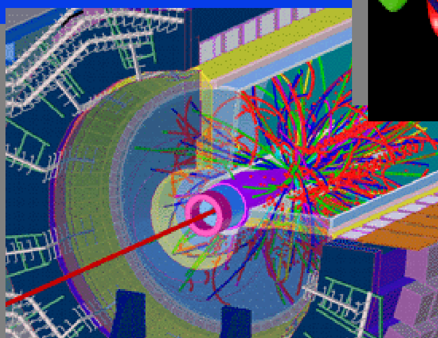
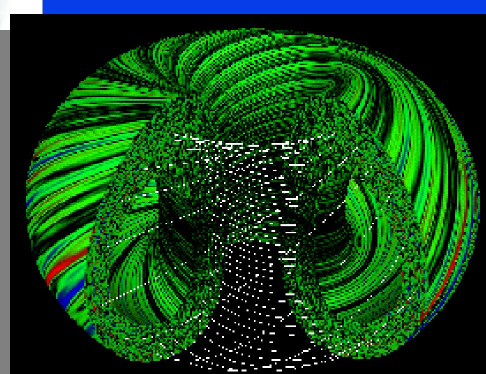
Report on NERSC Upgrade and Plans

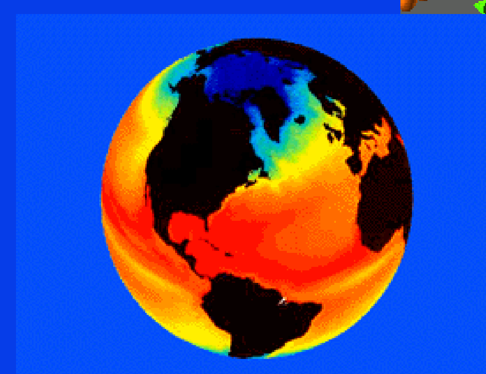
Horst D. Simon
Lawrence Berkeley National Laboratory

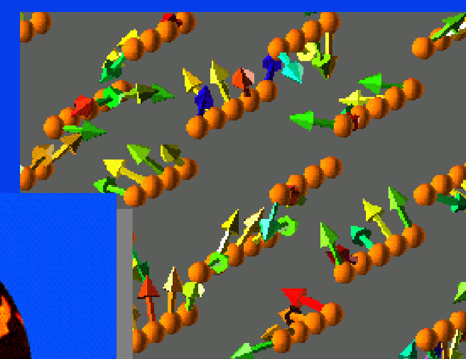
National Energy Research Scientific Computing Center

- Serves all disciplines of the DOE Office of Science
- ~2000 Users in ~400 projects
- Focus on large-scale computing









NERSC Center Division at LBNL

The National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory is one of the nation's most powerful unclassified computing resources and is a world leader in accelerating scientific discovery through computation.

Horst Simon, Division Director

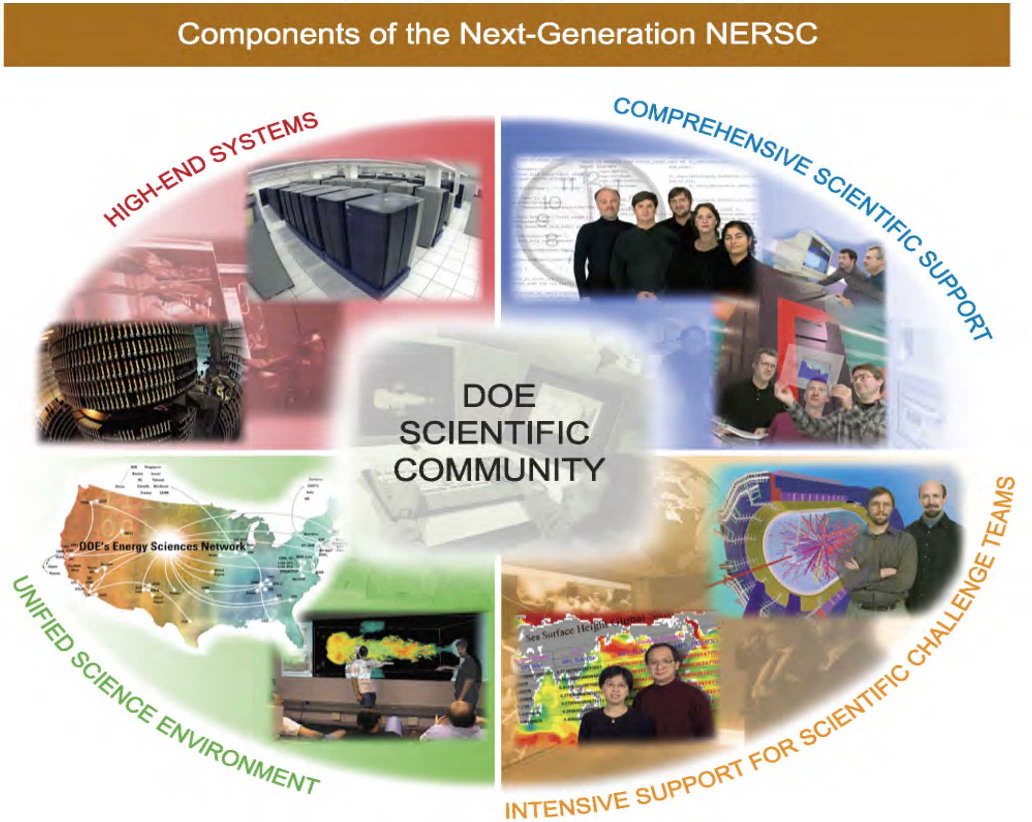
Bill Kramer, Deputy and Facility General Manager

Groups:

- Advanced Systems
- Computational Systems
- Computer Operations and Networking Support
- HENP Computing
- Mass Storage
- Networking and Security
- User Systems

Total Staff: 78

FY2002 New Strategic Plan



- **First full year of operation under new strategic plan**
- **Full review by DOE in 2001**
- **Defines NERSC as general purpose, full service, capability center**

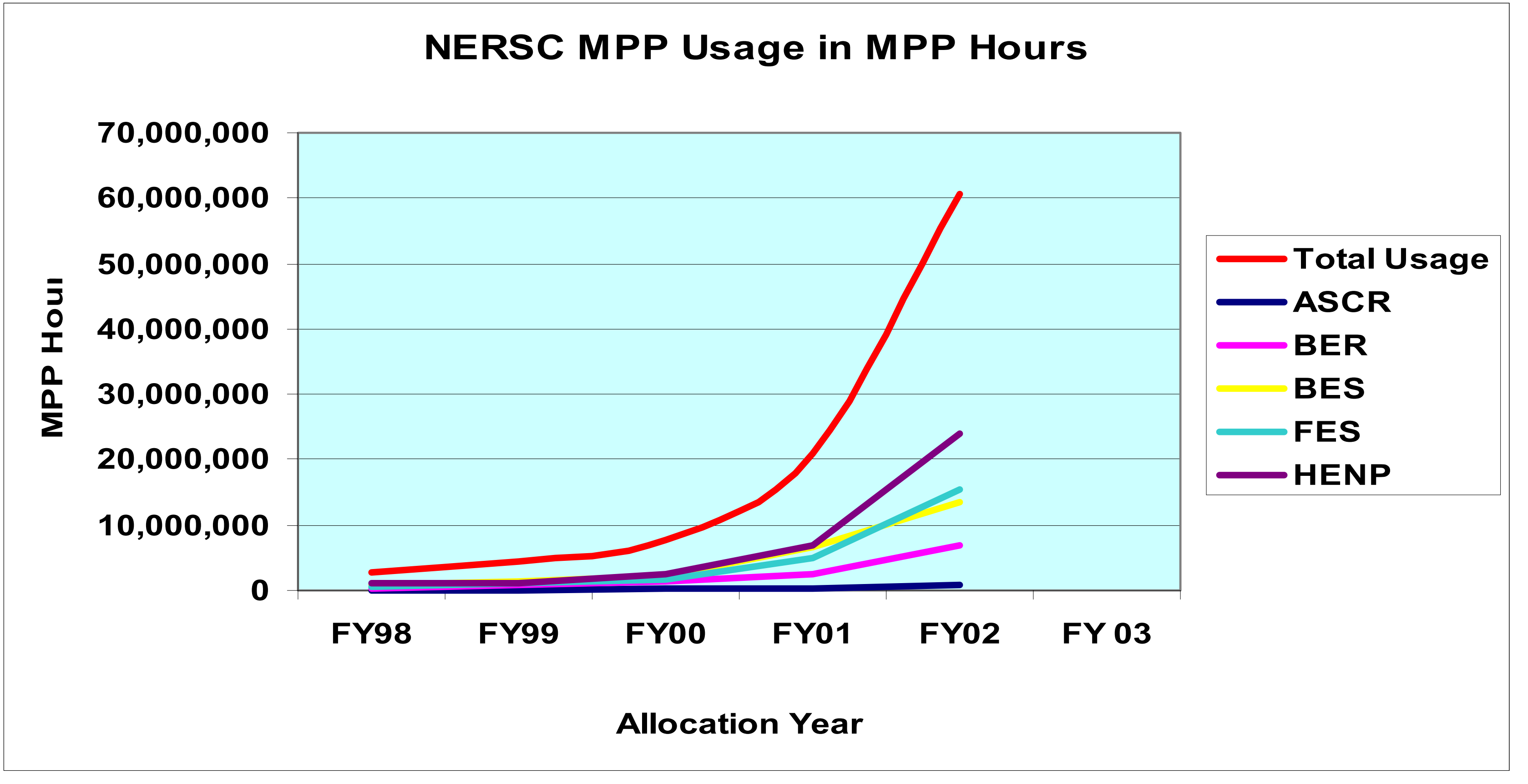
FY2002 Accomplishments

- High End Systems
  - — Upgraded NERSC 3 (“Seaborg”) to 10 Tflop/s system
  - — Increased HPSS storage capacity to 7PBytes
- Comprehensive Scientific Support
  - — Reached >95% utilization on Seaborg
  - — Received excellent ratings in User Survey
- Intensive Support for Scientific Challenge Teams
  - — Support of “Big Splash” users and SciDAC projects
- Unified Science Environment
  - — Introduced Grid services at NERSC
  - — MOU with IBM

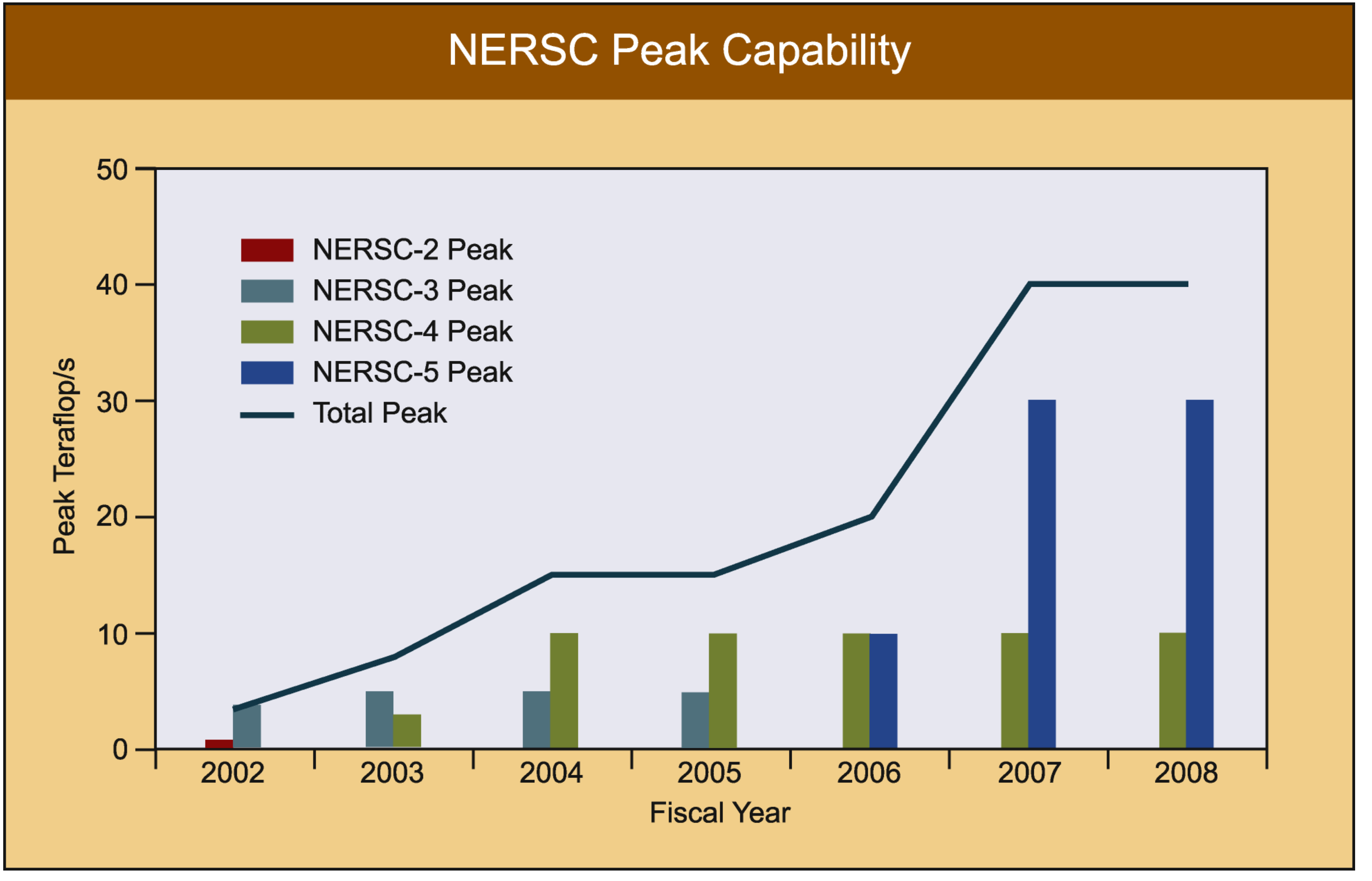
Expanding NERSC's Computational Capability

Increasing Demand for NERSC Resources

- SciDAC and new DOE programs created new demand for NERSC Resources
- SciDAC did not provide for additional facility resources



NERSC Peak Capability as Projected in the Strategic Plan



Current NERSC Strategy

- To achieve major increases in computational capability every three years to replace the Generation N-2 technology.
  - Two generations of systems in service at a time
- System requirements derived from the NERSC User Group Greenbook that represents all the computational, storage and service requirements of each DOE/SC program office and from strategic DOE thrusts
- Procurement done with the Best Value Method that uses only measurable/projected values based on the NERSC current and future scientific workload to determine the best value

NERSC 4 Became NERSC 3E

- NERSC 4 procurement did not produce a cost-effective independent new machine that could be installed in 2003
- Instead, NERSC decided to upgrade the current system and double its size
  - — NERSC 3E provides large capability available immediately
- There was no better solution available for a year or longer

Upgraded NERSC 3E Characteristics

- The upgraded NERSC 3E system has
  - 416 16-way Power 3+ nodes with each CPU at 1.5 Gflop/s
    - 380 for computation
  - 6,656 CPUs – 6,080 for computation
  - Total Peak Performance of 10 Teraflop/s
  - Total Aggregate Memory is 7.8 TB
  - Total GPFS disk will be 44 TB
    - Local system disk is an additional 15 TB
  - Combined SSP-2 is greater than 1.238 Tflop/s
  - NERSC 3E is in full production as of March 1,2003
    - nodes arrived in the first two weeks of November
    - Acceptance end of December 2002
    - 30-day availability test near completed Feb. 2003
    - In full production March 1, 2003

Comparison with Other Systems

	NERSC 3 E	ASCI White	ES	PNNL Mid 2003
Nodes	416	512	640	960
CPUs	6,656	8,192	5,120	1900
Peak(Tflops)	10	12	40	11.4
Memory (TB)	7.8	4	10	6.8
Shared Disk(TB)	60	150	700	53
SSP(Gflop/s)	~1,400	1,652	?	?

PNNL system available in Q3 CY2003; 53 TB SAN + 234 TB local disk

SSP = sustained system performance (NERSC applications benchmark)

Benefits of NERSC 3E for DOE/Office of Science Applications

- High Processor Count (6656 proc.)
  - Permits investigation of scalability of applications to new levels
  - Only open production system of this size world-wide
- Large memory (7.8 TB)
  - Permits innovative new “Big Splash” and EXCITE applications
  - Second largest memory on any open production system
- Same architecture and environment as NERSC 3 Base
  - Immediate productive use
- Combining the system
  - Reduces system administration cost, disk storage
  - Improves utilization

Selection Based on DOE Scientific Applications

Application (* Indicated code was part of SSP-2 calculation)	Scientific Discipline	Algorithm or Method	MPI Task	System Size
GTC*	Plasma Physics	Particle-in-cell	256	10^7 ions
MADCAP*	Cosmology	Matrix inversion	484	40000x 40000
MILC*	Particle Physics	Lattice QCD	512	32^3x64
NAMD	Biophysics	Molecular dynamics	1024	92224 atoms
NWChem	Chemistry	Density functional	256	125 atoms
Paratec*	Material Science	Density functional	128	432 atoms
SEAM*	Climate	Finite element	1024	30 days

* indicates codes that make up the Sustained System Performance (SSP) Metric
There are also tests for I/O, Networking, Throughput,
Effective System Performance, Variation, functionality and many others

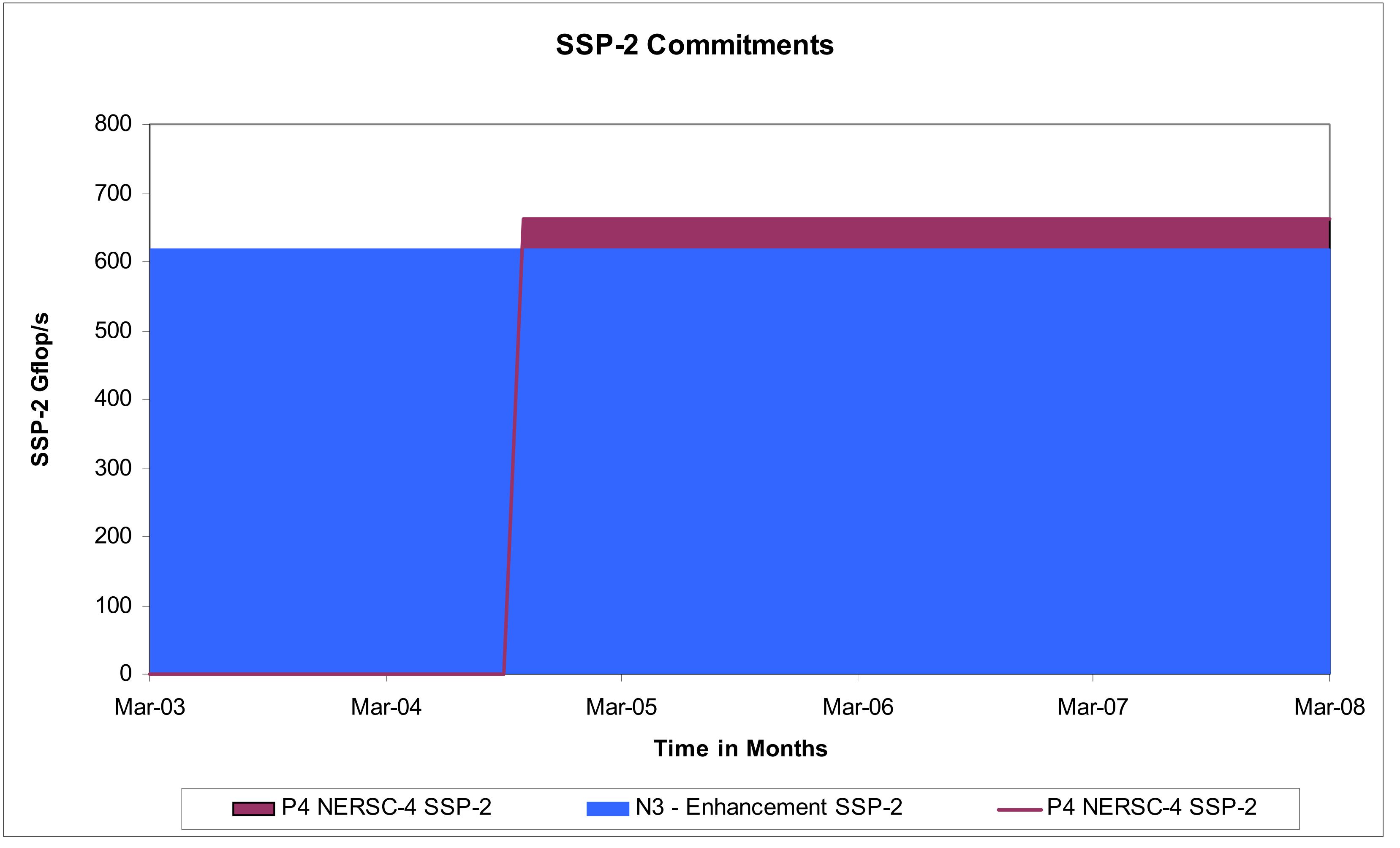
Power 4 vs. Power 3

- By simple measures a Power 4+/Federation should be 4 to 10 times better than an equal number of Power 3 CPUs
  — 4.5 times the Gflop/s per CPU, 9 times the GFlop/s per node, 8 times the interconnect bandwidth, 11 times the memory bandwidth, etc
- Measured performance did not track with peak improvements
  — Average improvement for real applications was only 2.5 times better
  — The integrated Sustained System Performance Metric was actually worse than on Power 3
    - Fewer CPUs for the same cost

Why?

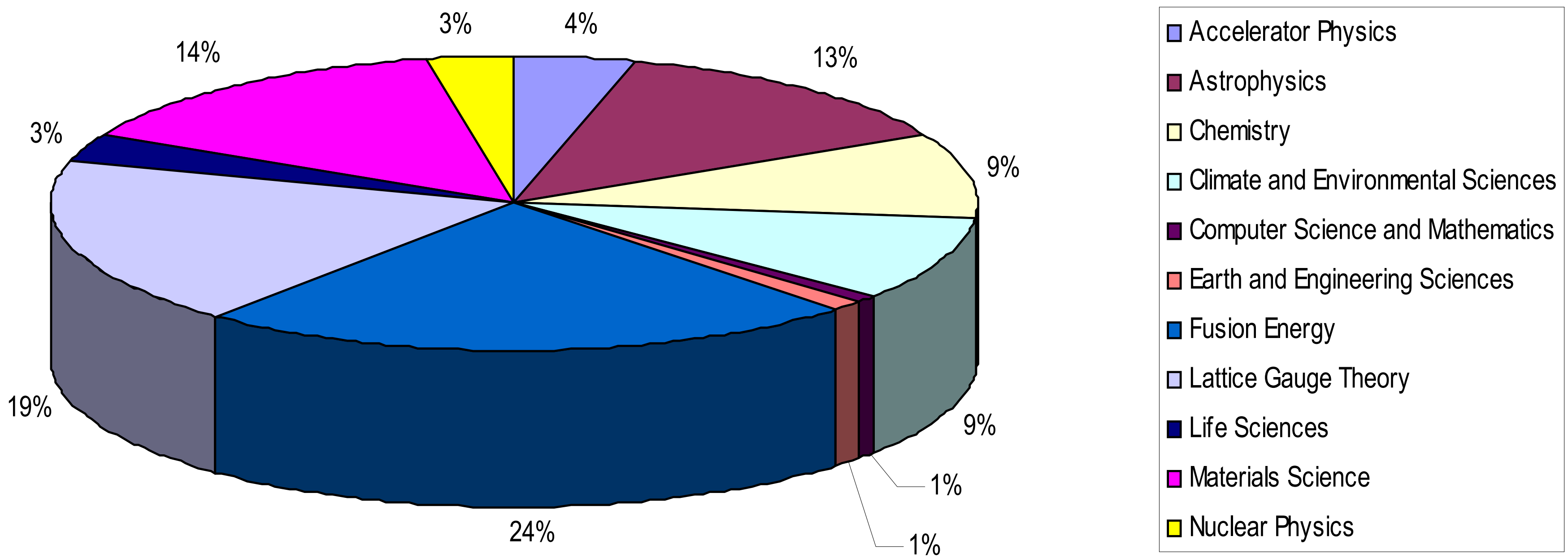
- Memory latency did not improve. In fact, it got relatively worse.
  — Aggravated by the lack of rename registers that generated more flushes of the instruction pipeline
- Power 4 nodes do not scale well for more than 16 scientific tasks

N3E Sustained System Performance (SSP) 36% better over five years

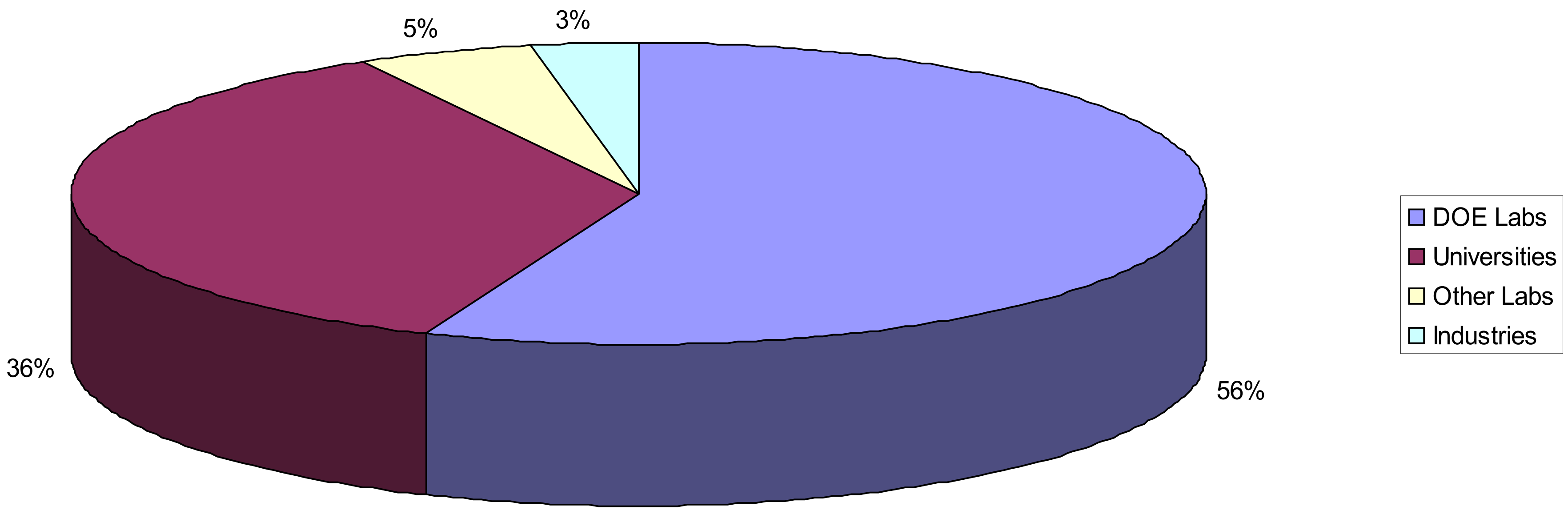


Users, Allocations, Utilization

FY02 Usage by Scientific Discipline



FY02 Usage by Institution Type



FY 2003 Allocations

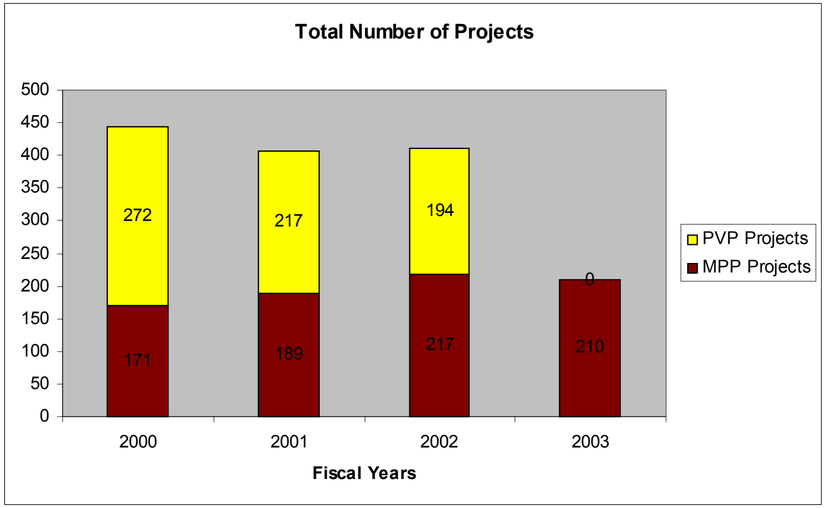
- DOE initiated a new allocations process for FY 2003.
- Open to all DOE Office of Science mission-relevant applications
- Computational Review Panel (CORP) conducts a computational review of all DOE Base requests.
- DOE Program Managers make all production (SciDAC and DOE Base) awards, considering CORP input
- NERSC makes all Startup awards
- Special selection process for “Big Splash”

FY 2003 NERSC Center Allocations

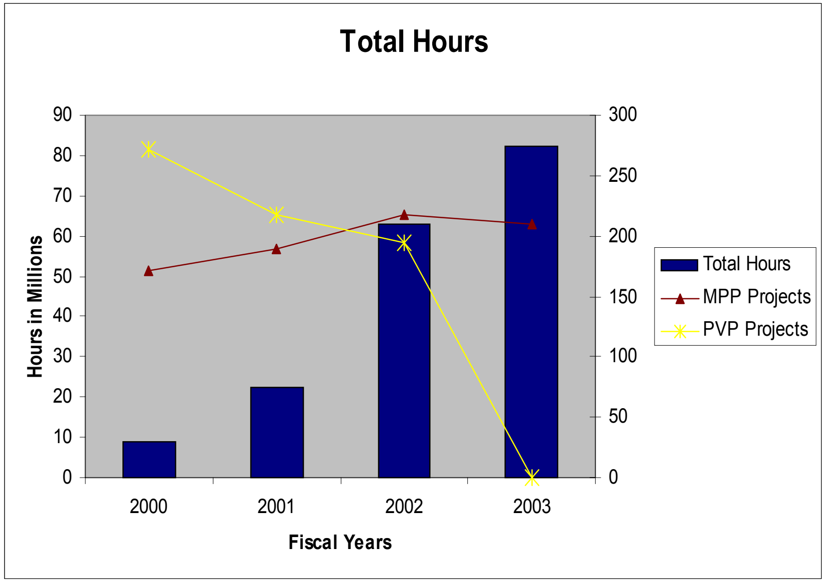
Award Category	Number of k MPP hours	Number of Projects
EXCITE	7,500 (9.4 %)	~5
Big Splash	5,780 (7.2%)	3
SciDAC	18,580 (23.2 %)	20
DOE Base	48,290 (60.2 %)	182

- Smaller number of projects compared to FY2002
- Focus on capability projects (EXCITE and Big Splash)

Increase in Capability Computing

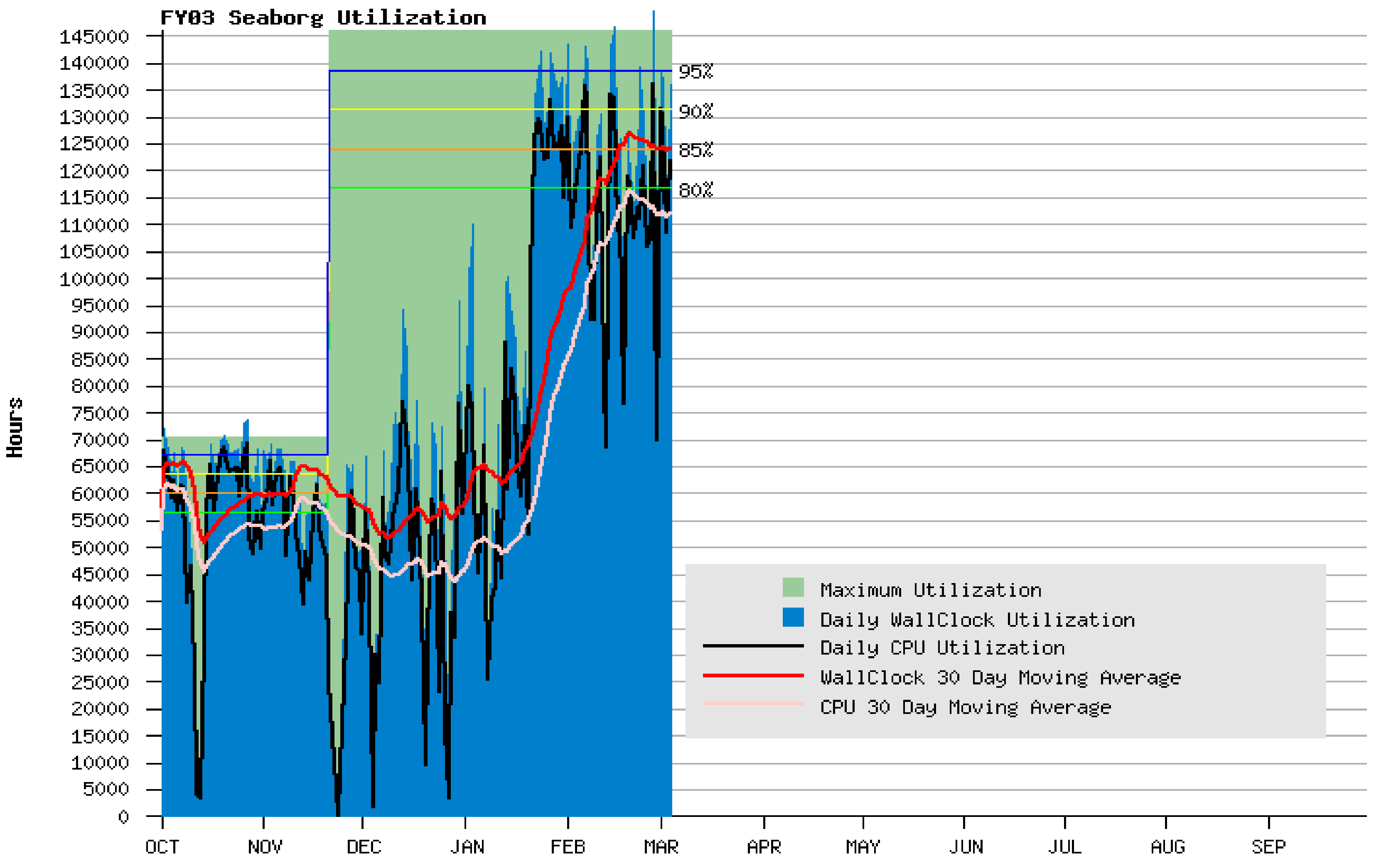


The number of projects at NERSC has significantly decreased.



The amount of available hours has significantly increased

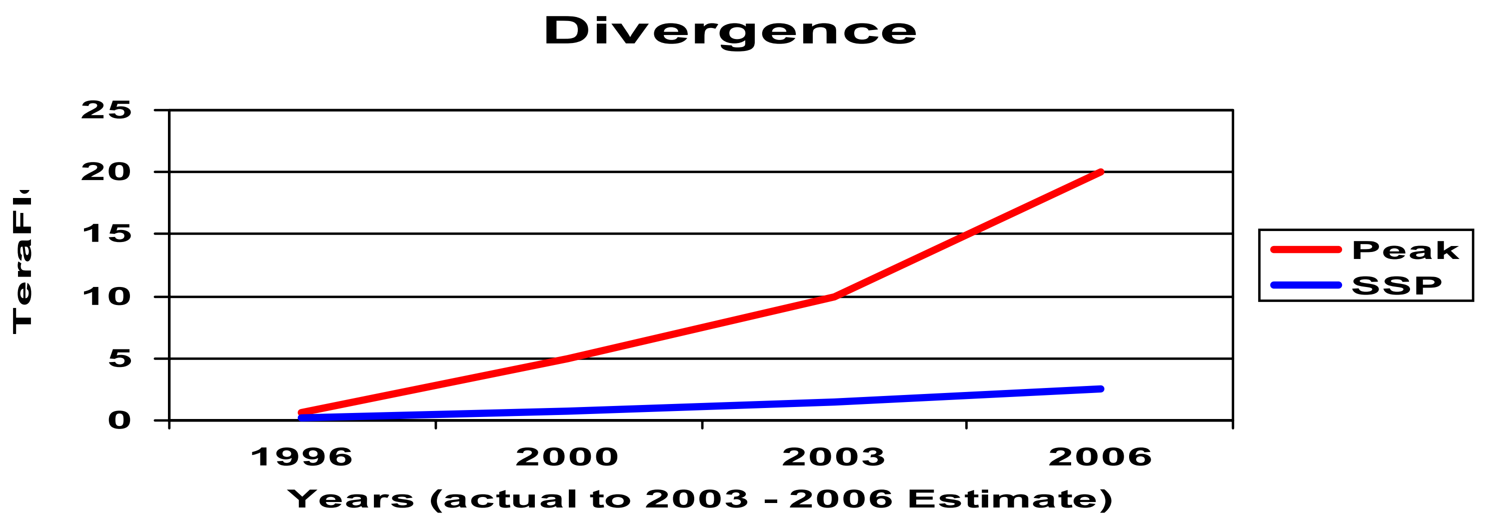
Seaborg Utilization



Future Hardware Strategy for NERSC

The Divergence Problem

- The requirements of high performance computing for science and engineering and the requirements of the commercial market are diverging.
- The commercial-clusters-of-SMP approach is no longer sufficient to provide the highest level of performance
  - Lack of memory bandwidth
  - High interconnect latency
  - Lack of interconnect bandwidth
  - Lack of high performance parallel I/O
  - High cost of ownership for large scale systems



Current System Designers do not Understand Scientific Needs

- Do not understand the memory usage of scientific applications
- Many things done to emphasize theoretical peaks
- Example: IBM designers had science codes that were 5-10 years old as their target applications
  - Assumptions could result in worst-case performance for a sparse DAXPY, common to many codes, of $1/16^{th}$ of peak for larger SMPs
- Memory subsystems designed for capacity
- Interconnects remain very problematic
- Large-scale I/O being ignored by many vendors and self-built systems
- Unjustified optimism for effectiveness of the design on sustained performance

There is a growing recognition in the U.S. vendor design community that this is a problem.

Cooperative Development NERSC/ANL/IBM Workshop

- Goal: Pursue a path(s) to provide a system that can have sustained performance in the range of 30-50% on systems with peak performances of more than one petaflop/s....
- Shorter term goal: By 2005, field a computer at twice the applications performance of the Earth Simulator that is on a sustainable path for scientific computing
- Held two joint workshops
  - Sept 2002 – defining the Blue Planet architecture
  - Nov. 2002 – IBM gathered input for Power 6
- Developed White Paper "Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership," available at http://www.nersc.gov/news/blueplanet.html

Selection is Based on Scientific Applications

	AMR	Coupled Climate	Astrophysics		Nanoscience	
			MADCAP	Cactus	FLAPW	LSMS
Sensitive to global bisection	X	X	X		X	
Sensitive to processor to memory latency	X	X			X	
Sensitive to network latency	X	X	X	X	X	
Sensitive to point to point communications	X	X				X
Sensitive to OS interference in frequent barriers				X	X	
Benefits from deep CPU pipelining	X	X	X	X	X	X
Benefits from Large SMP nodes	X					

A Multifaceted Response

- Goal is a system better able to support scientific applications
  - System design derived from scientific applications
- Blue Planet
  - A compromise between the best for science and what is cost-effective, practical deviation from “business as usual”
  - Goal is sustained scientific performance that is long-term and viable so cost and leverage are key
- Blue Gene
  - Not on standard roadmap
  - Higher risk and less certainty about the scope of applications that can be effective
- Cray X1
  - Standard offering that has potential
  - Unproven for cost effectiveness
- Room for others
  - Since the paper, we have had discussions with HP, Cray, Intel, AMD, SGI…

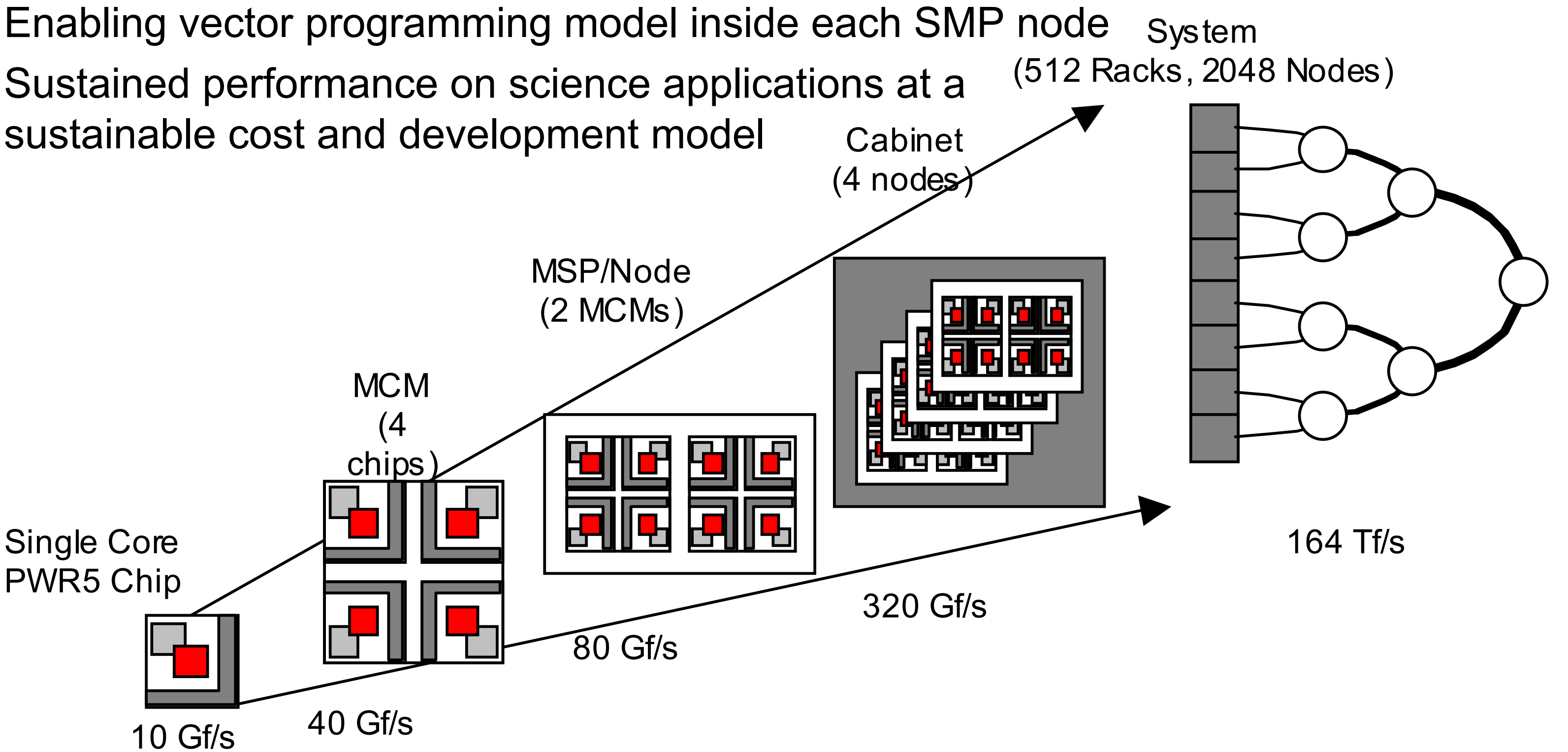
"Blue Planet": Extending IBM Power Technology and Virtual Vector Processing

Addressing the key barriers to effective scientific computing
- Memory bandwidth and latency
- Interconnect bandwidth and latency
- Programmability for scientific applications

- Getting "inside the box" of commercial servers (SMPs)
  - Increasing memory and switch bandwidth using commercial parts available over the the next two years
- Exploration of new architectures with the IBM design team
- Enabling the vector programming model inside an SMP node
- Changing the design of subsequent generations of microprocessors
- It is the first step, not the final result
  - Long lead times for chip designs means we can only influence N+2 and N+3 generations
    - 2.5 years for tweaks, 5 years for redesign
  - Near-term improvements will build momentum

Blue Planet: A Conceptual View

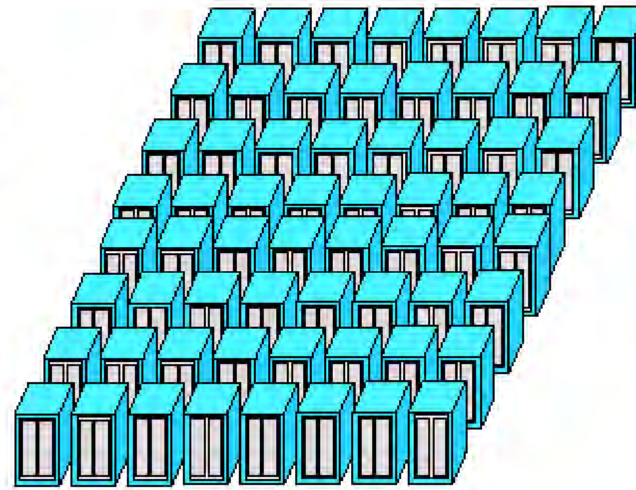
- Increasing memory bandwidth – single core
  - 8 single CPUs are matched with memory address bus limits for full memory bandwidth
- Increasing switch bandwidth – 8-way nodes
- Decreased switch latency while increasing span
- Enabling vector programming model inside each SMP node
- Sustained performance on science applications at a sustainable cost and development model



Ultracomputer Research:

Blue Planet

System
(256 racks/
2,048 nodes/
16,384 processors
+ 160 switch frames)



160 TF/s

Rack
(64 processors/
8 nodes)

640 GF/s

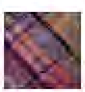
ViVA Node
(8 processors)

80 GF/s

MCM
(4 processors)

40 GF/s

POWER5+ Chip
(1 processor)



10 GF/s

Blue Planet Target Design:

- ✓ **POWER5+ GS single-core chip**
- ✓ **Approx 2.5 GHz**
- ✓ **0.10u 10S2 technology**
- ✓ **2005 availability**

http://www.nersc.gov/news/blueplanetmore.html

Slide courtesy of Peter Ungaro, IBM

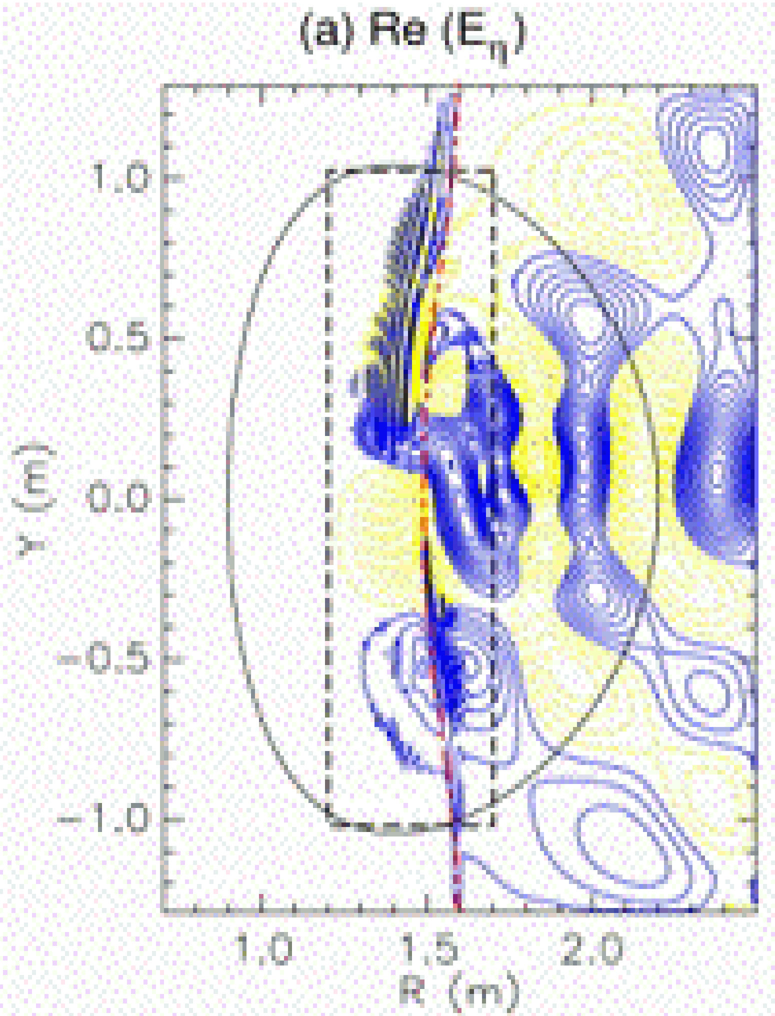
Science Results on NERSC 3 E

Linpack on N3E with 416 nodes

- Performance of original Linpack Benchmark Code (HPL): 6.135 Tflop/s on a matrix of order 409,600 (61.4% of peak).
- LBNL enhancements to the HPL code incorporating:
  - IBM specific non-blocking broadcast calls
  - Shared memory on nodes coupled with SMP-aware communication to reduce memory copies
  - Improved placement of tasks on nodes (used before)

Size of matrix	nodes	Rate (Tflop/s)	% Peak
368,000	208	3.53	70.7%
409,600	416	6.87	68.8%
512,000	416	7.21	72.2%

Science of Scale: Electromagnetic Wave-Plasma Interactions



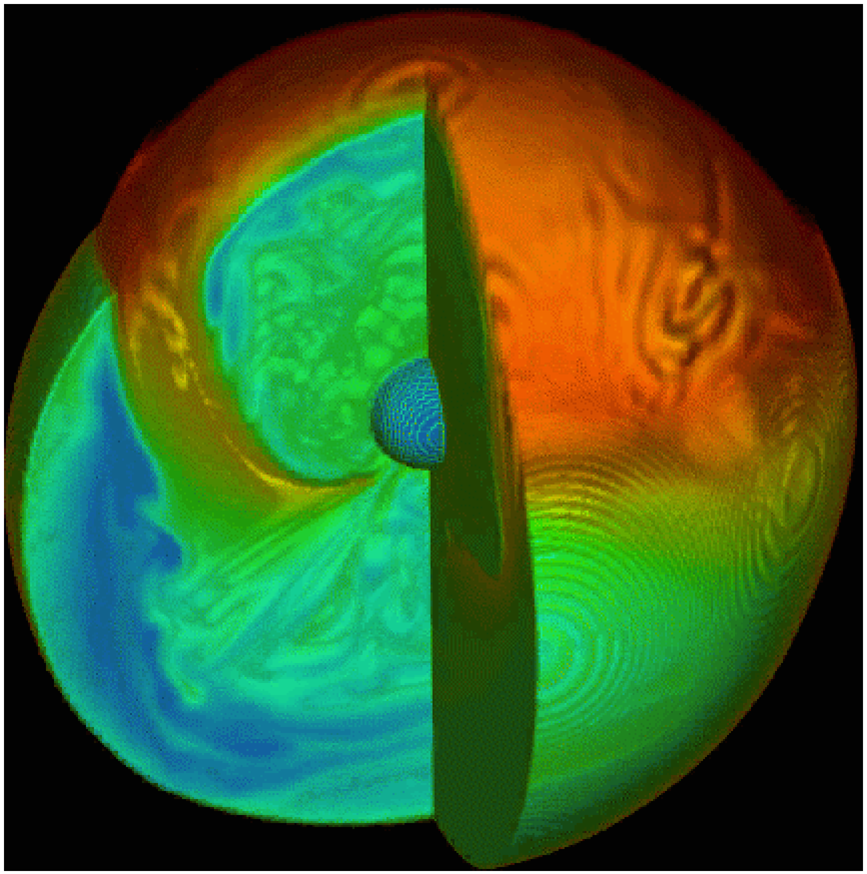
- **PI:** Don Batchelor, ORNL
- **Allocation Category**: SciDAC
- **Code**: all-orders spectral algorithms (AORSA)
- **Kernel**: ScaLAPACK
- **Performance**: 1.026 Gflop/s per processor (68% of peak)
- **Scalability**: 2 Tflop/s on 2,048 processors
- **Allocation**: 1.115 million MPP hours; requested and needs 3 million

Electromagnetic Wave-Plasma Interactions(cont.)

Recent accomplishments:

- Developed new full-wave models called "all-orders spectral algorithms" (AORSA) to take advantage of MPPS when solving the integral form of the wave equation in multi-dimensional plasmas.
- New models give higher resolution 2-D solutions in tokamak geometry and fully 3-D solutions for ion heating in stellarator geometry.
- Calculated poloidal flows that have been observed experimentally; such calculations enhance tokamak confinement regimes (submitted to Physical Review, Jan. 2003).

Science of Scale: Terascale Simulations of Supernovae



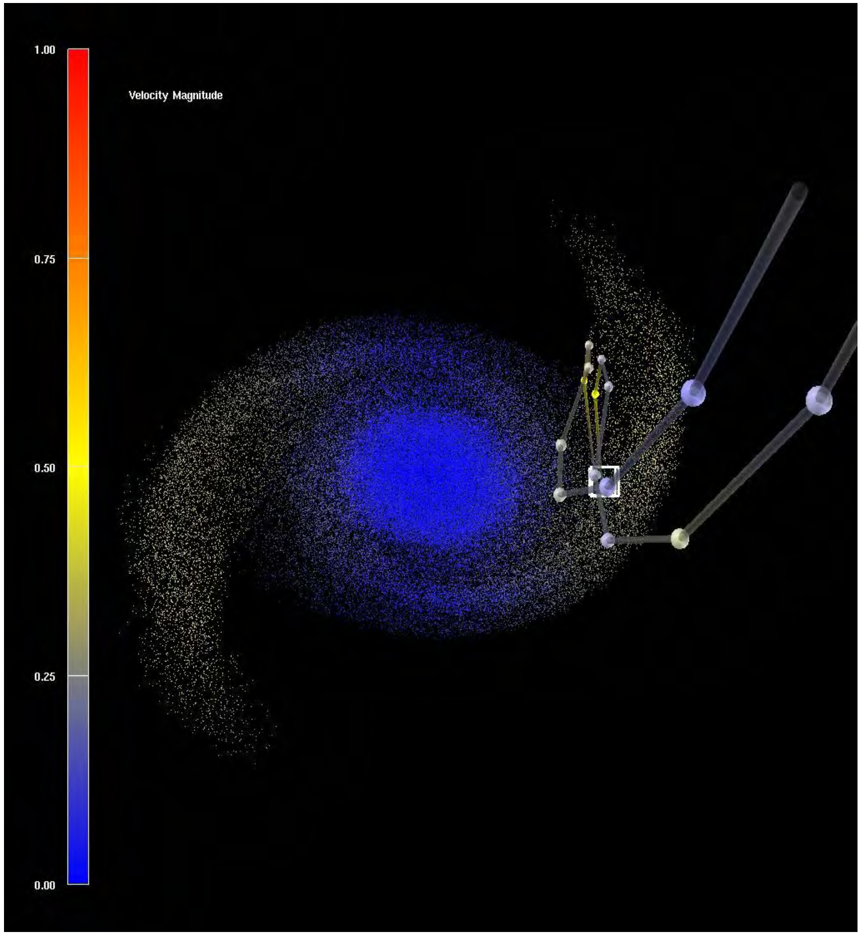
- **PI:** Tony Mezzacappa, ORNL
- **Allocation Category**: SciDAC
- **Code**: neutrino scattering on lattices (OAK3D)
- **Kernel**: complex linear equations
- **Performance**: 537 Mflop/s per processor (35% of peak)
- **Scalability**: 1.1 Tflop/s on 2,048 processors
- **Allocation**: 565,000 MPP hours; requested and needs 1.52 million

Terascale Simulations of Supernovae
(cont.)

Recent accomplishments:

- Developed the OAK3D code to study the electron capture and neutrino scattering on lattices of large arrays of nuclei that form during certain phases of star collapse.
- OAK3D became operational in the Fall of 2002 and has achieved sustained speeds of 1.1 teraflops on 2,048 processors.
- These runs required double precision complex solutions of linear equations of dimension 524,288.

Science of Scale: Accelerator Science and Simulation



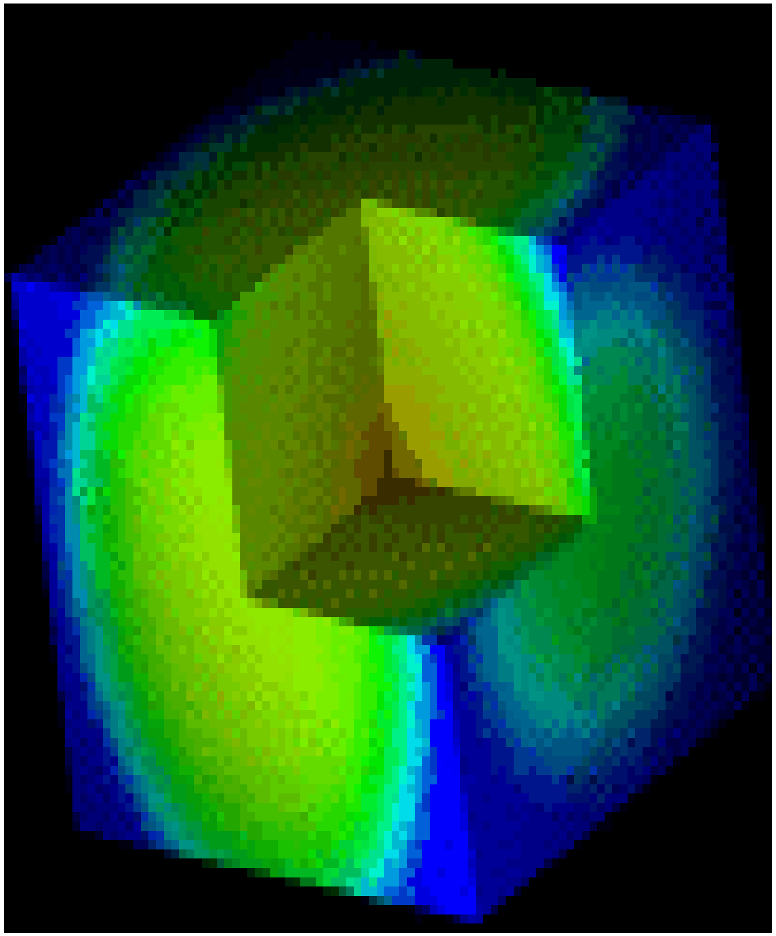
- **PIs:** Kwok Ko, SLAC & Robert Ryne, LBNL
- **Allocation Category**: SciDAC
- **Code**: Beam Dynamics
- **Kernel**: finite element 3D Poisson solver
- **Performance**: being worked on
- **Scalability**: scales to 4,096 processors
- **Allocation**: 1.5 million MPP hours; requested and needs 2.5 million

Accelerator Science and Simulation (cont.)

Recent accomplishments:

- The finite element 3D Poisson solver with semi-structured grids has been improved to scale perfectly up to 4,096 processors; they are confident this will scale to the full machine when MPI can go past 4,096 tasks. Numerical stability and accuracy have been verified. Performance is being worked on.
- Parallel beam-beam code scales up to 2,048 processors with 48% efficiency.
- Parallel MaryLie code achieved 375 Mflops/sec/proc (25% of peak) for 5th order Taylor series tracking (code optimization assistance provided by NERSC User Services group).
- Parallel PIC code of V. Decyk run with 12.4 billion particles, 1024^3 grid.

Science of Scale: Quantum Chromodynamics at High Temperatures



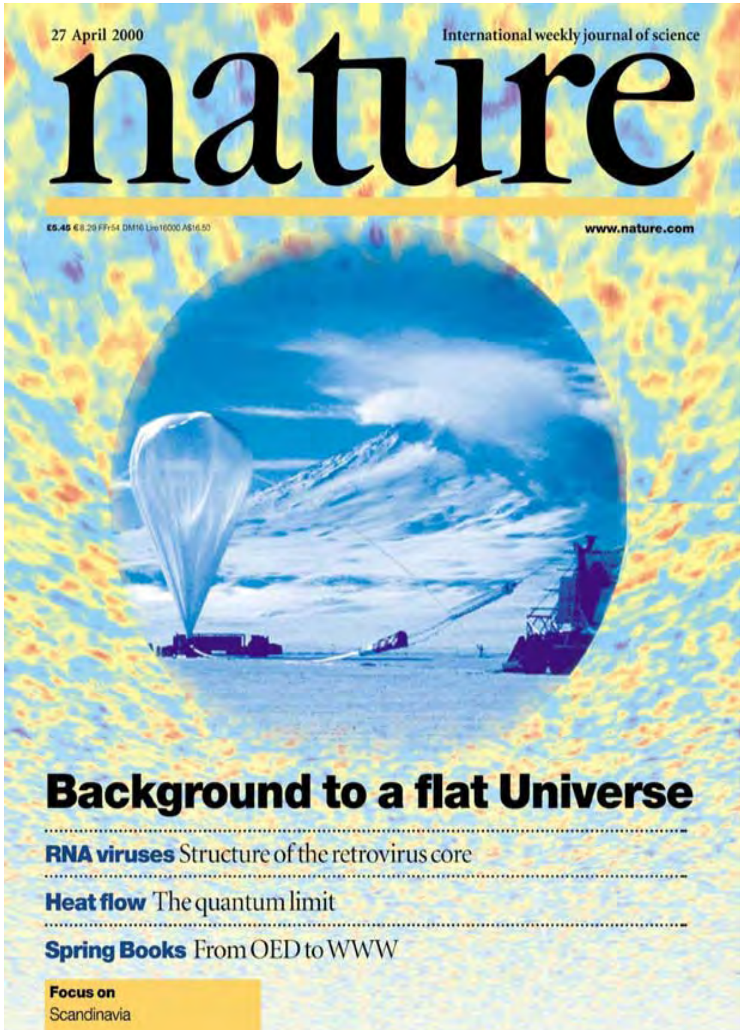
- **PI:** Doug Toussaint, Arizona University
- **Allocation Category**: Class A
- **Code**: hybrid Monte Carlo and Molecular Dynamics (MILC)
- **Kernel**: iterative sparse matrix inversion
- **Performance**: 190 Mflop/s per processor (13% of peak)
- **Scalability**: 200 Gflop/s on 1,024 processors
- **Allocation**: 2.3 million MPP hours; requested and needs 3.4 million

QCD at High Temperatures (cont.)

Recent accomplishments:

- Took advantage of free test time on Seaborg to start work on “next year’s problem”: trial runs of a QCD simulation with a quark mass that is closer to the physical quark masses than we could previously do on this fine a grid. Specifically, light quark masses at 1/10 the strange quark mass with a lattice spacing of 0.09 fm on a 64,000 by 96 lattice.
- Was able to run about 17 units of simulation time. 2,000 units will provide more accurate calculations of hadronic properties: topological structures; theoretical parameters needed for accelerator experiments.

Science of Scale: Cosmic Microwave Background Data Analysis



- **PI:** Julian Borrill, LBNL & UC Berkeley
- **Allocation Category**: Class B
- **Code**: Maximum likelihood angular power spectrum estimation (MADCAP)
- **Kernel**: ScaLAPACK
- **Performance**: 750 Mflop/s per processor (50% of peak)
- **Scalability**:
- 0.78 Tflop/s on 1024 proc
- 1.57 Tflop/s on 2048 proc
- 3.02 Tflop/s on 4096 proc
- **Allocation**: 1.1 million MPP hours; requested and needs 2 million

Cosmic Microwave Background Data Analysis (cont.)

Recent accomplishments:

- MADCAP extended to enable simultaneous analysis of multiple datasets and CMB polarization – the new frontier.
- MADCAP was rewritten to exploit extremely large parallel systems, allowing near-perfect scaling from 256 to 4,096 processors.
- MADCAP++ is being developed using approximate methods to handle extremely large datasets for which matrix multiplications are impractical, such as will be generated by the PLANCK satellite.
- Recent results from NASA's WMAP satellite observations of the whole CMB sky confirm MADCAP analyses of previous partial-sky balloon datasets.

New Results in Climate Modeling

- Recent improvements in hardware have reduced turnaround time for the Parallel Climate Model
- This has enabled an unprecedented ensemble of numerical experiments.
  - — Isolate different sources of atmospheric forcing
    - Natural (solar variability & volcanic aerosols)
    - Human (greenhouse gases, sulfate aerosols, ozone)
- Data from these integrations are freely available to the research community.
  - — By far the largest and most complete climate model dataset
  - — www.nersc.gov/~mwehner/gcm_data

Investigating Atmospheric Structure Changes with PCM

- The tropopause is that height demarking the troposphere and the stratosphere.
  - Below the tropopause, the temperature cools with altitude.
  - Above the tropopause, the temperature warms with altitude.

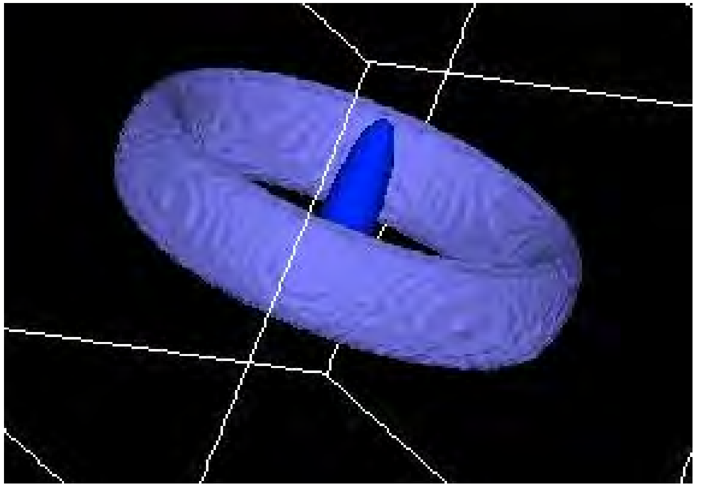
- A diagnostic that is robust to El Nino but sensitive to volcanoes.
- An indicator of the total atmospheric heat content
- Changes in natural forcings alone (blue) fail to simulate this feature of the atmosphere, but natural + anthropogenic changes (orange) do

Summary on NERSC 3E

- NERSC implemented upgrade to 10 Tflop/s successfully and is delivering a new capability to SC community
- Excellent scalability on many large scale applications
- High sustained performance on levels comparable to Earth Simulator
- New science results

More Scientific Results (backup)

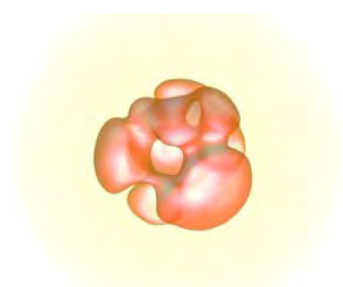
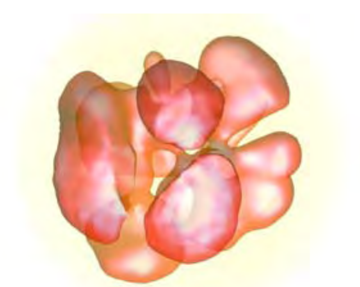
Big Splash Project: Supernova Explosions



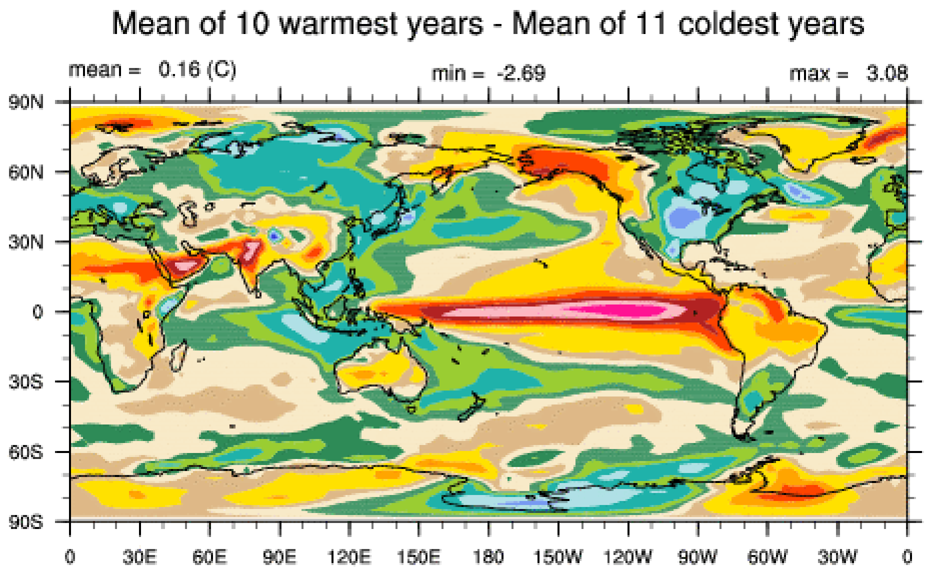
- **PIs:** Adam Burroughs, Arizona State; and Peter Nugent, Berkeley Lab
- **Current Requirements:**
  - — 20 iterations per star model; 20 to 30 models
  - — 1 million MPP hours for 3D simulations with simplified physics;
  - — 10 GB input and 1 GB output per iteration - 6 TB
- **NERSC Provided:** new 24-hour run queue, required to run one iteration and checkpoint

Big Splash Project: Supernova Explosions (cont.)

- **Science Results:** understanding of type 1-A supernovas; first 3D supernova explosion simulation based on computation at NERSC. This research eliminates some of the doubts about earlier 2-D modeling and paves the way for rapid advances on other questions about supernovae.
- **Near-Term Requirements:** figure out how to visualize the data
- **Future Requirements (next 2-3 years):**
  - — 100X CPU for 3D simulations with complex physics if no algorithmic improvements; maybe 10X if new algorithms.
  - — for Supernova Factory will need to receive 50GB daily into HPSS and Seaborg; retrieve 50GB from HPSS; store 25 GB back to HPSS.

SciDAC Project: Climate Change Prediction

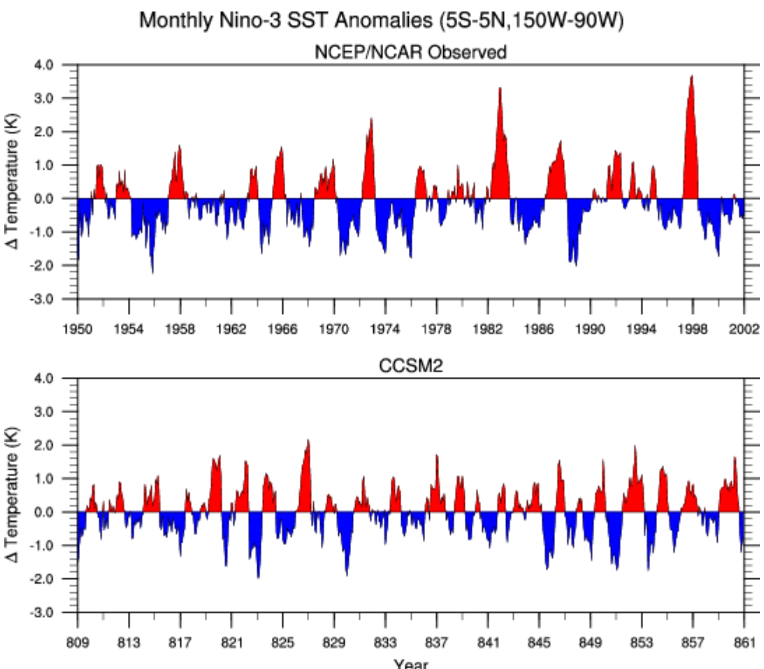


•**PI:** Warren Washington, NCAR

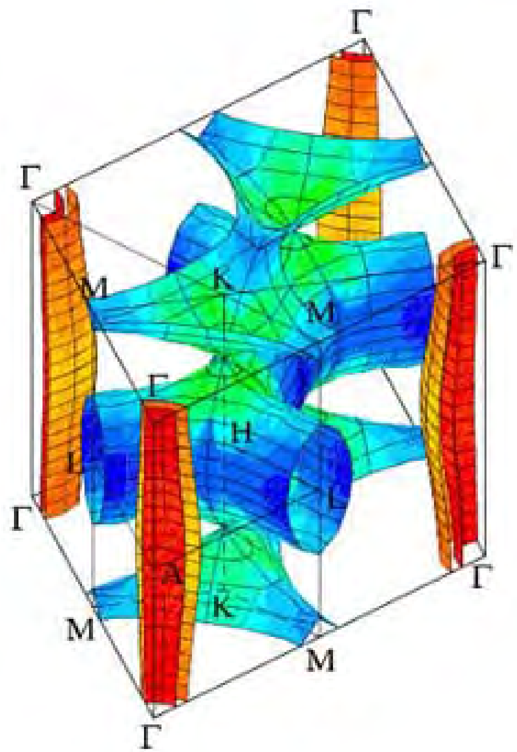
- **Current Requirements:**
  — 1.6 million MPP hours
  — good daily turnaround to process sequential events
  — 6 TB data in HPSS (6 GB per simulation)
  — Make data set available to community

SciDAC Project: Climate Change Prediction (cont.)

- **NERSC Provided:**
  - Prioritized queue scheduling to eliminate wait time between the 1,000 simulations that must be run sequentially
  - Consulting support for code debugging and effective system utilization
- **Science Results**: First 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.
- **Future Requirements (3 years):**
  - 6-8 million MPP hours
  - 12 TB in HPSS
  - Grid access to public data repository

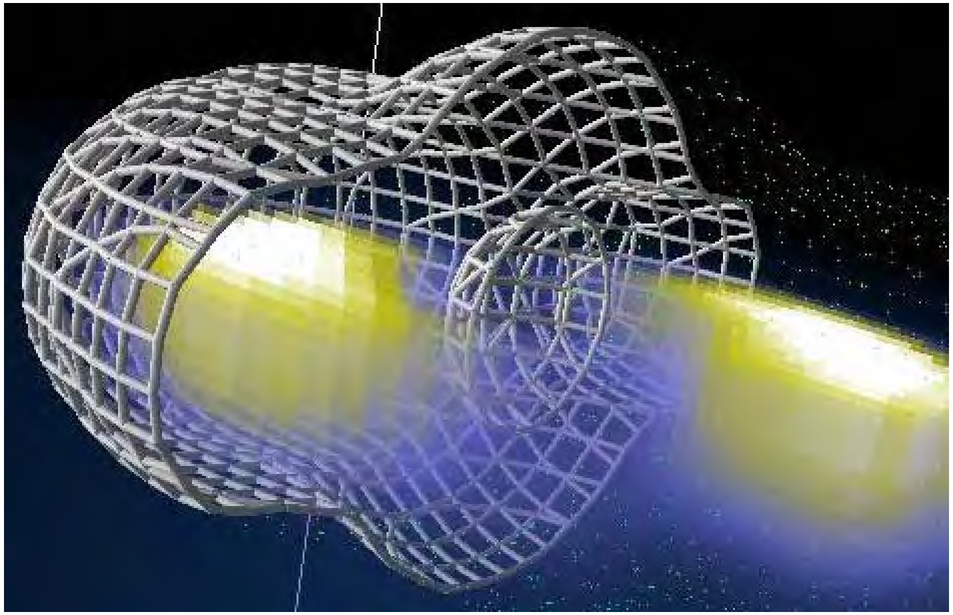


Base Program Project : HT Superconductors



- **PIs:** Marvin Cohen and Steve Louie, UC Berkeley
- **Current Requirements:**
  - — 400,000 MPP hours
- **NERSC Provided:** Collaboration on development of new parallel FFT algorithm
- **Science Result:** Calculated the properties of the unique superconductor MgB2 from first principles, revealing the secrets of its anomalous behavior, including more than one superconducting energy gap; published in *Nature*, August 2002.

Black Hole Merger Simulations

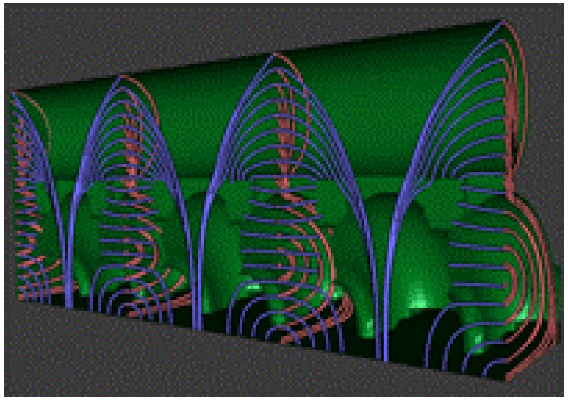


- **PI:** Ed Seidel, Max Planck Institute
- **Current Requirements**:
  — large memory ≥ 1.5 TB & 64-bit MPI
  – ≥ 1 million MPP hours
  — 2 TB scratch disk per run (8+ runs)
  — fast turnaround for parameter studies
- **NERSC Provided:**
  — 2 TB scratch space and 250,000 inodes
  — access to a special queue to improve turnaround
  — opened ports to allow remote-steering and grid access

Black Hole Merger Simulations (cont.)

- — consulting support for 64-bit integration and code debugging
- **Science Results**:
  - — Seaborg enabled the largest-ever black hole collision simulations
  - — confirmed the coalescense characteristics predicted by the French Meudon group over the Cook-Baugamarte model
  - — invaluable for understanding data from new gravitational wave observatories (LIGO, VIRGO)
- **Near-Term Requirements:**
  - — 10 TB disk for each run
  - — 5 TB uniform, user-available memory
  - — 15 million MPP hours

Accelerator Science

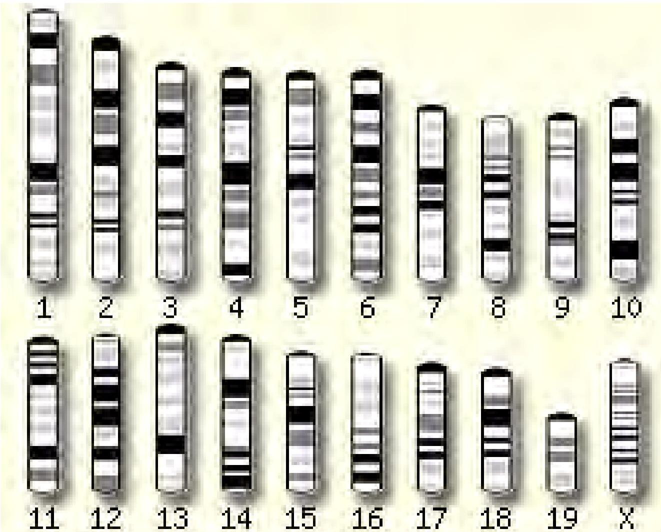


- **PI:** Robert Ryne, Berkeley Lab
- **Current Requirements:**
  - — 1.6 million MPP hours
  - — large memory: up to 2 TB
  - — 64-bit MPI
  - — visualize and post process up to 3 TB of data
- **NERSC Provided:**
  - — 3 TB scratch space
  - — consulting support for large memory management and performance analysis
  - — CVS support and web hosting

Accelerator Science (cont.)

- **Science Results:**
  - — understand beam heating for PEP-II (SLAC) upgrade
  - — help design the Next Linear Collider accelerating structure
  - — understand emittance growth in high intensity beams
  - — study laser wakefield accelerator concepts for future accelerator design
- **Future Requirements (3 years)**:
  - — 15-20 million MPP hours
  - — 5+ TB scratch space
  - — continued consulting support

JAZZ Genome Assembler



- **PI:** Dan Rokhsar, Joint Genome Institute
- **Current Requirements**: Fugu assembly required 30 GB for database files and 150 GB of scratch space.

- **NERSC Provided:**
  — porting of JAZZ assembler, BLAST alignment tool, cross_match alignment tool, and MySQL client to *the IBM SP*
  — a dedicated MySQL server
  — resolved issues installing a MySQL server on the IBM SP

JAZZ Genome Assembler (cont.)

— consulting support for parallelization of BLAST and cross_match tool

- **Science Results:** Assembly of Fugu genome from 3.1 million reads, and initial preparation of mouse genome data.
- **Near-Term Requirements:** Initial mouse assembly will require 75 GB for database files and 500 GB of intermediate data. As more raw data is added, this could easily double.