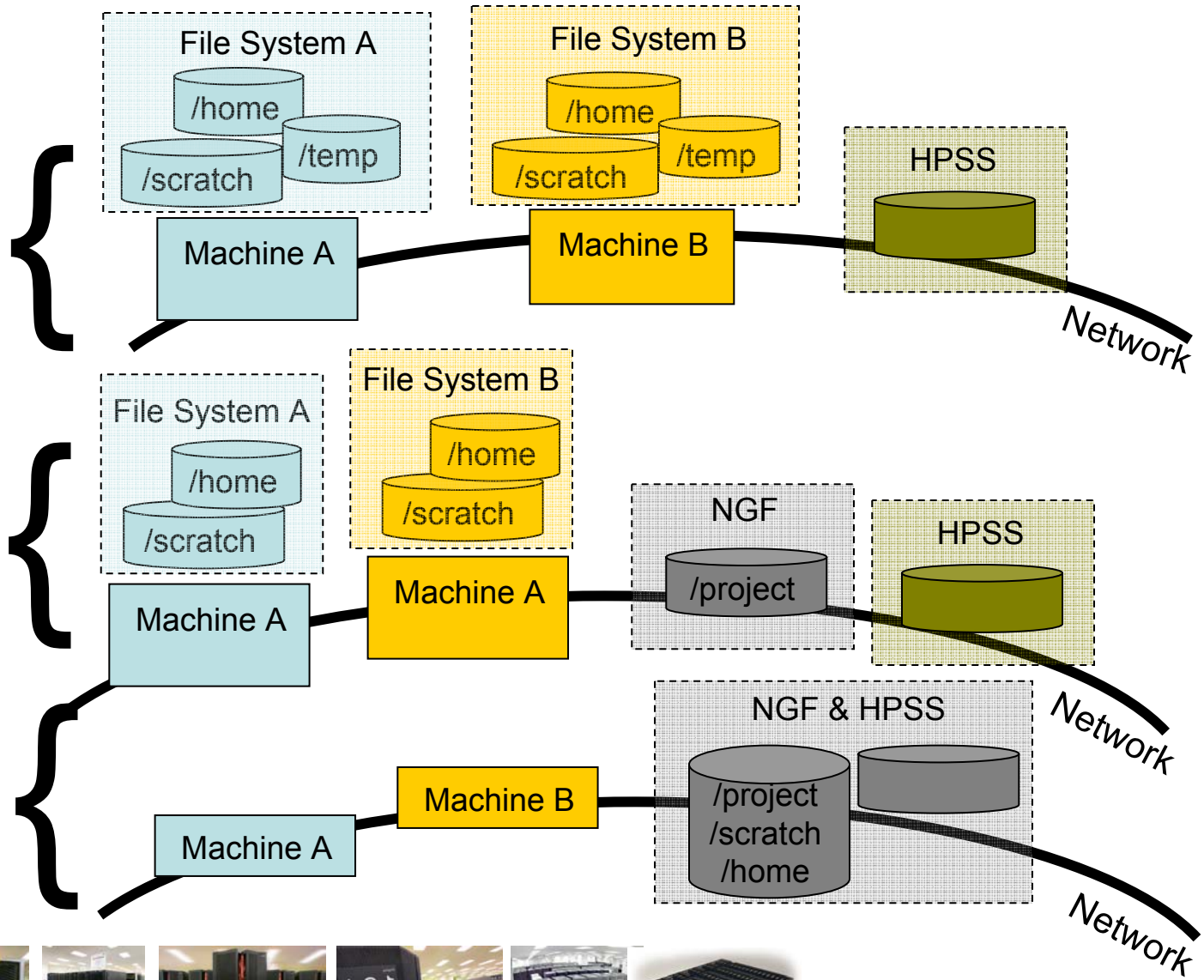
- Entered prototype in SC05 StorCloud Challenge
- Separate computational resources coupled via WAN-GPFS
- Winner: Best Deployment of a Prototype for a Scientific Application

William P. Baird, Wes Bethel, Jonathan Carter, Cristina Siegerist, Tavia Stone, and Michael Wehner



NERSC Storage Roadmap

- **Past**
  - **Local Disk**
    - **/scratch**
    - **/home**
    - **/tmp**
  - **HPSS**
- **Now**
  - **Local Disk**
    - **/scratch**
    - **/home**
  - **NGF**
    - **/project**
  - **HPSS**
- **Future**
  - **Local Disk**
    - **/home**
  - **NGF-HPSS**
    - **/scratch**
    - **/project**



Summary

- NERSC continues to enable outstanding computational science through
  - a highly reliable, efficient, integrated production environment
  - provision of the whole spectrum of resources (computers, storage, networking)
- NERSC 5 promises to be a significant increase in production capability
- NERSC taking bold steps for the furture

The Real Result of NERSC's Science-Driven Strategy

SCIENCE-DRIVEN SYSTEMS

Each year on their allocation renewal form, PIs indicate how many refereed publications their project had in the previous 12 months.

Year of request renewal	Number of refereed publications
2007	1,437
2006	1,448
2005	1,270

SCIENCE-DRIVEN SERVICES

SCIENCE-DRIVEN ANALYTICS

Some References

- The Landscape of Parallel Computing Research: A View from Berkeley
  - http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html
- Science-Driven Computing: NERSC's Plan for 2006-2010
  - https://www.nersc.gov/news/reports/LBNL-57582.pdf
- How Are We Doing? A Self-Assessment of the Quality of Services and Systems at NERSC, 2005-2006
  - https://www.nersc.gov/news/reports/LBNL-62117.pdf
- Software Roadmap to Plug and Play Petaflop/s
  - https://www.nersc.gov/news/reports/LBNL-59999.pdf
- National Facility for Advanced Computational Science: A Sustainable Path to Scientific Discovery
  - https://www.nersc.gov/news/reports/PUB-5500.pdf
- Creating Science-Driven Architecture: A New Path to Scientific Leadership
  - https://www.nersc.gov/news/reports/ArchDevProposal.5.01.pdf
- Parallel Scaling Characteristics of Selected NERSC User Project Codes.
  - http://www-library.lbl.gov/docs/PUB/904/PDF/PUB-904_2006.pdf
- ESP: A System Utilization Benchmark
  - https://www.nersc.gov/news/reports/espsc00.pdf
- The NERSC Sustained System Performance (SSP) Metric.
  - http://www-library.lbl.gov/docs/LBNL/588/68/PDF/LBNL-58868.pdf






Backup

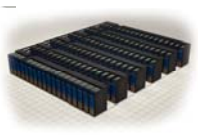




NERSC 5 Requirements





NERSC-5 Minimum Requirements

- General
  - A complete, integrated system for a multi-user, multi-application parallel scientific workload
  - System shall not exceed 3200 gross square feet of floor space and consume no more than 2.5 MW of electrical power
- Performance
  - A proposal for and a commitment to deliver application performance as measured by the Sustained System Performance (SSP) metric
  - A high-performance interconnect with scalable performance characteristics over the entire system
  - 10 Gigabit Ethernet connectivity to NERSC infrastructure
- Effectiveness
  - A filesystem accessible system-wide via a single, unified namespace
  - A scalable, robust, effective and comprehensive system administration, and resource management environment
  - An application development environment consisting of at least: standards compliant Fortran, C, and C++ compilers, and an MPI library
  - Ability to effectively manage system resources with high utilization and throughput under a workload with a wide range of concurrencies

Note – Blue items are discussed

NERSC-5 Minimum Requirements
(Continued)

- **Reliability**
  - Comprehensive maintenance and 24x7 support for all hardware and software components
  - Demonstrated ability to produce and maintain the proposed system
- **Variability**
  - Consistent and reproducible execution times in dedicated and production mode
  - The Offeror shall document the amount of run time variation the system shall have, both in dedicated and general user modes
- **Usability**
  - Correct, consistent and reproducible computation results
  - Compliance with 32- and 64-bit IEEE 754 floating point arithmetic
- **Facility Wide File System**
  - The system shall be integrated with NERSC's GPFS based Facility Wide File System system
  - All system shared storage and storage fabric shall be standards based and packaged independently. Acceptable standards are Fiber Channel, Ethernet, and Infiniband

NERSC-5 Performance Features

- **General**
  - Low power, cooling, and floor space
  - Ease of seismic bracing
  - Credible roadmap for future hardware and software products
- **Performance**
  - Documented performance characteristics and benchmark results
  - High bandwidth and low latency interconnect
  - Large amount of aggregate user addressable memory
  - Ability to use a large amount of memory by a serial or multithreaded program, containing no explicit calls to an API enabling distributed-memory access (e.g. MPI, shmem, LAPI), on a portion of the machine
  - Sustained I/O bandwidth to global shared disk storage
  - 300 TB of formatted disk space with initial delivery, with an option for 150 TB of additional formatted disk disk after 1 year
  - High sustained aggregate external network bandwidth
- **Effectiveness**
  - Ability to run a single application instance over all the compute nodes in the system
  - Minimal intrusion upon memory available to application data structures by system libraries, daemons, operating system and/or kernel
  - High performance MPI collective operations and support for overlapped computation and communication activity

NERSC-5 Performance Features
(Continued)

- **Effectiveness (cont)**
  - Parallel file system capable of being accessed at high performance both from within the system, and from other NERSC systems
  - Advanced resource management functionality; e.g. checkpoint-restart, job migration, backfill, gang scheduling, advanced reservation and job preemption
  - The system shall be partitionable - at least in half - through either logical or physical features so that the partitions operate as independent systems. All functionality shall exist and shall properly operate in an identical manner whether the system is partitioned or not. Performance for codes with concurrency less than the partition size shall be no less than in the full system configuration. The offeror shall describe how the system is partitioned and indicate how long it takes to partition and to rejoin the system, as well as any extra costs
- **Reliability**
  - Commitment to achieving specific quality assurance, reliability and availability goals
  - A clear plan documenting how the vendor will effectively respond to software defects and system outages at each severity level, and how a problem or defect will be escalated if not fixed in a timely manner
  - Provide information concerning the number of defects filed at each severity level and average time to problem resolution for all major software and hardware compnents
  - An effective methodology for system upgrades, repairs and testing. Provide a description of how it addresses issues of system availability and user productivity

NERSC-5 Performance Features
(Continued)

- **Variability**
  - Minimal intrusion on CPU resources available to application processes by system libraries, daemons, operating system and/or kernel
  - Minimal intrinsic architectural barriers to application scaling such as system jitter or synchronization mismatches across nodes boundaries
- **Usability**
  - Native 64-bit support within libraries, compilers and the operating system
  - User access to performance counters on the processor, storage subsystems and interconnect via a documented API
  - Support for centralized configuration management/change management
  - Capability for remote administration including hardware reset, power management, booting, and remote console
  - Fully featured application development environment, including: vendor optimized serial and parallel scientific libraries (e.g LAPACK, BLAS); MPMD MPI; GNU tools and utilities; a parallel debugger such as Totalview; performance profiling and tuning tools
  - Standards compliant MPI-1, MPI-2 and OpenMP (if appropriate)
  - Accounting and activity tracking functionality, e.g., job containers, which assist in job, session and unix process tracking for security and resource management purposes
  - Support for global addressing, e.g. CoArray FORTRAN and UPC, and remote data access with put/get semantics
  - Online documentation of all system software and hardware available to NERSC staff

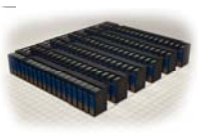
NERSC-5 Performance Features
(continued)

- **Usability**
  - Online documentation of all user visible system features available to all NERSC users (OS/Scheduler user interfaces, filesystems, libraries, programming environments, debugging and performance monitoring/profiling tools.)
  - Training for NERSC System management and user support staff.
  - Details of how the proposed system architecture will enhance latency tolerance to non-unit stride memory accesses
  - Ability to integrate with grid environments running current software implementations, for example, Globus Toolkit 2.6, OGSA 4.0
- **Facility Wide File System**
  - Offeror shall provide a plan for integrating, supporting and achieving high performance parallel access to the GPFS based Facility Wide File System system
  - All storage support nodes shall be capable of being reconfigured into computational nodes
  - Offeror shall provide engineering assistance with the re-allocation of storage hardware from the NERSC 5 system to the GPFS-based Facility Wide File System system
  - Maintenance and required licenses shall continue on the storage and storage fabric after being connected to the GPFS based Facility Wide File System system

Kernel Benchmarks





Kernel Benchmarks

- Test specific system features
  - Processor
  - Memory
  - Interconnect
  - I/O
- Support our performance modeling activities
  - 3 Packages (Memory, Interconnect, I/O)

Kernel Benchmarks

- **Processor: NAS Parallel Benchmarks (NPB)**
  - **Serial: NPB 2.3 Class B**
    - **best understood code base**
  - **Parallel: NPB 2.4 Class D at 64-256 processors**
    - **Class D is not available with 2.3**
- **Memory**
  - **Streams**
  - **APEX-Map – serial**
    - **For a more thorough characterization of system**

Kernel Benchmarks (cont.)

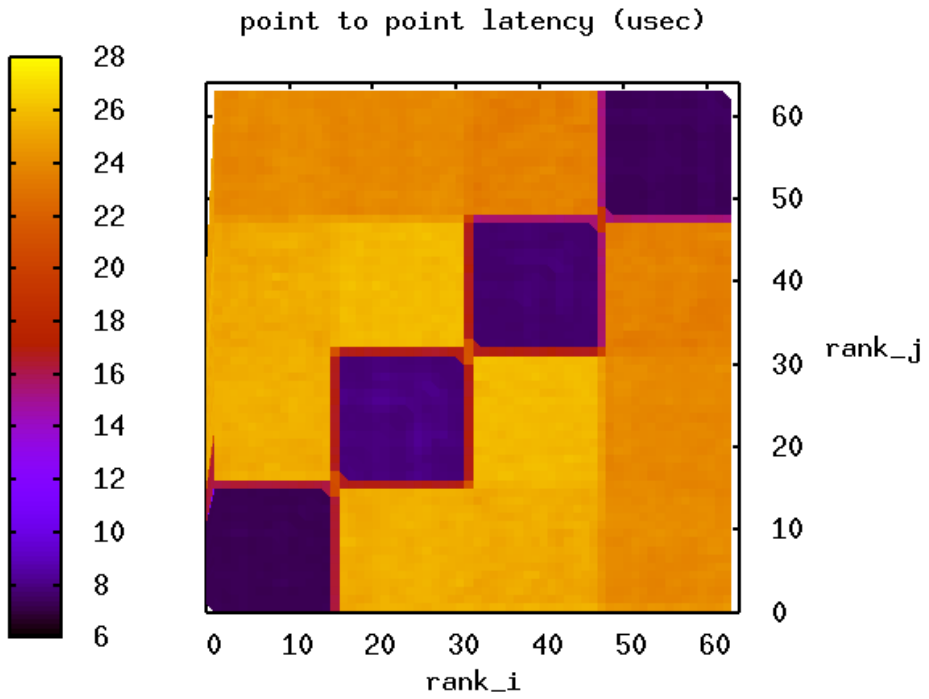
- **Interconnect**
  - **MultiPong**
    - **Maps out switch topology latency and bandwidth**
  - **APEX-Map parallel**
    - **Random message exchanges**
- **Network performance**
  - **netperf benchmark**

Interconnect Testing: MultiPong

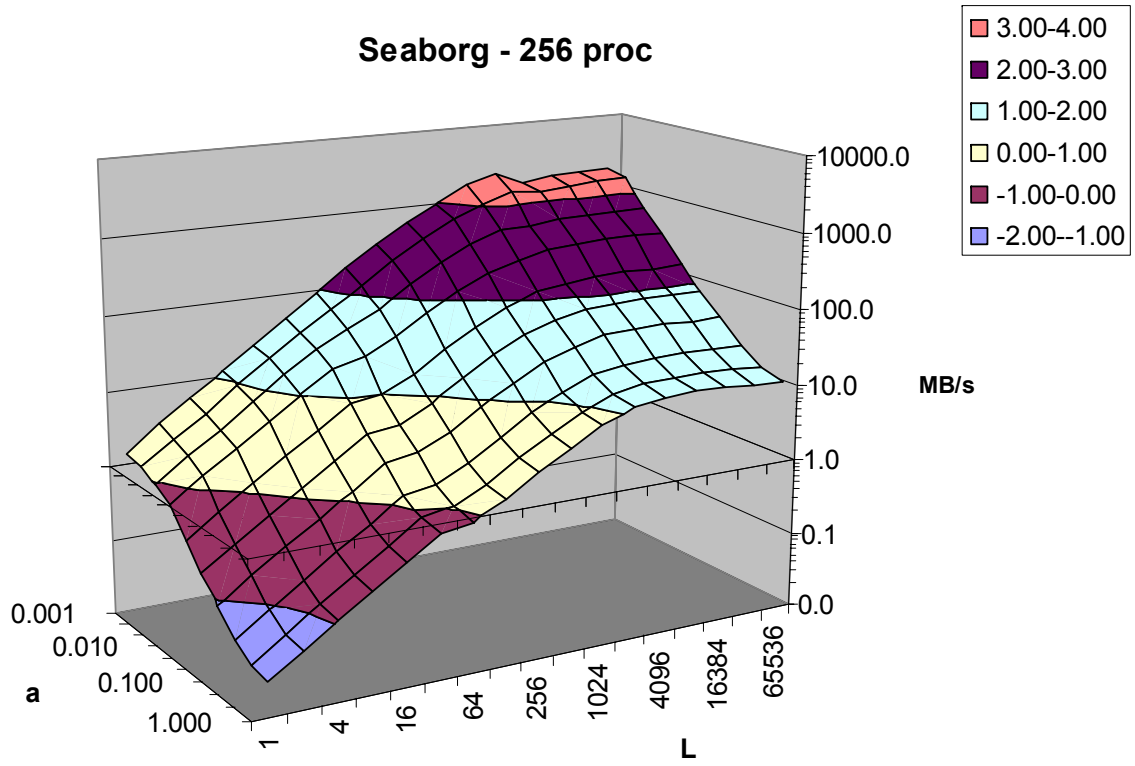
Switch performance is more complex than a single latency + bandwidth

MultiPong maps the interconnect's hierarchy of connections

More detailed understanding of communication topology



Parallel APEX-Map



L measures spatial locality (vector length)
a measures temporal locality, related to the probability we will jump in memory

Application Benchmarks

Application Benchmarks

- **Selection of benchmarks take several considerations**
  - **Representative of the workload**
  - **Represent different algorithms and methods**
  - **Are portable to likely candidate architectures with limited effort**
  - **Work in a repeatable and testable manner**
  - **Are tractable for a non-expert to understand**
  - **Can be instrumented**
  - **Ability to distribute**
- **Started with approximately 20 candidates**

Application Benchmarks

- CAM3
  - Climate model, NCAR
- GAMESS
  - Computational chemistry, Iowa State, Ames Lab
- GTC
  - Fusion, PPPL
- MADbench
  - Astrophysics (CMB analysis), LBNL
- Milc
  - QCD, multi-site collaboration
- Paratec
  - Materials science, developed LBNL and UC Berkeley
- PMEMD
  - Computational chemistry, University of North Carolina-Chapel Hill

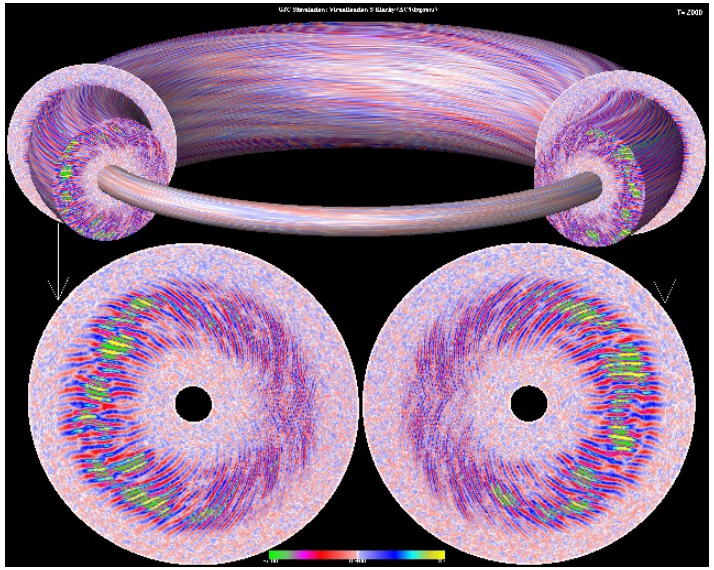
CAM3

- Community Atmospheric Model version 3
  - Developed at NCAR with substantial DOE input, both scientific and software
- The atmosphere model for CCSM, the coupled climate system model
  - Also the most time-consuming part of CCSM
  - Widely used by both American and foreign scientists for climate research
    - For example, Carbon, bio-geochemistry models are built upon (integrated with) CAM3
    - IPCC predictions use CAM3 (in part)
  - About 230,000 lines codes in Fortran 90
- 1D Decomposition, runs up to 128 processors at T85 resolution (150Km)
- 2D Decomposition, runs up to 1680 processors at 0.5 deg (60Km) resolution

GAMESS

- Computational chemistry application
  - Variety of electronic structure algorithms available
- About 550,000 lines of Fortran 90
- Communication layer makes use of highly optimized vendor libraries
- Many methods available within the code
  - Benchmarks are DFT energy and gradient calculation, MP2 energy and gradient calculation
  - Many computational chemistry studies rely on these techniques
- Exactly the same as DOD HPCMP TI-06 GAMESS benchmark
  - Vendors will only have to do the work once

GTC



3D visualization of electrostatic potential in magnetic fusion device

- **Gyrokinetic Toroidal Code**
- **Important code for Fusion SciDAC project and for ITER, the international fusion collaboration**
- **Transport of thermal energy via plasma microturbulence using particle-in-cell approach (PIC)**

MADbench

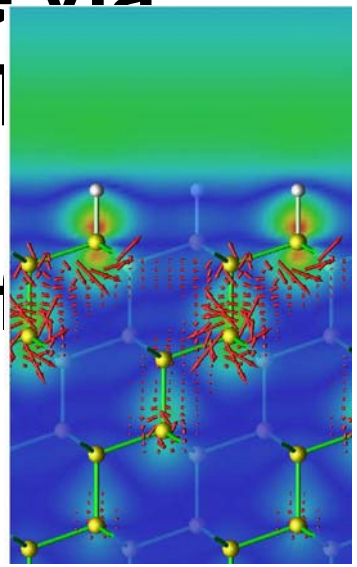
- Cosmic microwave background radiation analysis tool (MADCAP)
  - Used large amount of time in FY04 and one of the highest-scaling codes at NERSC
- MADBench is a benchmark version of the original code
  - Designed to be easily run with synthetic data for portability.
  - Used in a recent study in conjunction with Berkeley Institute for Performance Studies (BIPS).
- Written in C making extensive use of ScaLAPACK libraries
- Has extensive I/O requirements

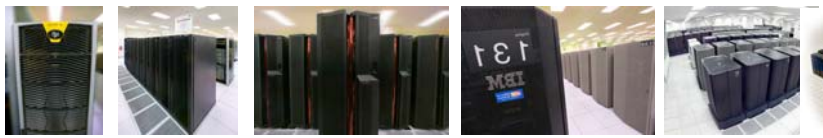
MILC

- Quantum ChromoDynamics application
  - Widespread community use, large allocation
  - Easy to build, no dependencies, standards conforming
  - Can be setup to run on wide-range of concurrency
- Conjugate gradient algorithm
- Physics on a 4D lattice
- Local computations are 3 x 3 complex matrix multiplies, with sparse (indirect) access pattern

PARATEC

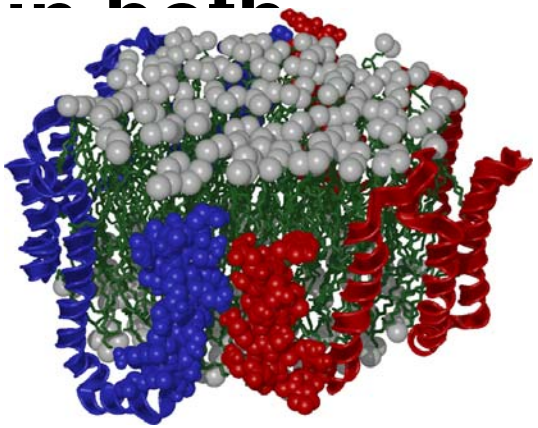
- Parallel Total Energy Code
- Plane Wave DFT using custom 3D FFT
- 70% of Materials Science computation at NERSC is done via Plane Wave DFT codes. PARAT captures the performance of a wide range of codes (VASP, CPMD, PETOT)





PMEMD

- Particle Mesh Ewald Molecular Dynamics
  - An F90 code with advanced MPI coding should test compiler and stress asynchronous point to point messaging
- PMEMD is very similar to the MD Engine in AMBER 8.0 used in both chemistry and biosciences
- Test system is a 91K atom blood coagulation protein



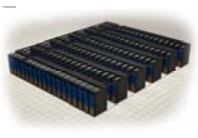
Application Summary

- **Benchmark deliverables**
  - **Timings at medium (64 processors, 54 for CAM3) and large (256 processors for most, 384 for GAMESS, 240 for CAM3)**
  - **Projections for extra large tests (1024 MADbench and 2048 MILC)**
  - **Variation in runtime as measured by coefficient of variation**

Composite Benchmarks and Metrics





Composite Benchmarks

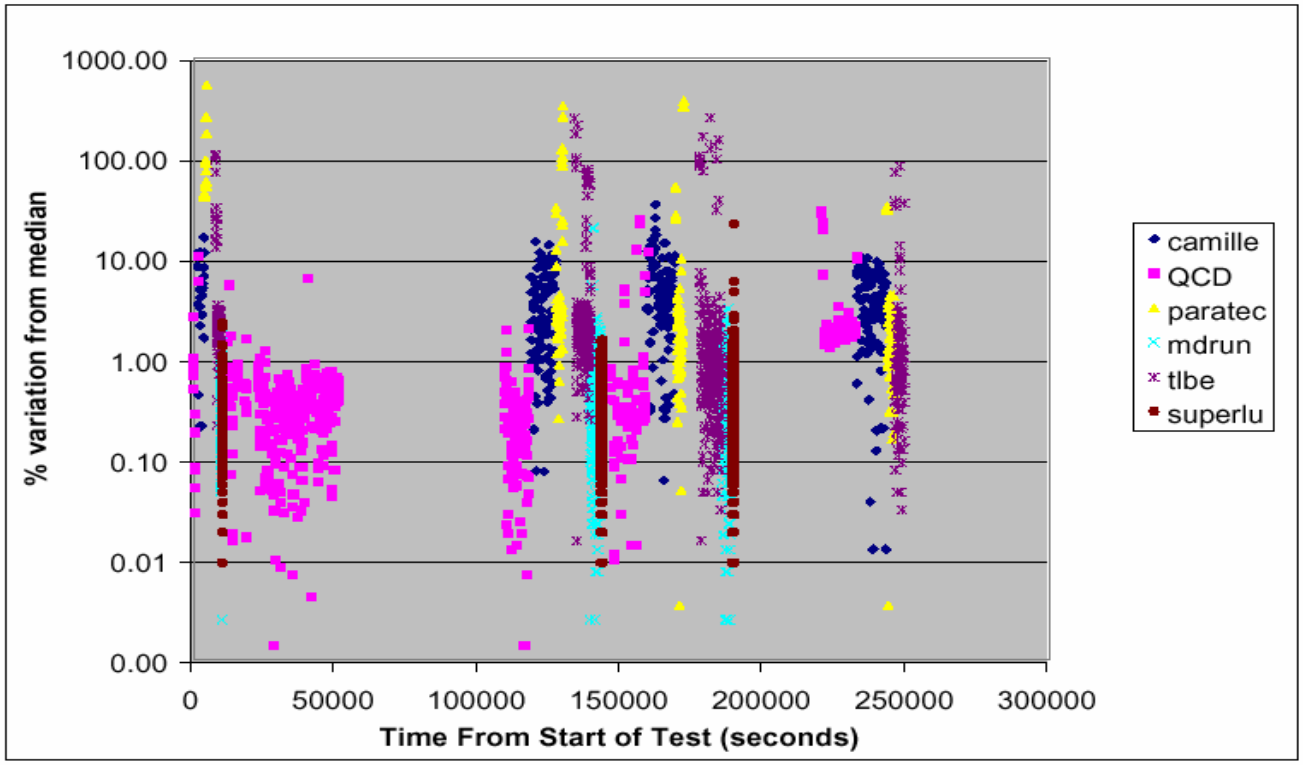
- Sustained System Performance (SSP)
- Throughput
  - Test simple job scheduling ability of system
  - Set of medium concurrency jobs
  - Use application benchmarks
- Full configuration
  - Large-scale FFT calculation
- Effective System Performance (ESP)
  - Approximate measure of the efficiency of the system in production
  - Mixture of medium- and large-scale jobs, large-scale priority job, shutdown and reboot

SSP

- **Reflects performance of NERSC scientific applications**
  - **Measure number of Flops for each application benchmark on reference architecture**
  - **Application benchmark concurrencies chosen to be representative of normal use**
  - **Vendors time (or project) application benchmarks and compute Flop/s for the proposed system**
  - **SSP value is geometric mean of performance across application benchmark suite**
  - **SSP generalized for heterogeneous systems/processors**
- **Predicted runtime variance required**
- **Sizes system for vendors**

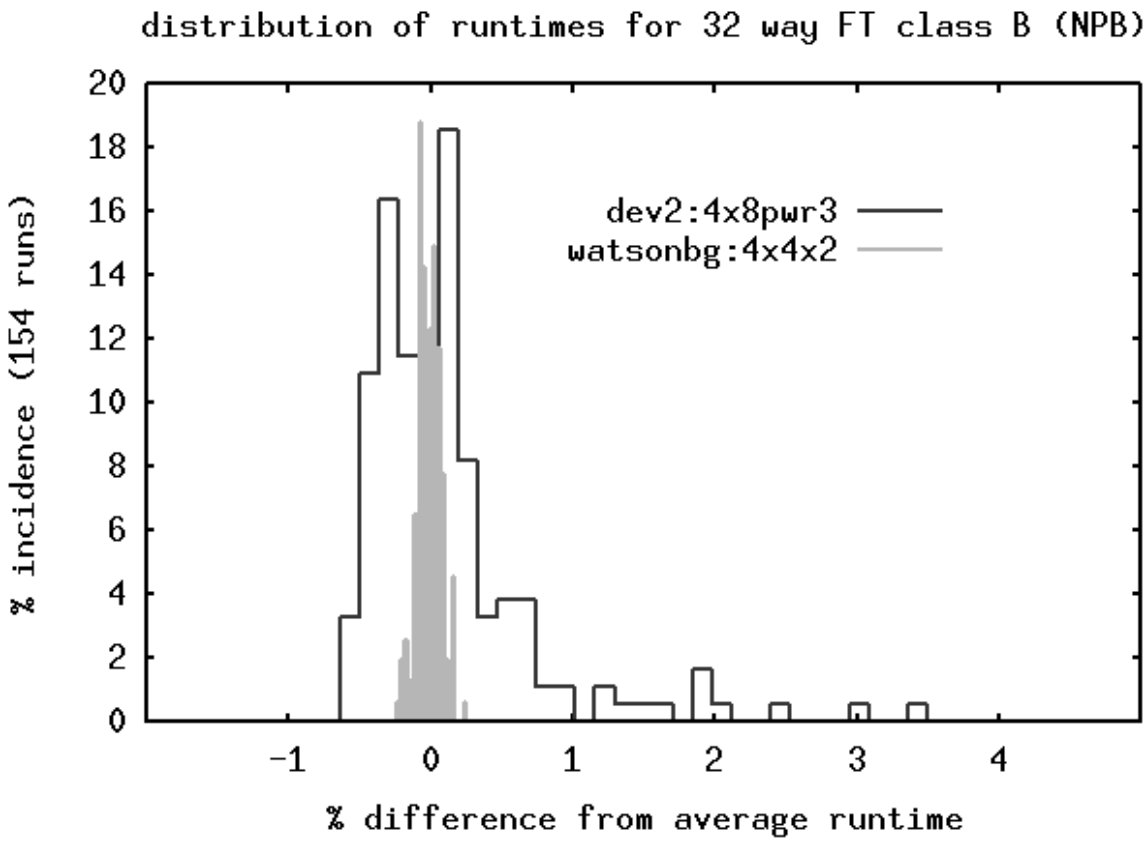
Variability

Consistency Sometimes is lacking



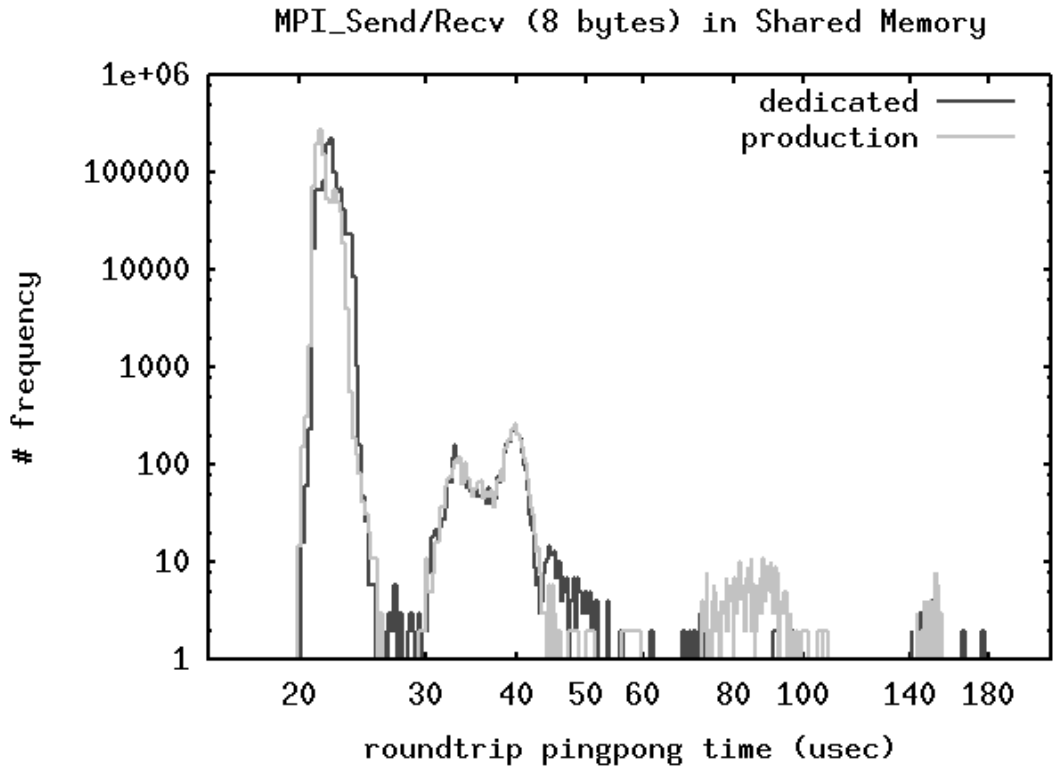
The variation in performance of 6 full applications that were part of the NERSC IBM running with 256 way concurrency SSP benchmark suite used for system acceptance. The codes were run over a three day period with very little else on the system. The run time variation shows that large-scale parallel systems exhibit significant variation unless carefully designed and configured

System Design Influences Consistency



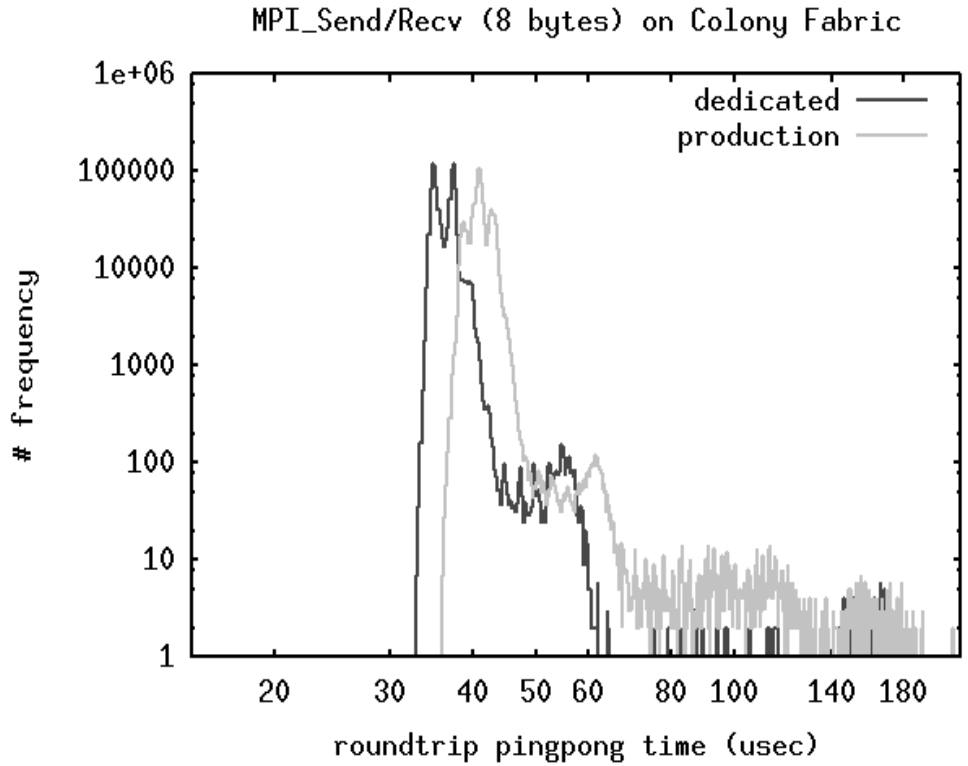
Relative distributions of runtimes for the class B FT NPB compared between an AIX / IBM SP cluster (dark) and a micro kernel based Blue Gene System (light).

MPI Influences Consistency



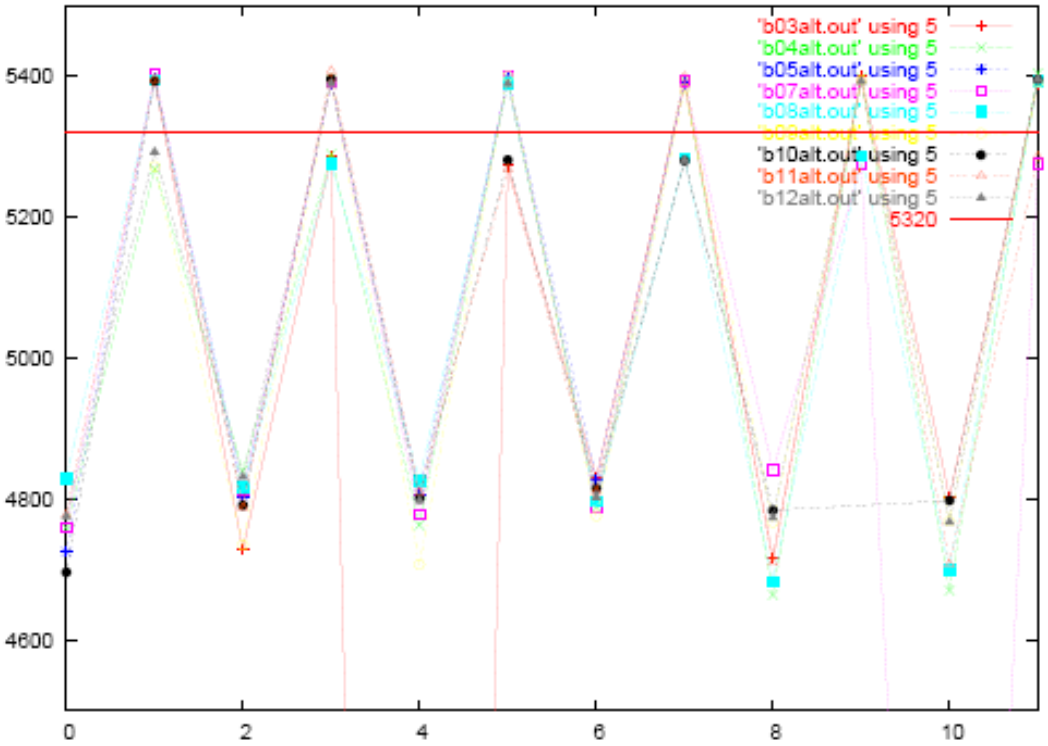
Distribution of intra-node roundtrip MPI_Send/MPI_Recv times through shared memory in dedicated and production modes. P0 shows the nominal performance and X1 ,X2 ,X3 show modes of variability that detract from P0

Consistency is Not Just Due to Busy Systems



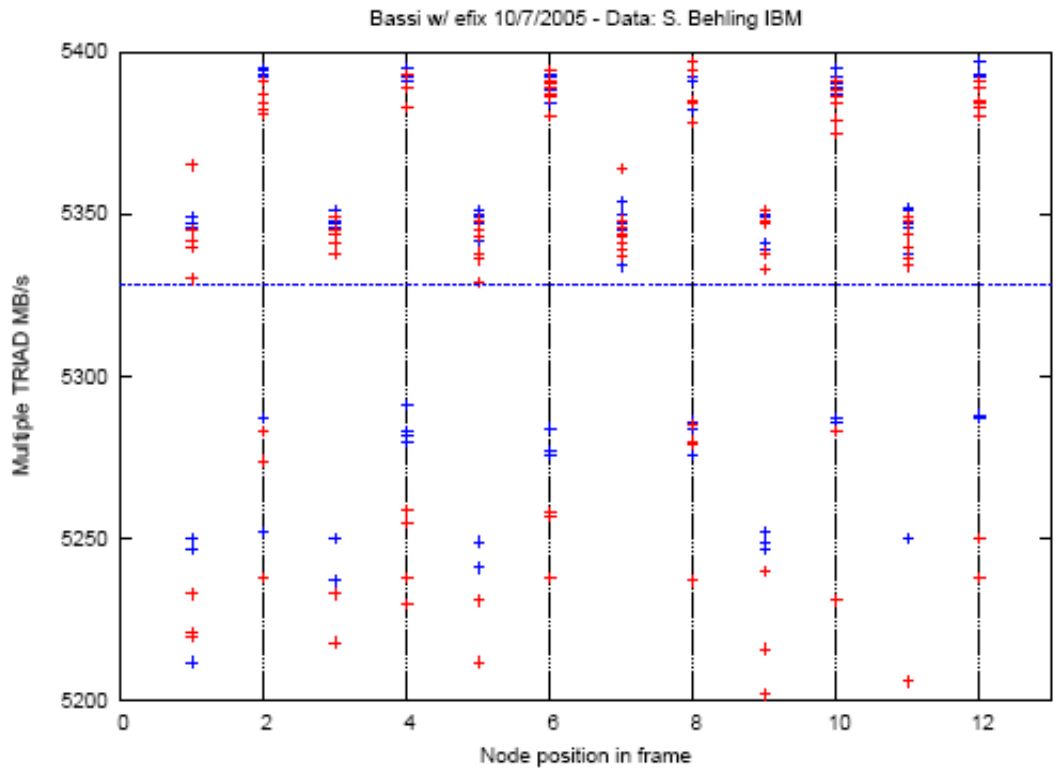
Distribution of intra-node roundtrip MPI_Send/MPI_Recv times through the colony switch fabric in dedicated and production modes.

Consistency is due to hardware configuration choices



Memory test performance depends where the adaptor is plugged in.

Consistency should be expected



Changing the assignments of large pages improved the problem.