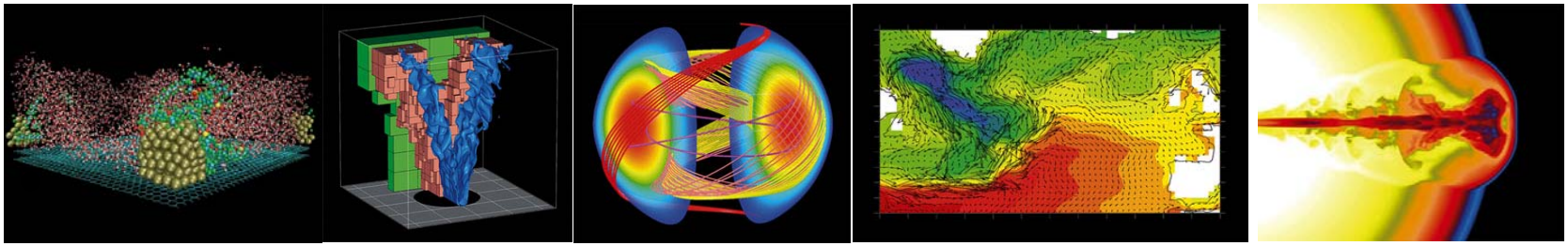
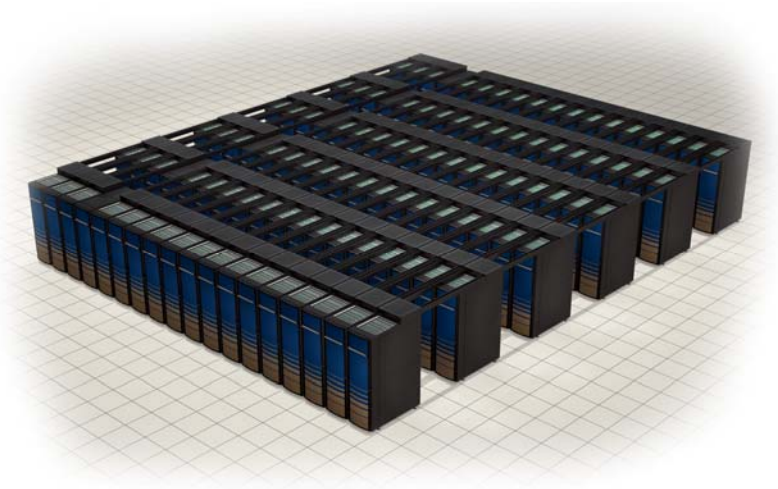


NERSC 5



Bill Kramer

kramer@nersc.gov

(510) 486-7577

Ernest Orlando Lawrence
Berkeley National Laboratory

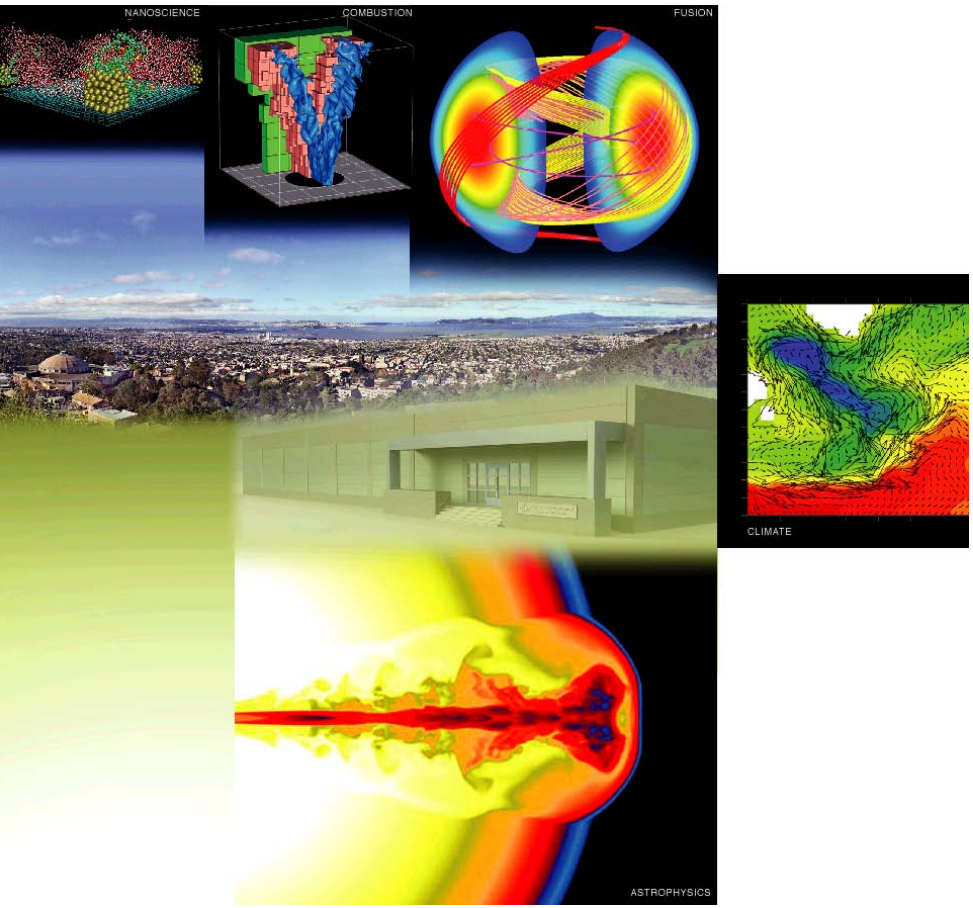
NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

BERKELEY LAB

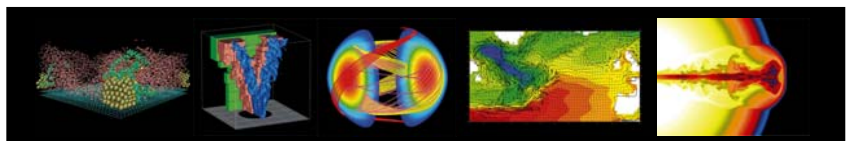
Office of Science
U.S. DEPARTMENT OF ENERGY

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.

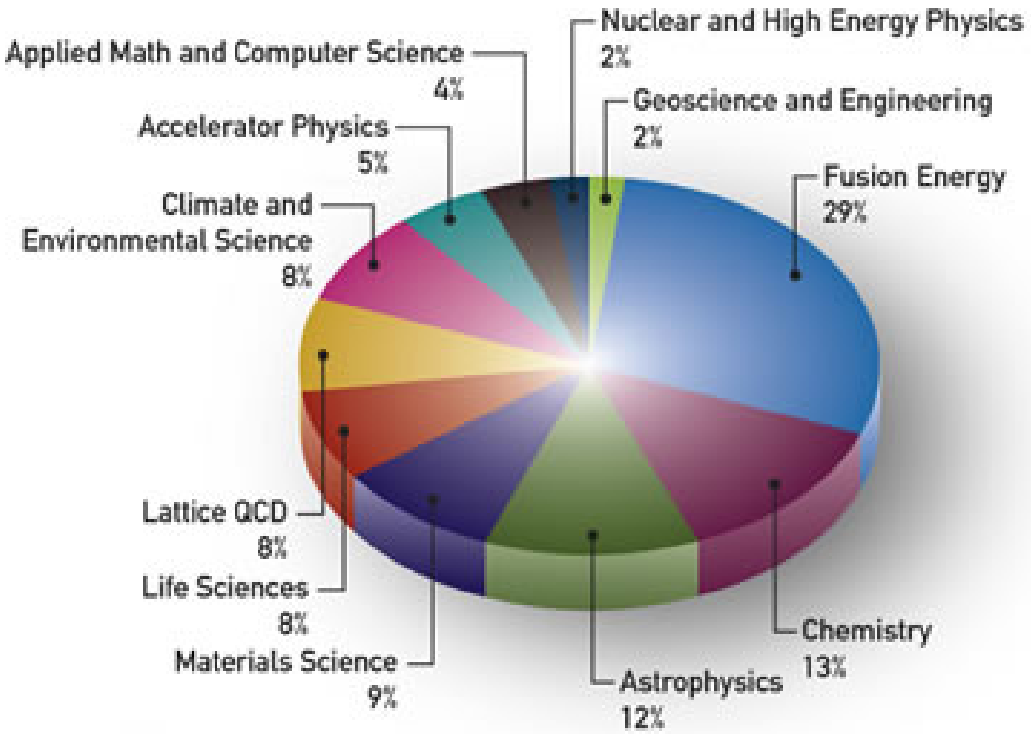
Outline

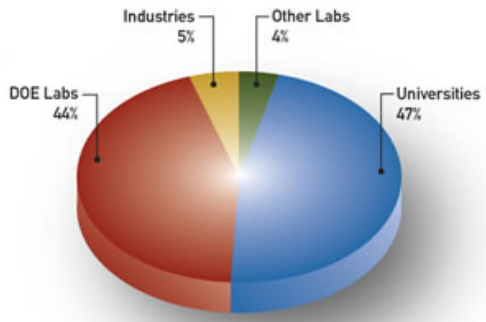


- NERSC Background
- Goals for NERSC 5 RFP
- Technology Observations
- The End Result

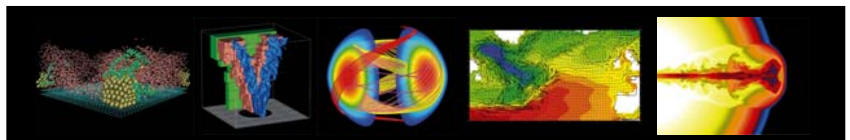


NERSC Usage for AY 2005



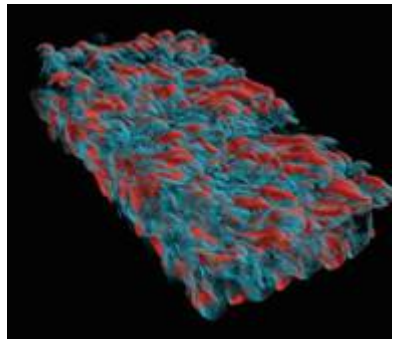


http://www.nersc.gov/news/annual_reports/annrep05/

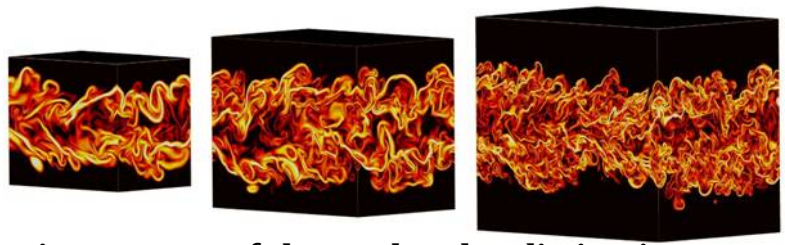


INCITE: Direct Numerical Simulation of Turbulent Non-premixed Combustion

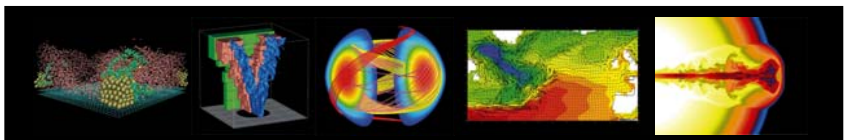
- First direct 3D simulations of a turbulent nonpremixed H2/CO–air flame with detailed chemistry. The simulations, included 11 chemical species and 33 reactions.
- Project used 11.5M MPP hours
- Generated 10TB of raw DNS data that then was analyzed.
- Investigators - Jacqueline Chen, Evatt Hawkes, and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



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.

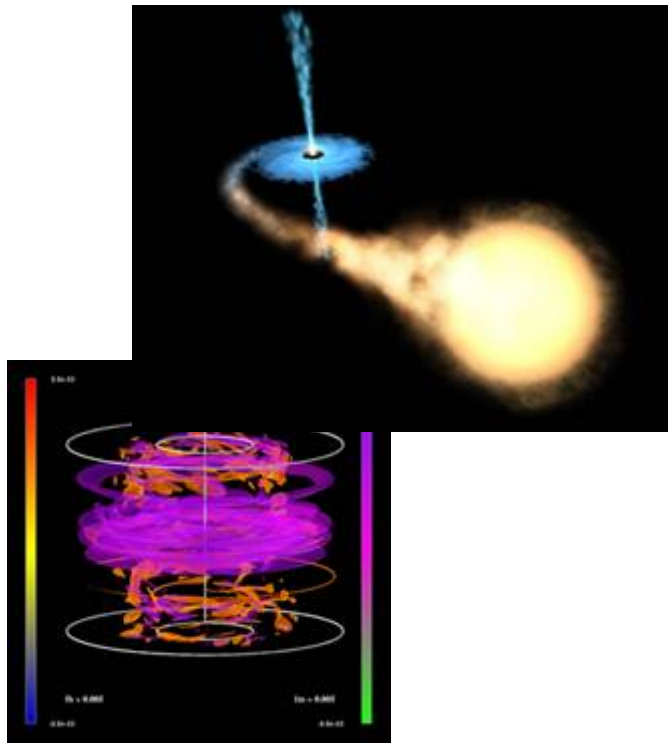


Instantaneous isocontours of the total scalar dissipation rate field for successively higher Reynolds numbers at a time when re-ignition following extinction in the domain is significant.

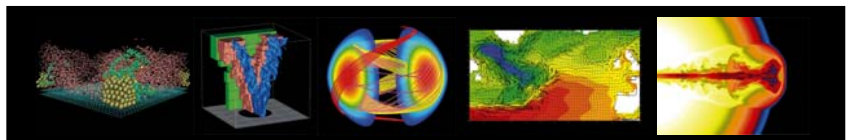


INCITE: Magneto-rotational instability and turbulent angular momentum transport

- 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

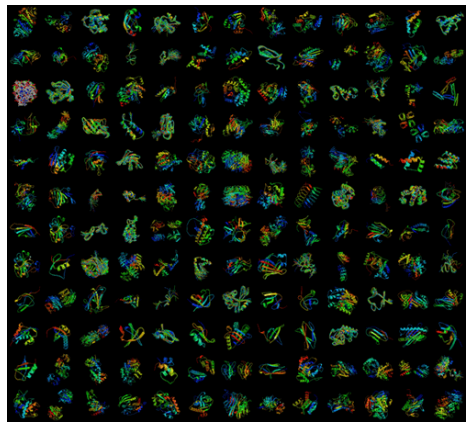


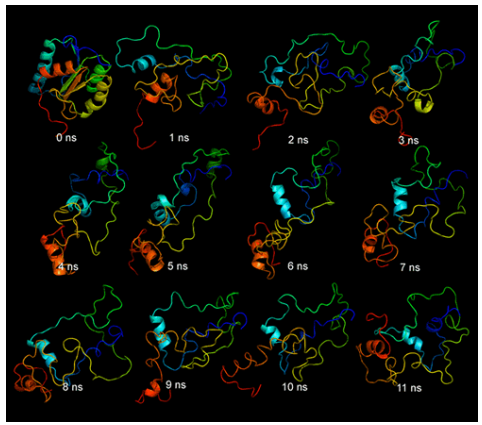
Visualization of the time evolution of the outward transport of angular momentum in a magnetic fluid bounded by rotating cylinders. The two colors correspond to the transport by hydrodynamic (orange) and hydromagnetic (purple) fluctuations.



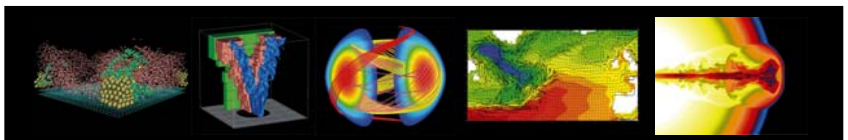
INCITE: Molecular Dynameomics

- **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**



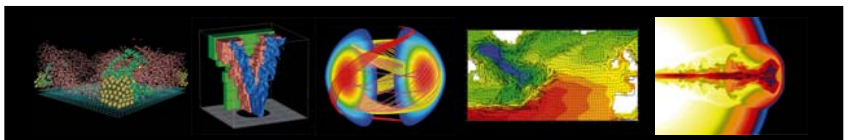


Schematic representation of secondary structures taken at 1 ns intervals from a thermal unfolding simulation of inositol monophosphatase, an enzyme that may be the target for lithium therapy in the treatment of bipolar disorder.



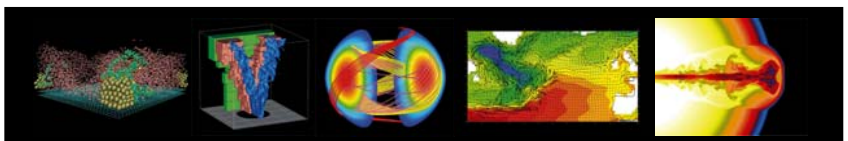
NUG Greenbook 2005
General Recommendations

- …The upcoming procurement must ensure a large increase in compute cycles available, as well as an appropriate balance of cache memory, processor memory, memory bandwidth, internode communication speed, intranode communication speed,…
- …minimize the time-to-completion of large jobs, as well as maximize the overall efficiency of the hardware. …"large" can refer to jobs requiring long running times, a large number of processors, exceptionally large memory, or any combination of these
- Significantly strengthen the computational science "infrastructure" at NERSC…local disk and archival storage, networking between NERSC's supercomputers and local storage and between NERSC and the WAN, and capabilities for local and remote data analysis and visualization must all be developed…
- …current and potential future scientific applications are especially data or I/O intensive. These requirements should be carefully evaluated in order to support as wide a range of science as possible while also realizing significant benefits in both performance and cost in the computer configuration.



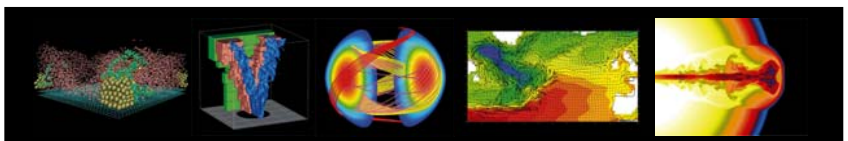
Original NERSC-5 Goals

- **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**
  - **Integrate with the NERSC Global Filesystem (NGF)**
- **Expected to significantly increase computational time for NERSC users in the 2007 Allocation Year**
  - **Dec 1, 2006 – November 30, 2007**
  - **Have full impact for AY 2008**
  - **Can arrive in FY 2006**



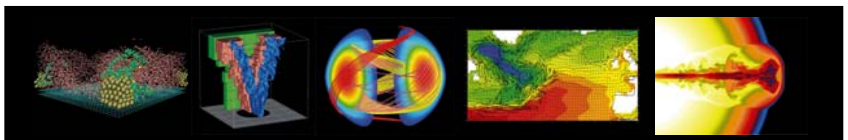
NERSC-5 Benchmarks

- **Selection of benchmarks - 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**
  - **Authors agree we can use and distribute it**
- **NERSC-5 started with approximately 20 candidates – settled on 7**



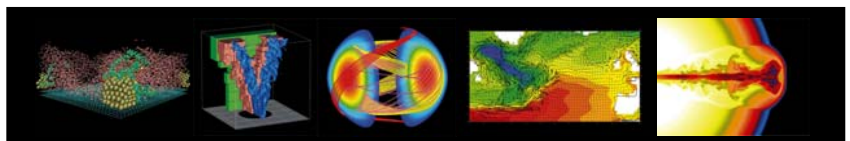
Application Summary

Application	Science Area	Basic Algorithm	Language	Library Use	Comment
CAM3	Climate (BER)	CFD, FFT	FORTRAN 90	netCDF	IPCC
GAMESS	Chemistry (BES)	DFT	FORTRAN 90	DDI, BLAS	
GTC	Fusion (FES)	Particle-in-cell	FORTRAN 90	FFT(opt)	ITER emphasis
MADbench	Astrophysics (HEP & NP)	Power Spectrum Estimation	C	Scalapack	1024 proc. 730 MB per task, 200 GB disk
MILC	QCD (NP)	Conjugate gradient	C	none	2048 proc. 540 MB per task
PARATEC	Materials (BES)	3D FFT	FORTRAN 90	Scalapack	Nanoscience emphasis
PMEMD	Life Science (BER)	Particle Mesh Ewald	FORTRAN 90	none	

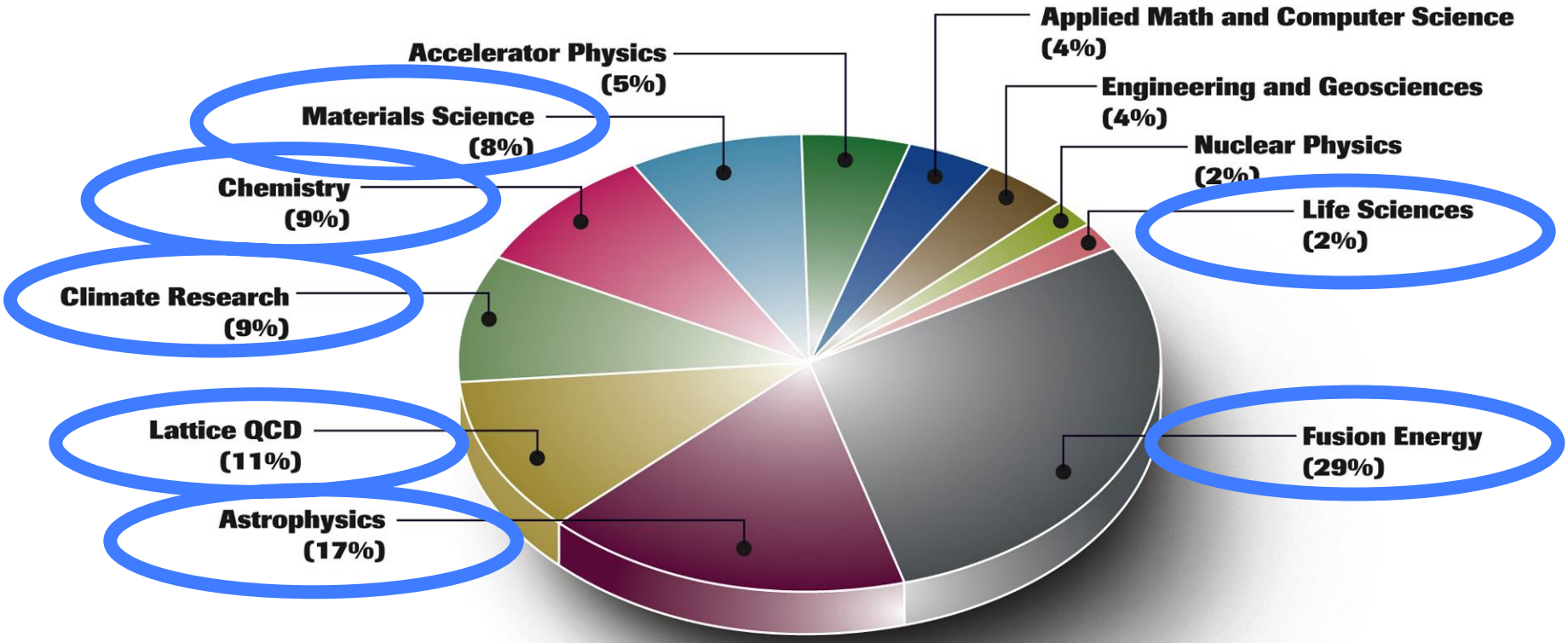


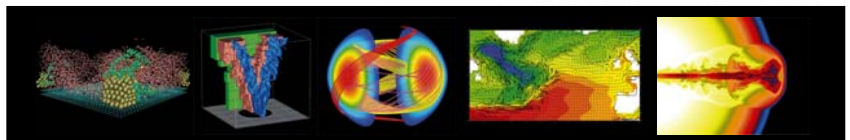
NERSC 5 Benchmarks

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



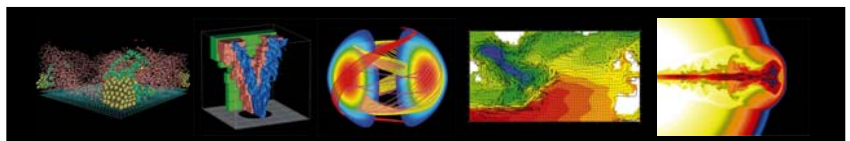
Application Benchmarks represent 85% of the Workload



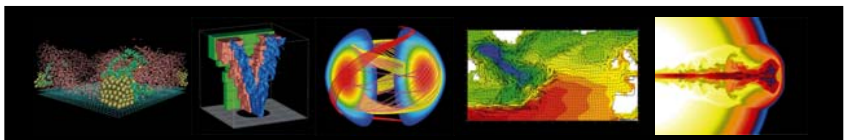


Comments

- Applications tests represent over 85% of the NERSC workload by discipline area.
- Cover most frequently used programming libraries and programming languages.
- For each benchmark, at least two test cases were prepared and validated on two or more architectures prior to the release of the RFP.
- The two tests comprised medium, run on 64 processors, and large, run on 256 processors.
  - For technical reasons, the CAM3 benchmark was run on 56 and 240 processors
  - The GAMESS L benchmark was run on 384 processors to be compatible with the DOD HPCMP TI-06 benchmark
- Extra-large benchmarks were prepared for MADbench (1024 processors) and MILC (2048 processors).

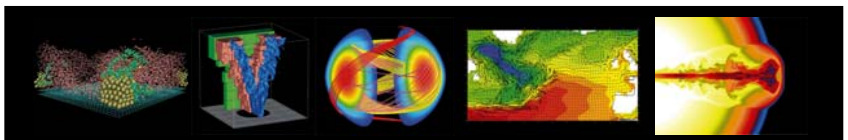


General Observations



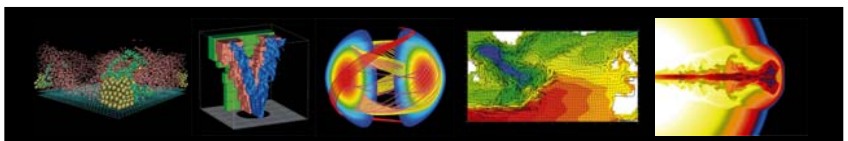
Technology Observations

- All bids were for multi core chips
  - Clock speed increasing at a much slower rate
  - The performance penalty is not as bad as we thought it might be
- Power and cooling continue to increase
  - Flop/s per $ improving faster than Flop/s per Watt or Flop/s per sf
- All proposals
  - Hybrid systems with Proprietary interconnects
  - High processor counts
  - One phase delivery - Influence of Sarbanes-Oxley?
  - Ran most to all benchmarks
  - Vertically integrated SW



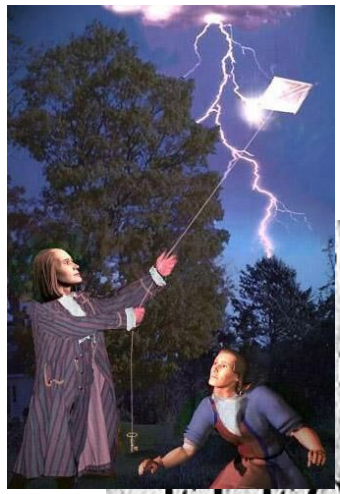
Technology Observations

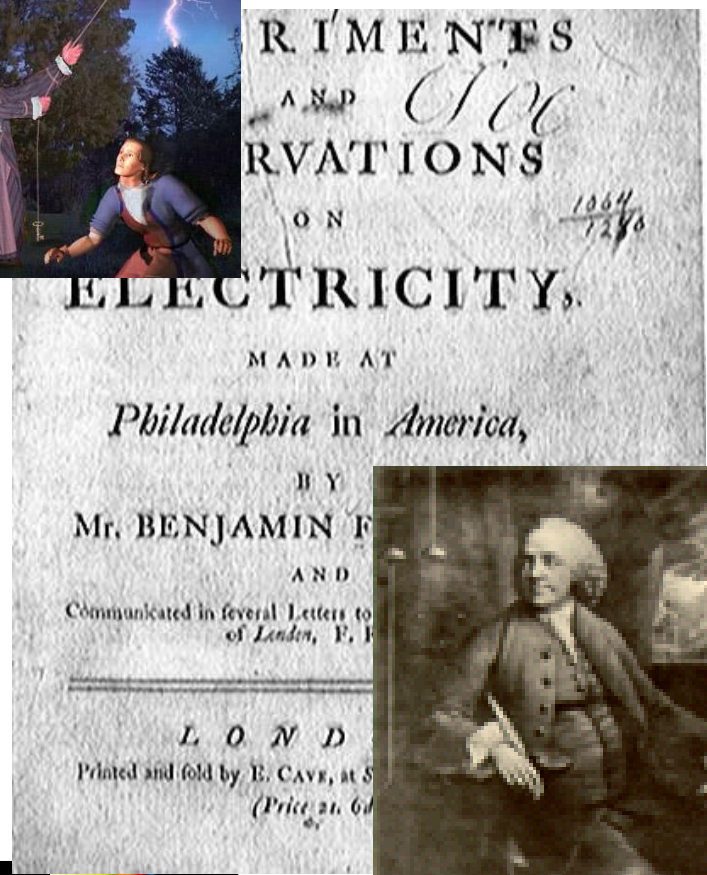
- Variety of topologies – between and within nodes
- No vector or CPU accelerated systems proposed
- Non commodity memory is very expensive
- All proposed Data Direct Networks disk
- External storage cost getting cheaper for capacity
- Delivery dates all at last moment
- All proposers can move disk drives off the single system
  - It means they all use standards compliant storage
- Declining viable bidders interested for full system and support of this size
- Is SW risk getting better? Maybe
- Efficiencies were stable and better than projected
- ESP got much better commitments
- No new technology for computer security
- No innovative technology offered

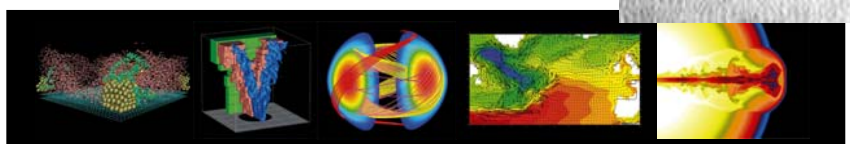


franklin.nersc.gov

- **Named after Benjamin Franklin – America's First Scientist**
- **Worked in almost every area of interest to DOE**
  - **electricity, thermal dynamics, energy efficiency, climate and global warming, ocean currents, weather, materials, population growth, medicine and health, and many other areas.**
- **We expect this system to be as productive in as many different areas of science**

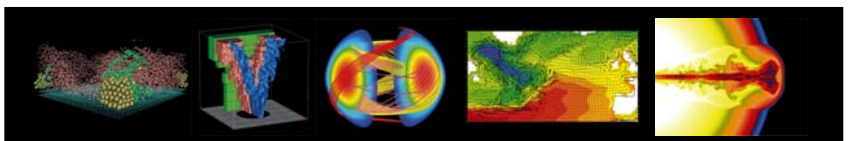






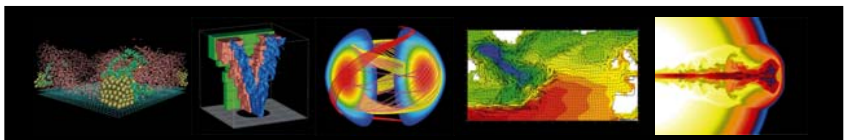
NERSC-5 - A New System

- **NERSC will field the largest Cray XT4 system.**
  - **Dual Core AMD processors at 2.6 GHz**
    - **This is a “Node” or PE**
  - **9762 Nodes = 19,524 CPUs**
    - **40 are “service node”**
  - **4 GB of memory per node**
  - **Seastar 2.1 Interconnect**
    - **A 3D Torus**
    - **Twice the injection rate as the Seastar 1.0**
    - **50 nanoseconds per hop**

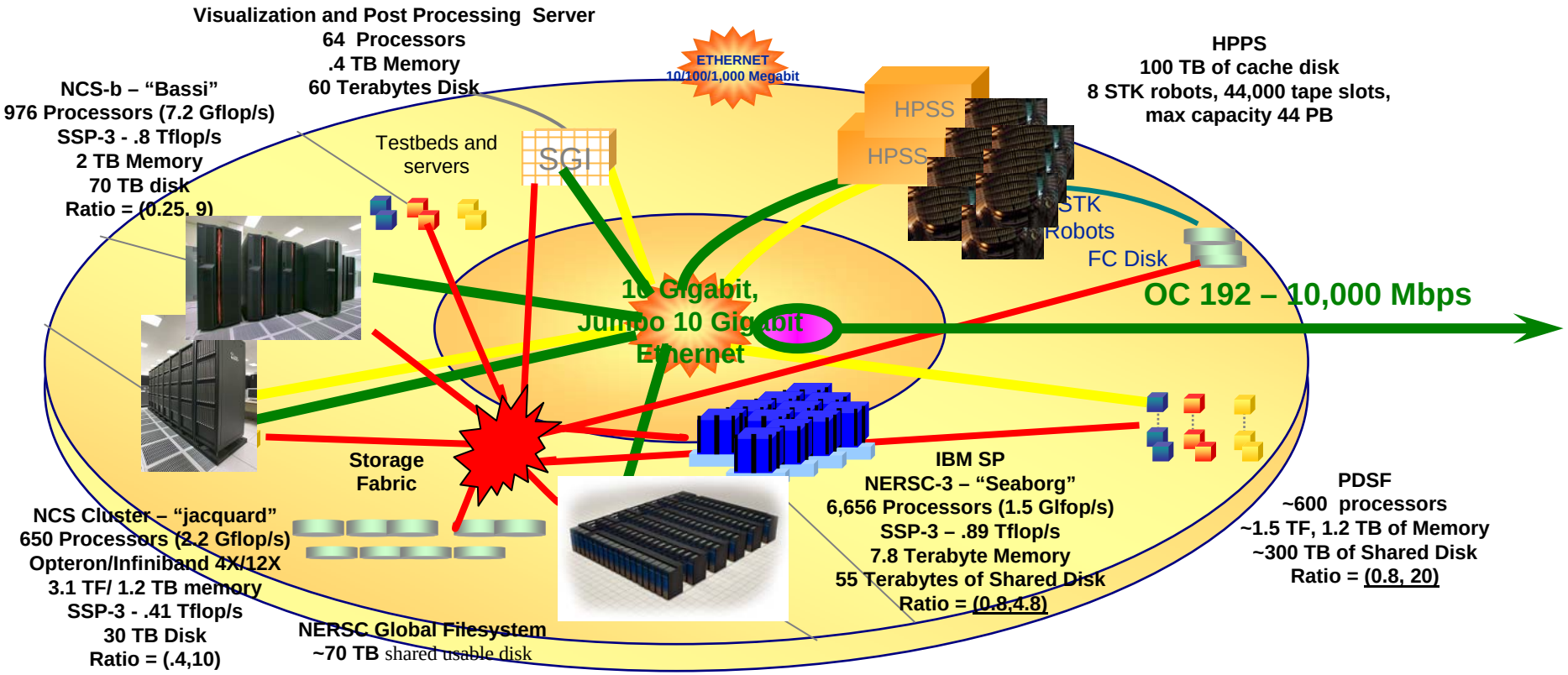


Franklin is Almost 10 Times all of NERSC's Sustained Performance!

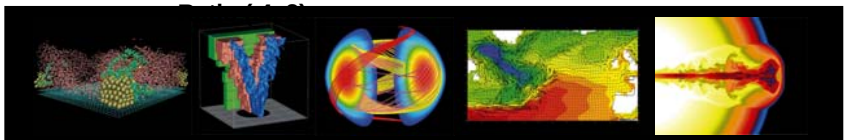
- 16.09 TF Sustained System Performance
  - Geometric Mean
  - Seaborg = .89 TF
  - Bassi ~ .8 TF
- 6.3 TB/s Bi-Section Bandwidth
  - 7.6 GB/s peak bi-directional bandwidth per link
- 402 TB of usable disk
  - DDN SA 9500 controllers with 32 tiers of 290 GB/10K RPM drives in a 8+1 Raid configuration
- 4 - 10 GigE connections
- 32 – 1 GigE connections
- 56 – 4 Gbps FibreChannel Connections



2007

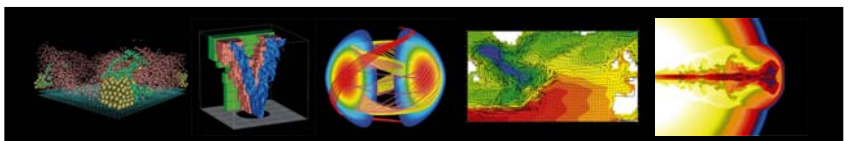


Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)



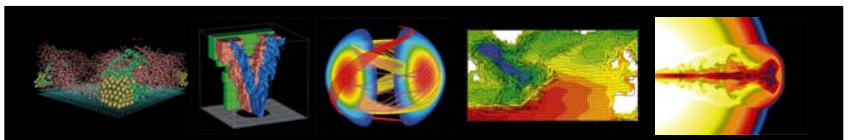
The Phasing of NERSC-5

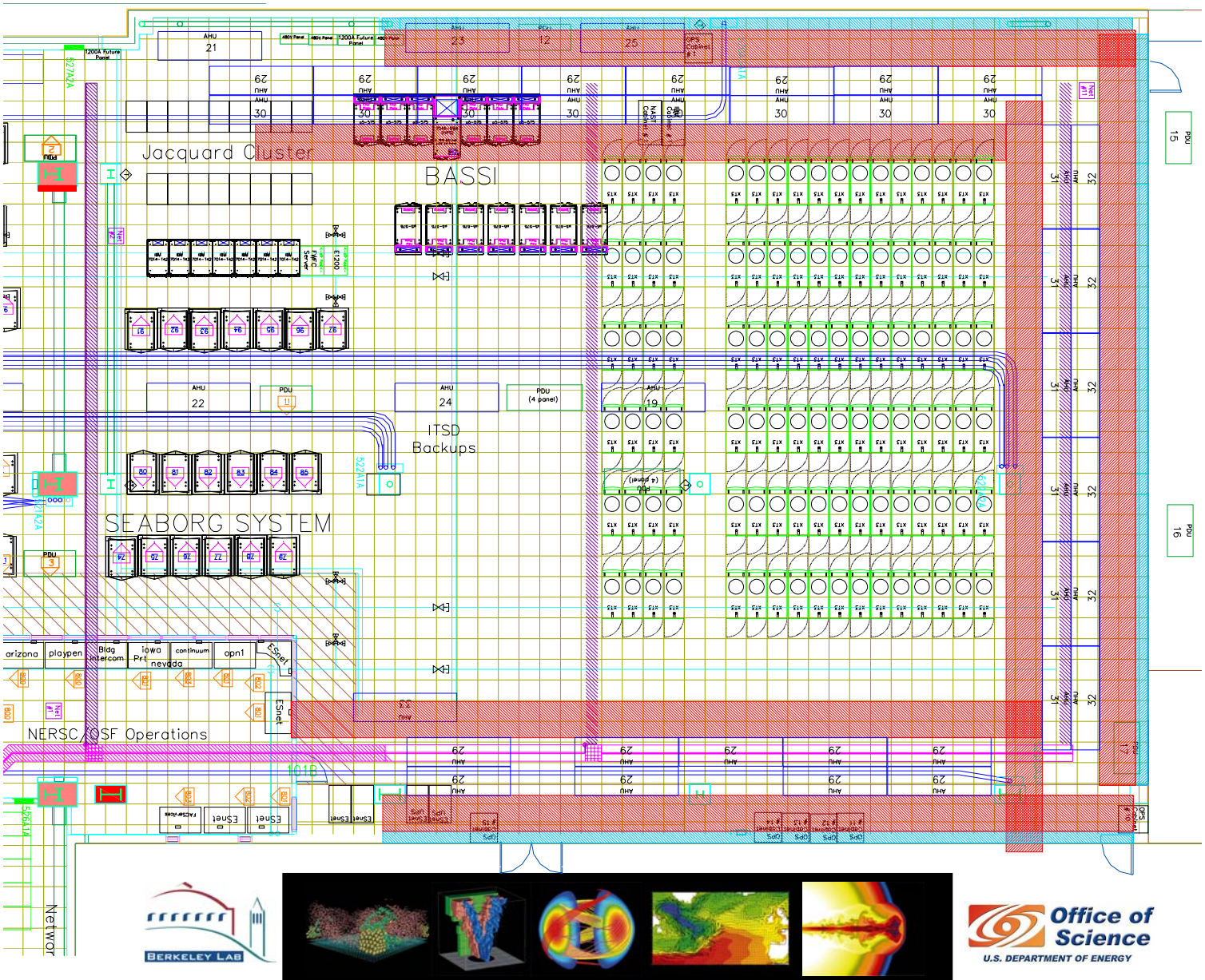
- **Small Test System**
  - **Summer 2006 – small 52 (44 compute) node XT3**
  - **Fall 2006 – upgrade to XT4**
- **January 2007 - Phase 1**
  - **36 racks**
  - **All I/O and Service Nodes**
  - **Most of the disk – 330 TB**
  - **6 x 24 x 24 Torus**
- **February 2007 – Phase 2**
  - **66 more compute rack**
  - **Most disks and controller – 402 TB total**
    - **71 TB and one controller move to NGF after Phase 2 acceptance**
  - **17 x 24 x 24 Torus**
- **Winter 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**
- **Summer 2008 – Major software upgrade**
- **Winter/Spring 2009 – *option for a 1 Petaflop/s system***



Configuration

Quantity	Type
9672	Compute Nodes – 4 GB memory
32	Spare Compute Nodes.
16	Login Nodes. Each node configured with 8 GB of memory, 1 dual port GigE Ethernet adapter (copper). And configure d with 1 Single port 4 gb/sec Fiber Channel Host Bus Adapter.
20	I/O Server nodes. Each node configured with 8 GB of memory, 2 Single port 4 gb/sec Fiber Channel Host Bus Adapter.
2	Boot Nodes. Each configured with 8 GB of memory, 1 GigE Ethernet adapter (copper) and 1 Dual port 2 gb/sec Fiber Channel Host Bus Adapter.
2	Syslog and System Database Nodes. Each configured with 8 GB of memory, 1 Dual port 2 gb/sec Fiber Channel Host Bus Adapter.
4	Network Nodes. Each configured with 8 GB of memory, 1 10 GigE Ethernet adapter (optical). And configured with 1 Single port 4 gb/sec Fiber Channel Host Bus Adapter.



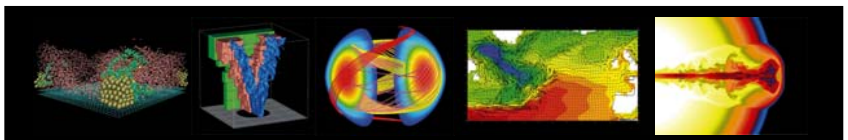


BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

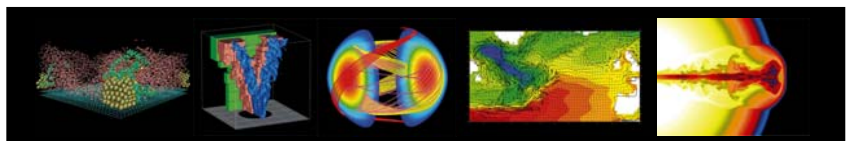
Initial Software Configuration

- SuSE SLES 9.0 Linux on Service Nodes
- Catamount (SNL) Virtual Node O/S for all compute nodes
- Portals communication layer
  - MPI, Shmem
- Lustre for scratch and some other solution for Homes
  - Homes not directly accessible by compute nodes
    - “copy in and out”
- PBS with Moab
  - Most expected functions including Backfill, Fairshare, advanced reservation
- Application Development Environment
  - PGI compilers - assembler, Fortran, C, UPC, and C++
  - Parallel programming models include MPI, and SHMEM.
  - Libraries include SCALAPACK, SuperLU, ACML, Portals 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 to 1,024 tasks
  - Craypat and Cray Apprentice
  - PAPI and Modules



Final Software Configuration

- SuSE SLES 9.0 Linux on Service Nodes
- Compute Node Linux O/S for all compute nodes
  - Cray's light weight Linux kernel
- Portals communication layer
  - MPI, Shmem
- Filesystems
  - NGF directly accessible from compute nodes with a "Petascale I/O Interface"
- PBS with Moab
  - Most expected functions including Backfill, Fairshare, advanced reservation
- Checkpoint Restart
  - Based on Berkeley Linux 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 to 1,024 tasks
  - Craypat and Cray Apprentice
  - PAPI and Modules



Questions