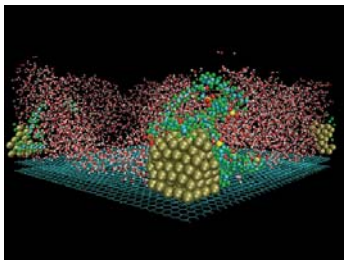
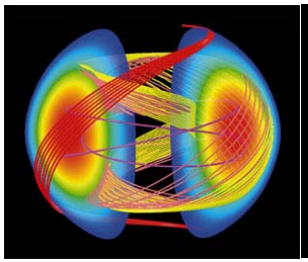
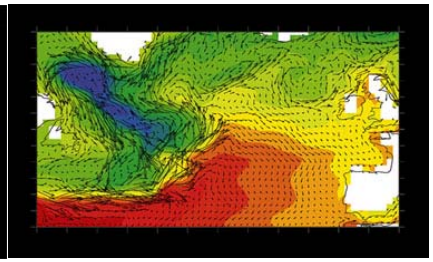
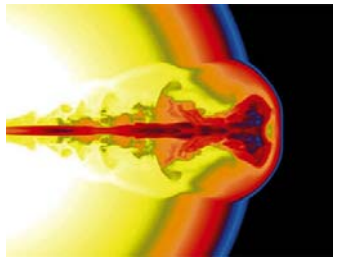
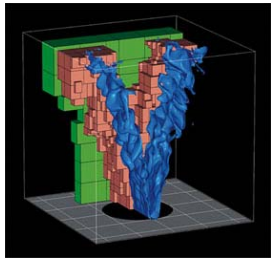








Creating Science Driven Architectures



William T.C. Kramer
NERSC
kramer@nersc.gov
(510) 486-7577

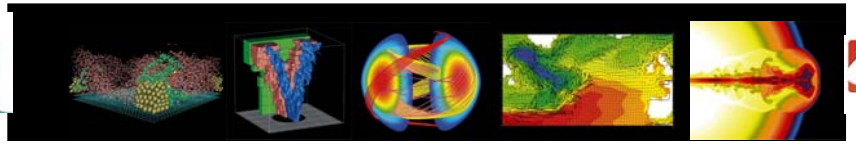
NERSC

BERKELEY LAB

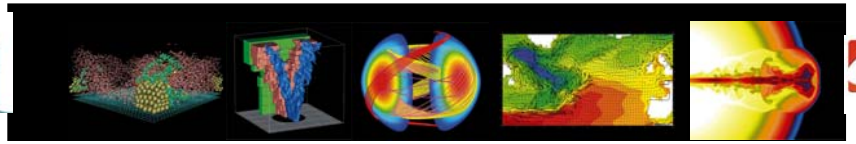
Office of Science
U.S. DEPARTMENT OF ENERGY

Agenda

- **Brief Update on NERSC and Plans**
- **The Science Driven Architecture Process**

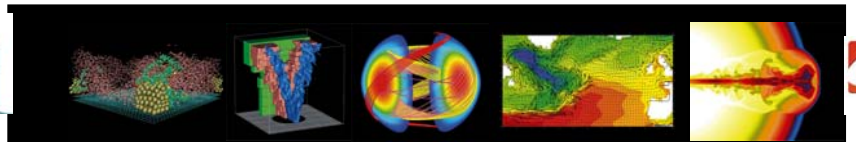


NERSC Update and Plans

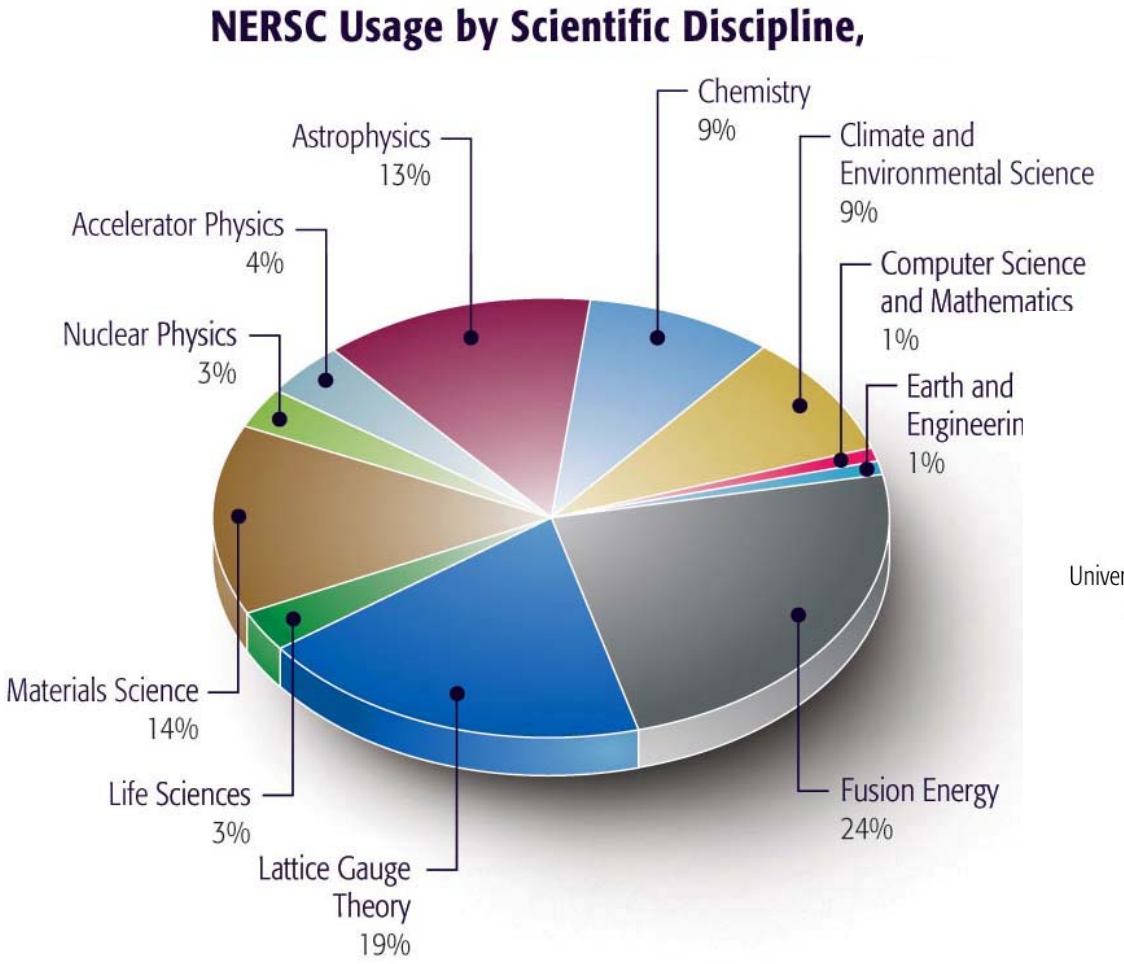


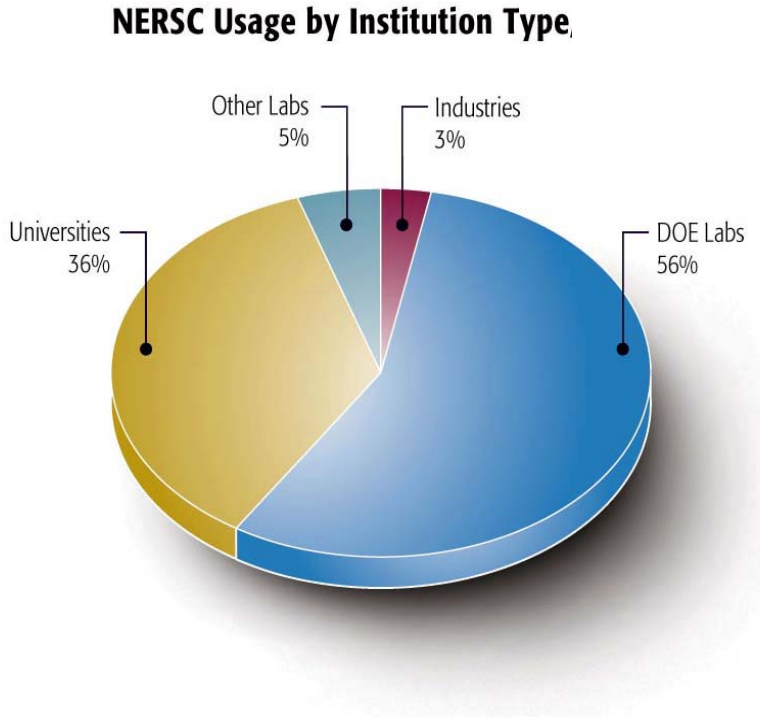
NERSC is Flagship Computational Facility for DOE Office of Science

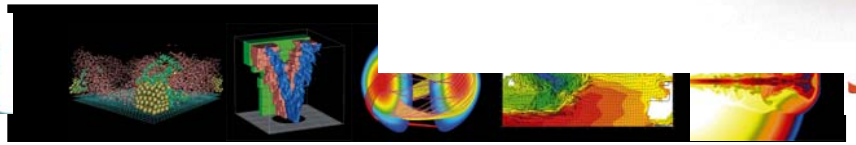
- NERSC is a user facility for the DOE Office of Science
  - Provides leading edge platforms and services
  - Introduces best new computational science tools
  - Provides high quality "intellectual" services that make complicated technologies useful
- More than 2,400 users nationwide with several hundred projects
- Focus on capability science: ~ 50% of cycles for large jobs
- Terascale computational systems, petabyte storage, gigabit networks
- NERSC provides resources to all researchers regardless of organization or funding agency
- NERSC is totally open – with no restricted or classified work.



NERSC Supports A Diverse Science Community

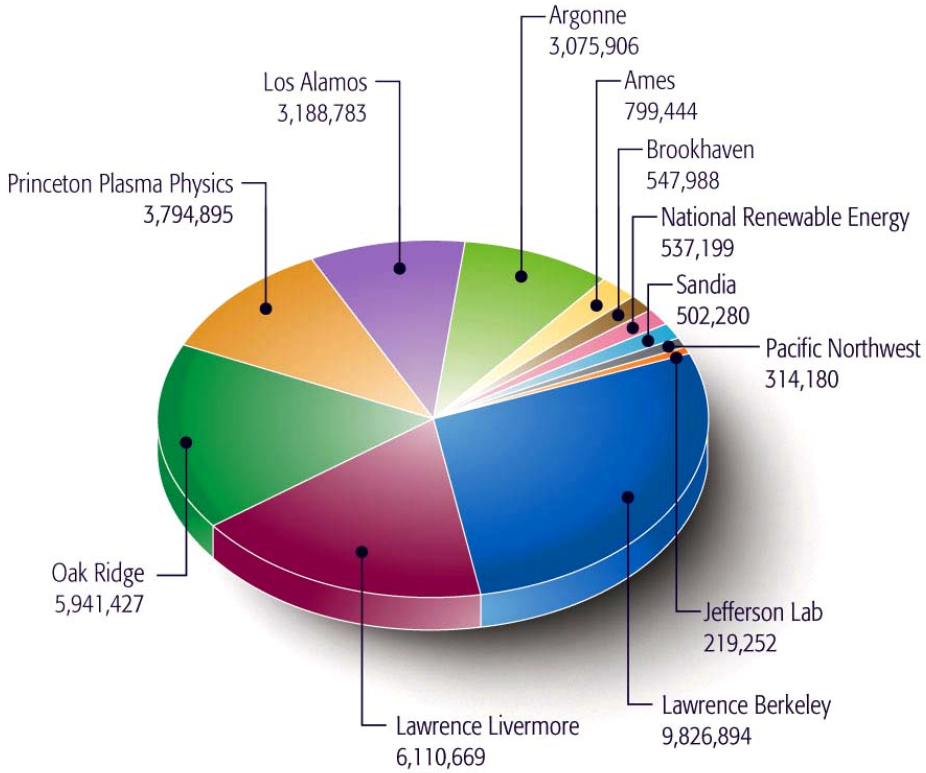




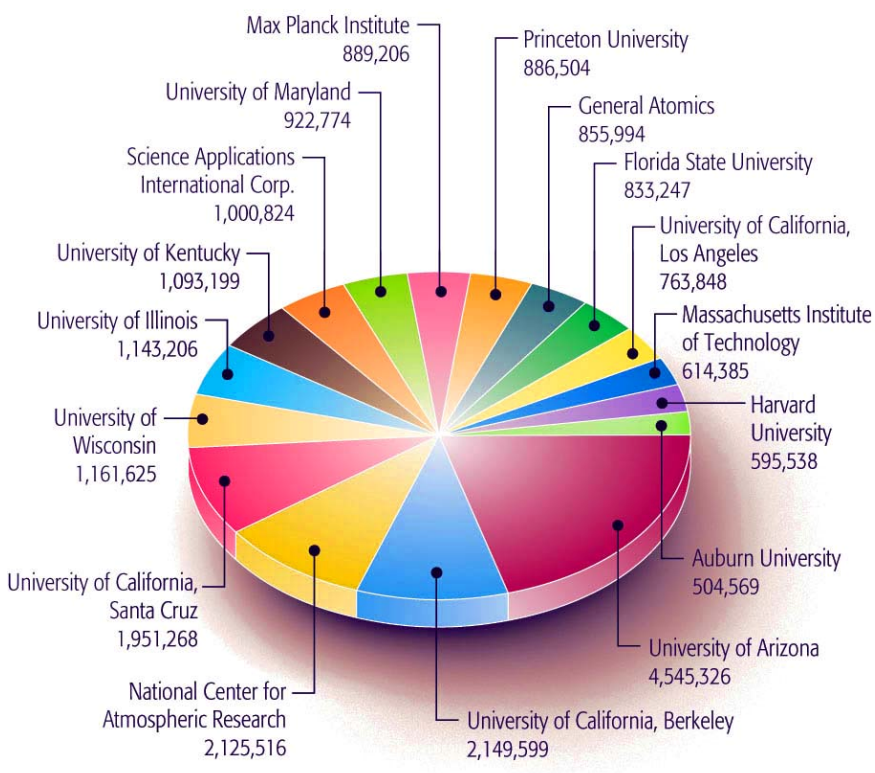


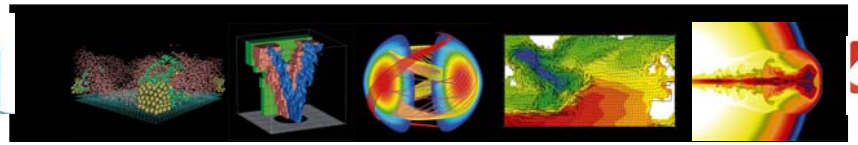
NERSC is a National Facility

Leading DOE Laboratory Usage at NERSC, FY02
(>200,000 processor hours)

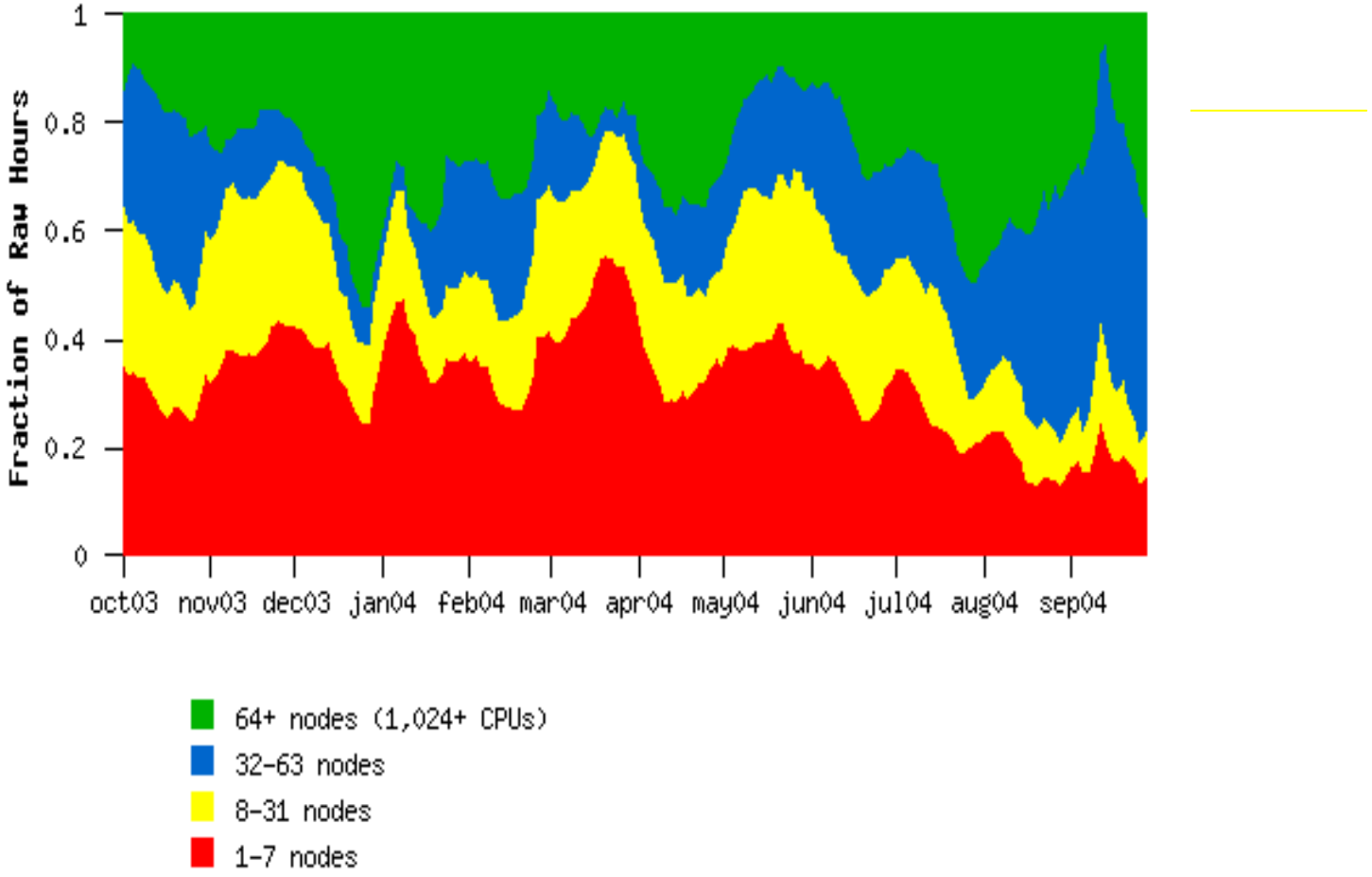


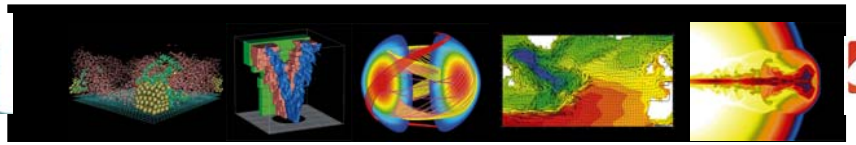
Leading Academic and Related Usage at NERSC, FY02
(>500,000 processor hours)



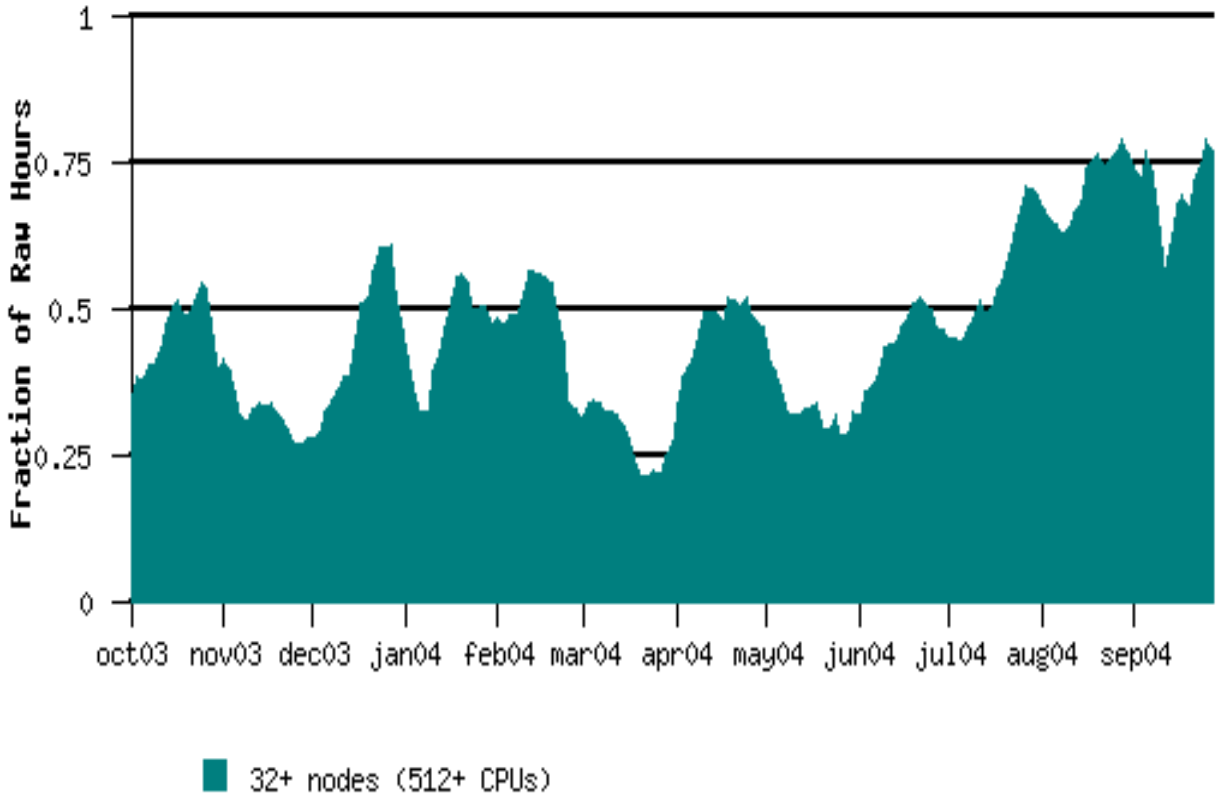


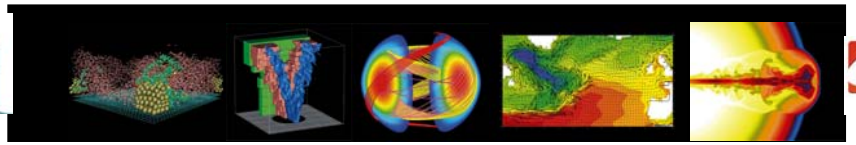
Large Jobs
Run Regularly





Large Jobs
Run Regularly

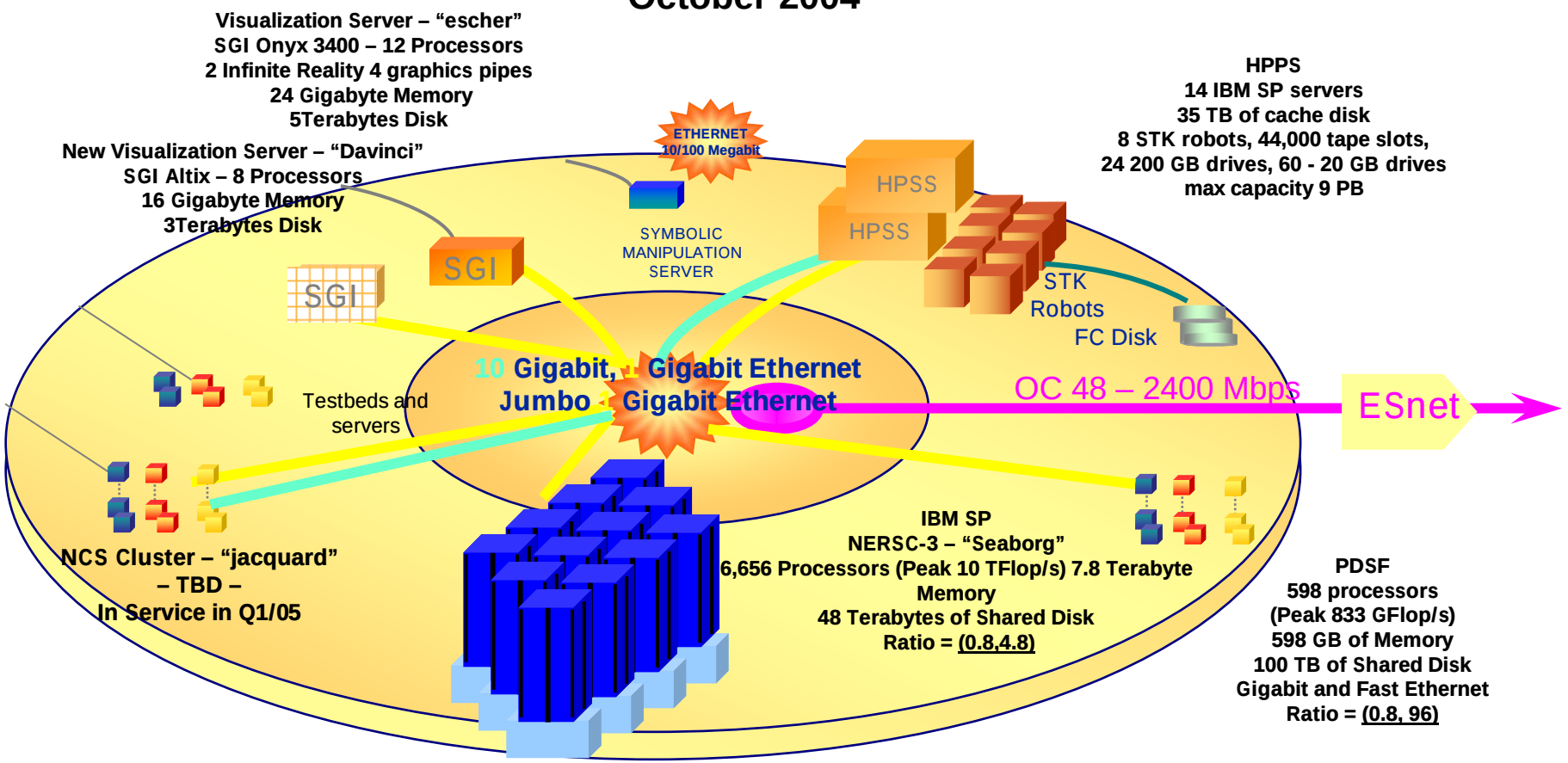




Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC System Architecture

October 2004



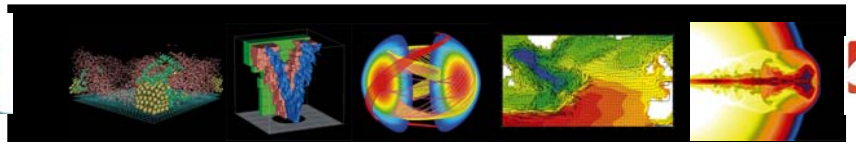
Flop, Disk Bytes per Flop)

NERSC – A Successful User Facility

- In 2003, NERSC users reported 2,404 peer reviewed papers in 199 different projects that were published based, at least in part, on work done at NERSC.
- NERSC is the only large scale center that does production computing for all areas of science for all research organizations in the US
- Excellent ratings by NERSC users via user survey, regular user meetings, peer reviews, etc.
- NERSC is measured by the DOE and outside organizations

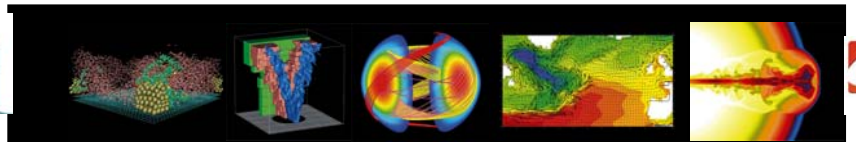
"NERSC simply is the best-run centralized computer center on the planet. I have interacted with many central computer centers and none are as responsive, have people with the technical knowledge available to answer questions, and have the system/software as well configured as does NERSC."

—2003 NERSC User Survey Respondent



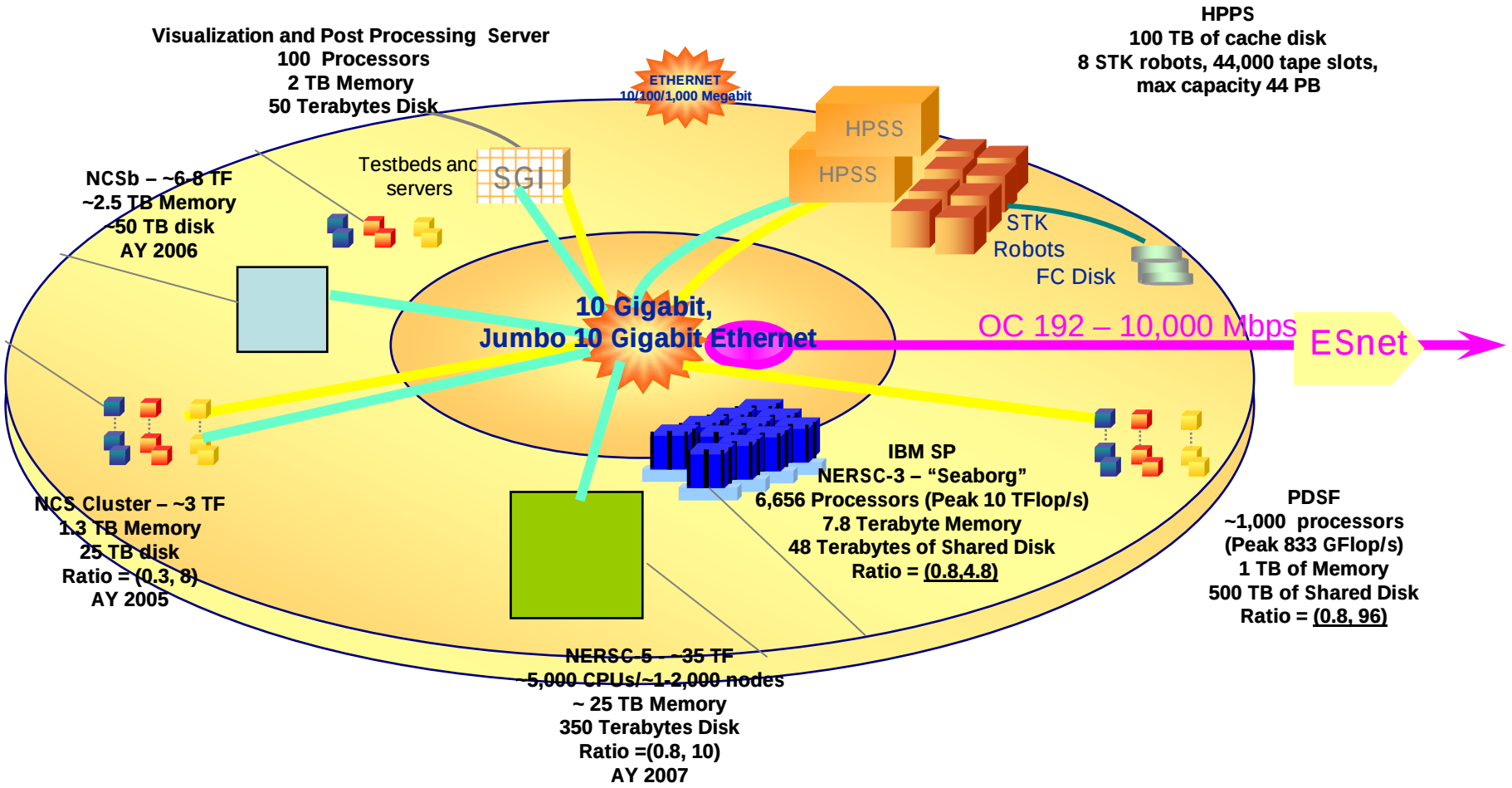
NERSC's Nominal Facility Improvement Plans

- Two major systems in place at a time – with modest size systems arriving in between the major systems
- NERSC – 5: 2006 initial delivery with possibly a phasing of delivery
  - 3 to 4 times Seaborg in delivered performance
  - Used for entire workload and has to be balanced
- NERSC-6: 2009 - initial delivery
  - 3 to 4 times NERSC-5 in delivered performance
  - Used for entire workload and has to be balanced
- NCS and NCS b
  - Moderate, focused systems that come between major systems
  - 2005 - >15 - 20% of seaborg
  - 2006 – >30 - 40% of seaborg
- HPSS and Network will scale in proportion to computational systems
- PDSF – continue to double every year in processing power
- Servers – Visualization, specialized work, etc. improve as needed

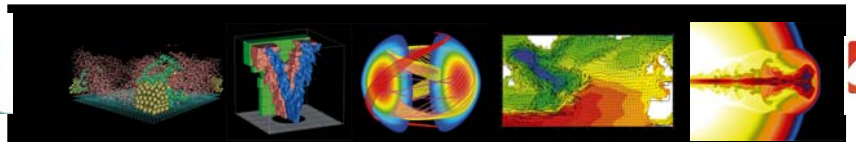


Future NERSC System Architecture

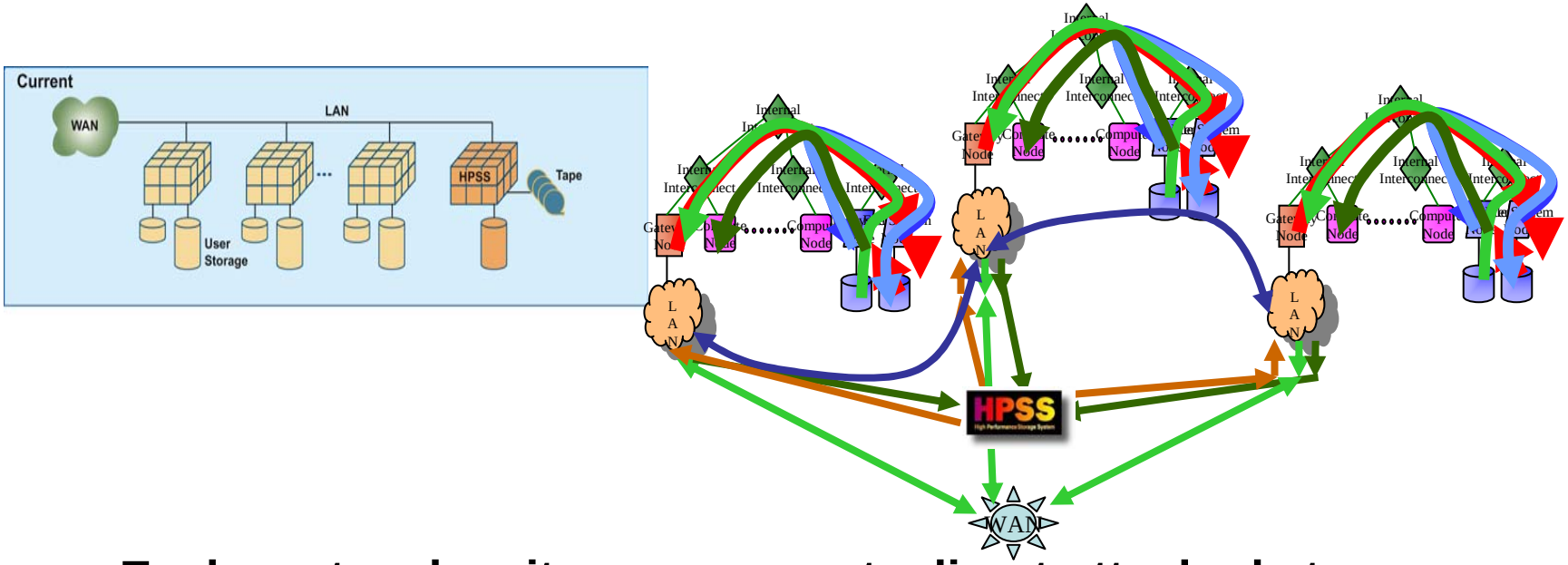
October 2007 Projection



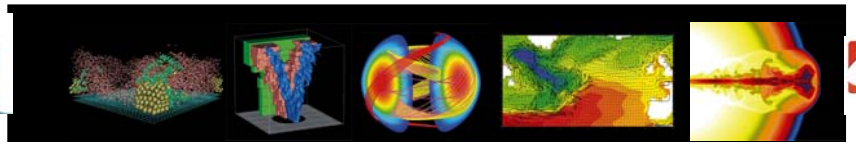
Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)



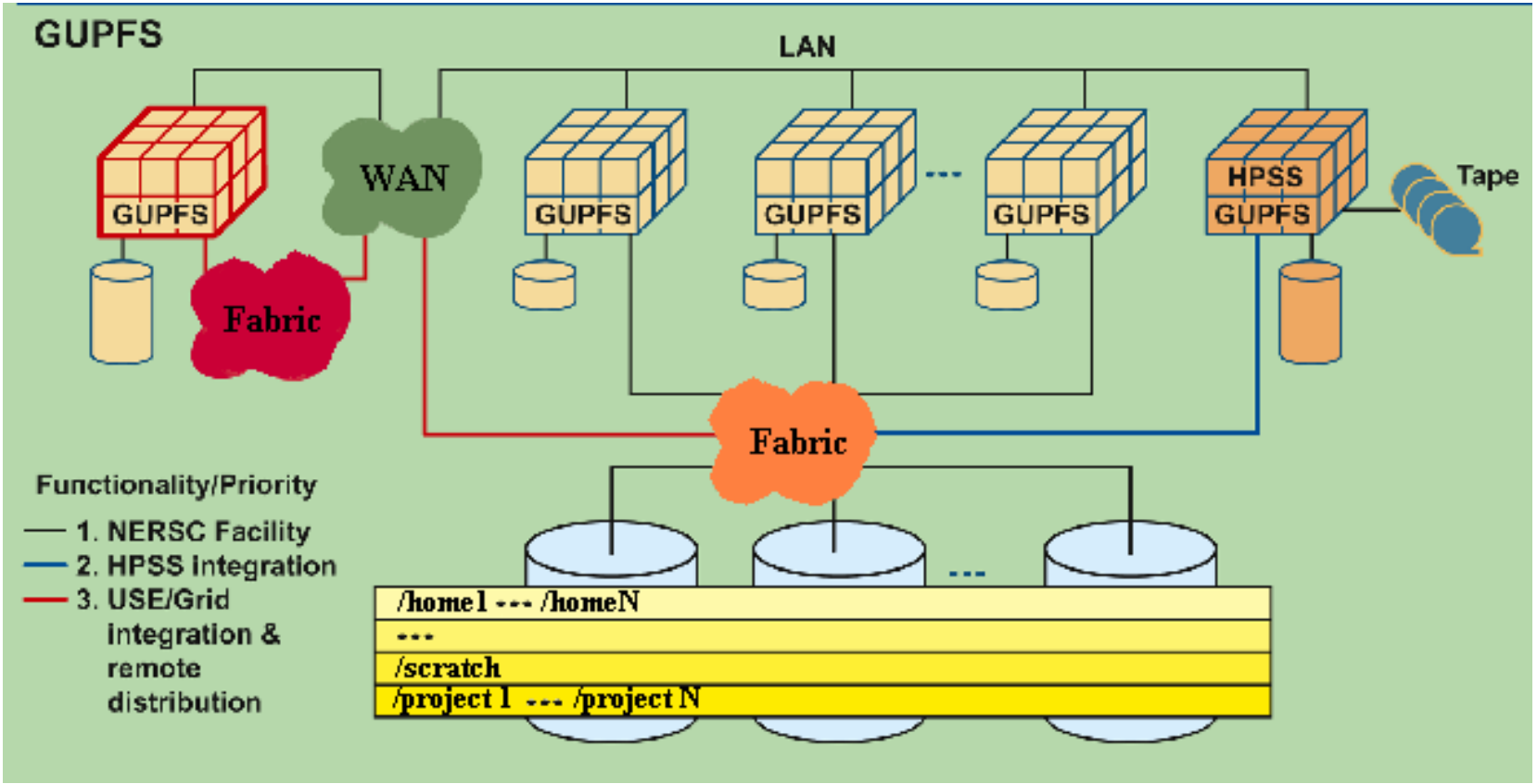
But Storage is not Scaling unless something new is done

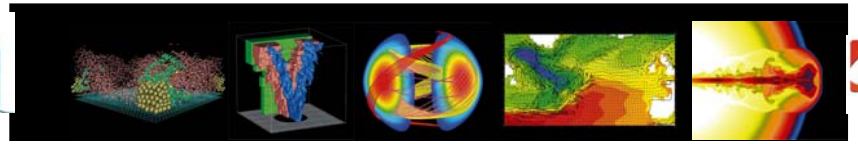


- Each system has its own separate direct-attached storage
- Each system has its own separate user file system and name space
- Data transfer between systems is over the LAN
- Includes large computational systems, small systems, and support systems



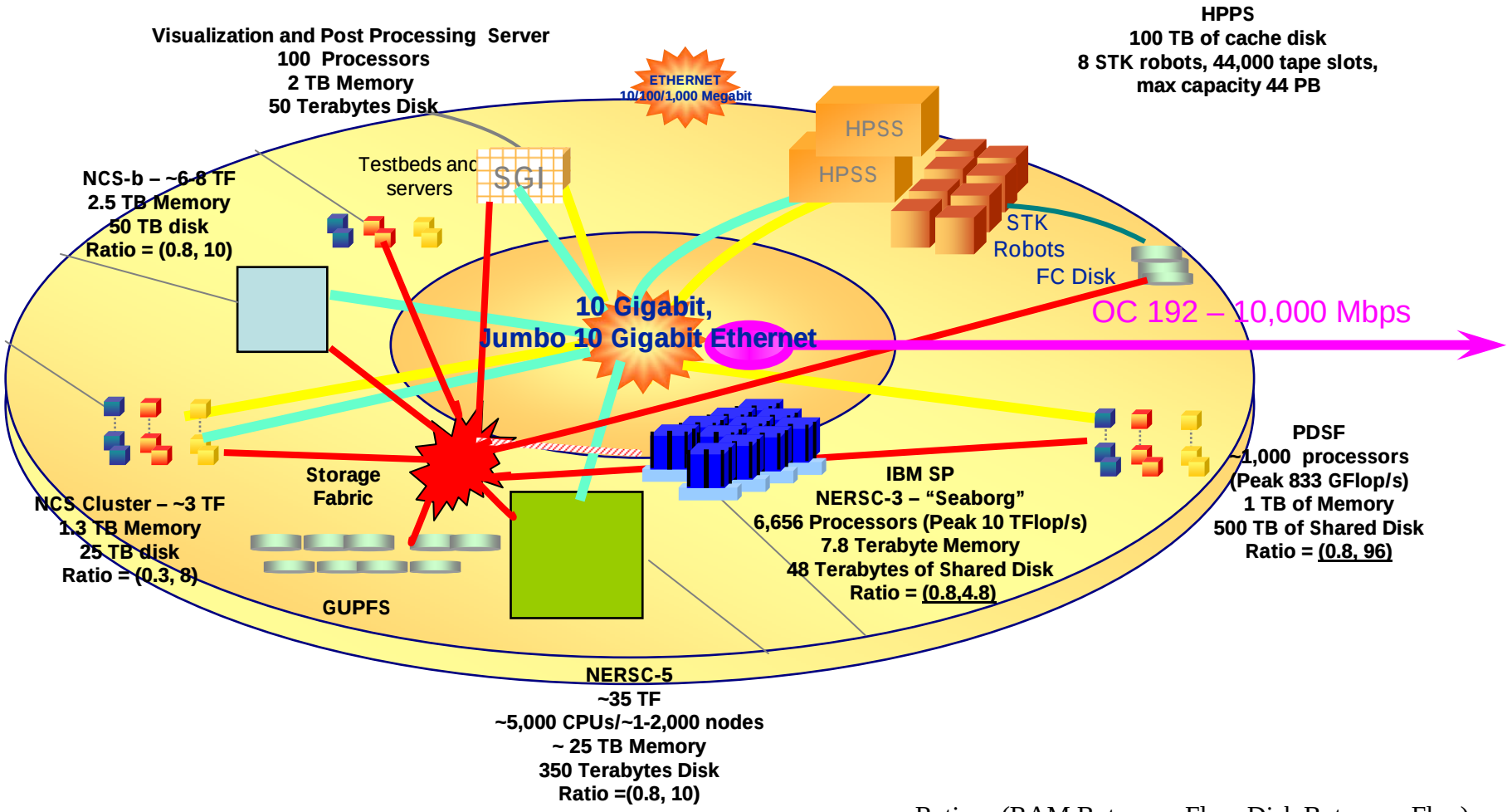
Envisioned NERSC Storage Configuration (GUPFS)



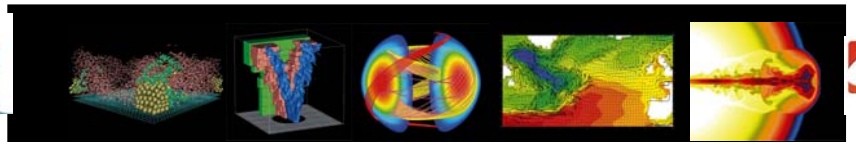


Future NERSC System Architecture with GUPFS

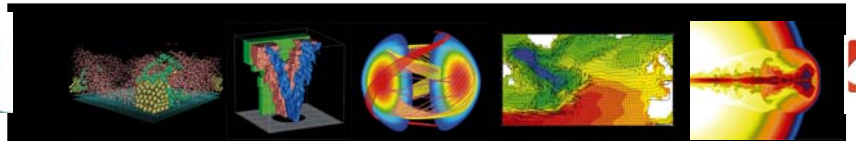
October 2007 Projection



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

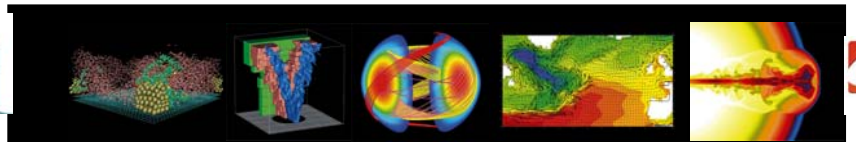


Science Driven Architectural Design



Science-Driven Technology
Recent Trends

- **Computational System components designed for applications other than high performance computing (web servers, desktop applications, databases, etc.)**
- **More cost pressure to move to commodity components – essentially to take things "as is"**
- **The general acknowledgement that Moore's Law scaling for increase clock rate is ending**
  - **Moore's Law, as we think of it, is a very simple 2D extrapolation.**
  - **The challenge to building faster and more powerful CPUs are essentially in the third dimension and that is no longer scaling**
    - **Dr. Bernard Meyerson – IBM**
    - **Dr. Shekhar Borkhar –Intel**



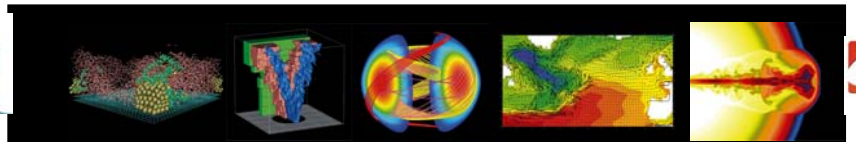
Science-Driven Technology

Result:

- **Scientists are increasingly frustrated by low sustained performance rates on real-world, high-end problems**

What is needed:

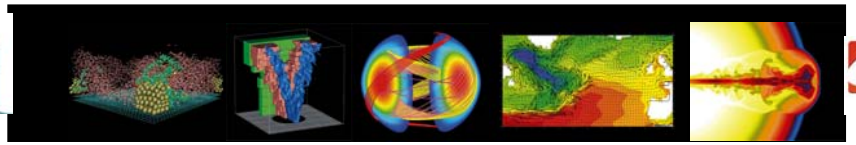
- **A science-driven system architecture that achieves high sustained performance on a broad range of scientific applications but also leverages commodity components.**



A Sustainable Path to HEC Leadership

HEC investments must lead to a widely deployable technology for high end computing – not just purchase a few "big" machines

Improved technology will not appear just by itself – sites need to take an active approach with vendors to develop the best possible solutions. This is a change for both sides.

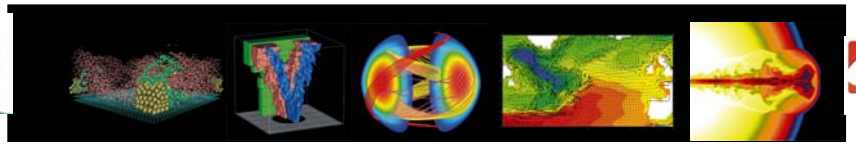


A Sustainable Path to HEC Leadership (cont.)

HECRTF, SCaLeS, and OASCR strategic plan all conclude that scientific applications must influence machine design in a repeating cycle

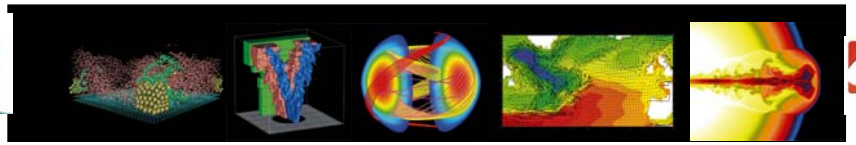
"We must develop a government-wide, coordinated method for influencing vendors. The HEC influence on COTS components is small, but it can be maximized by engaging vendors on approached and ideas five years or more before commercial products are created. Given these time scales, the engagement must also be focused and sustained. We recommend that academic and government HEC groups collect and prioritize a list of requested HEC-specific changes for COTS components, focusing on an achievable set."

–HECRTF Workshop Report, June 2003, (ed. Dan Reed), p. 23



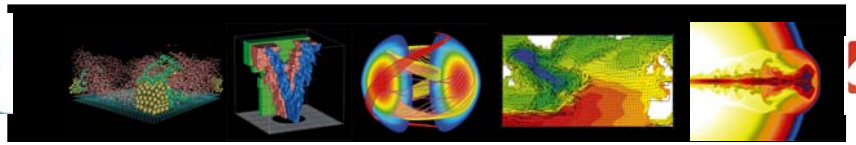
A Sustainable Path to HEC Leadership (cont)

- Requires partnerships with vendors with resources and an R&D track record AND a national partnership of laboratories, computing facilities, universities, and computational scientists
- There has to be clear understanding of the limitations of current architectures and exploration for improvement



Some Potential SDCA Science Breakthroughs

Science Areas	Goals	Computational Methods	Breakthrough Target (50Tflop/s)
Nanoscience	Simulate the synthesis and predict the properties of multi-component nanosystems	Quantum molecular dynamics Quantum Monte Carlo Iterative eigensolvers Dense linear algebra Parallel 3D FFTs	Simulate nanostructures with hundreds to thousands of atoms as well as transport and optical properties and other parameters
Combustion	Predict combustion processes to provided efficient, clean and sustainable energy	Explicit finite difference Implicit finite difference Zero-dimensional physics Adaptive mesh refinement Lagrangian particle methods	Simulate laboratory scale flames with high fidelity representations of governing physical processes
Fusion	Understand high-energy density plasmas and develop an integrated simulate of a fusion reactor	Multi-physics, multi-scale Particle methods Regular and irregular access Nonlinear solvers Adaptive mesh refinement	Simulate the ITER reactor
Climate	Accurately detect and attribute climate change, predict future climate and engineer mitigations strategies	Finite difference methods FFTs Regular & irregular access Simulation ensembles	Perform a full ocean/atmospheres climate model with 0.125 degree spacing, with an ensemble of 8-10 runs
Astrophysics	Determine through simulations and analysis of observational data the origin, evolution and fate of the universe, the nature of matter and energy, galaxy and stellar evolutions	Multi-physics, multi-scale Dense linear algebra Parallel 3D FFTs Spherical transforms Particle methods Adaptive Mesh Refinement	Simulate the explosion of a supernova with a full 3D model



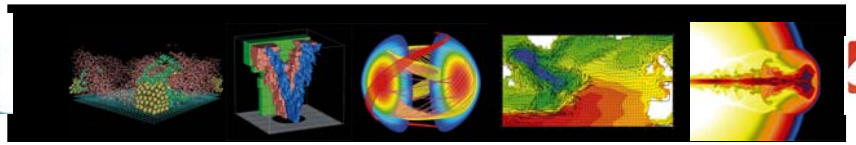
How Science Drives Architecture

State-of-the-art computational science requires increasingly diverse and complex algorithms

Science Areas	Multi Physics and Multi-scale	Dense Linear Algebra	FFTs	Particle Methods	AMR	Data Parallelism	Irregular Control Flow
Nanoscience	X	X	X	X		X	X
Combustion	X			X	X	X	X
Fusion	X	X		X	X	X	X
Climate	X		X		X	X	X
Astrophysics	X	X	X	X	X	X	X

Only balanced systems that can perform well on a variety of problems will meet future scientists' needs!

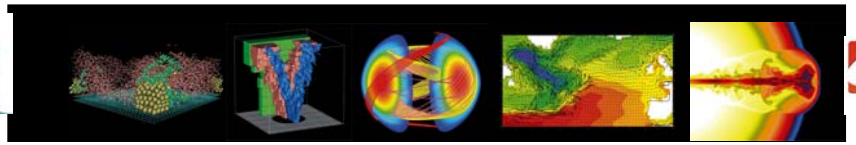
Data-parallel and scalar performance are both important



New Science Presents New Architectural Challenges

Leadership computing requires an architecture capable of achieving high performance across a spectrum of key state-of-the-art applications.

- **Data parallel algorithms do well on machines with high memory bandwidth (vector or superscalar)**
- **Irregular control flow requires excellent scalar performance**
- **Spectral and other methods require high bisection bandwidth**



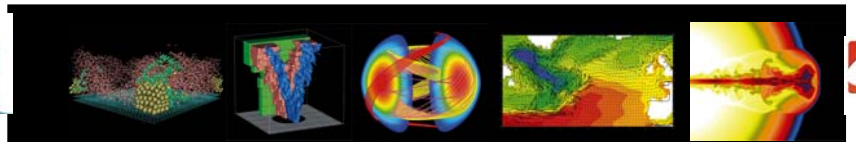
NERSC BERKELEY LAB Office of Science U.S. DEPARTMENT OF ENERGY

Scalar Performance Increasingly Important

- **Cannot use dense methods for largest systems because of N^3 algorithm scaling. Need to use sparse and adaptive methods with irregular control flow**
- **Complex microphysics results in complex inner loops**

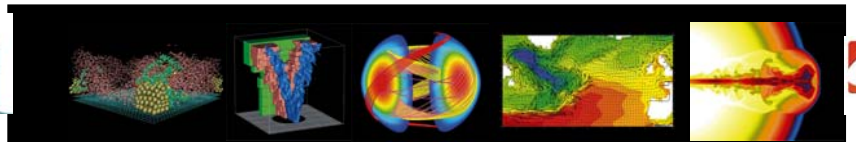
"It would be a major step backward to acquire a new platform that could reach the 100 Tflop level for only a few applications that had 'clean' microphysics. Increasingly realistic models usually mean increasingly complex microphysics. Complex microphysics is not amenable to [simple vector operations]."

– Doug Swesty, SUNY Stony Brook



SDCA Goals

- **Broadest, large-scale application base runs very well on SDCA solutions with excellent *sustained* performance per dollar**
- **Even applications that do well on specialized architectures should performance relatively near optimal on a SDCA Architectures**



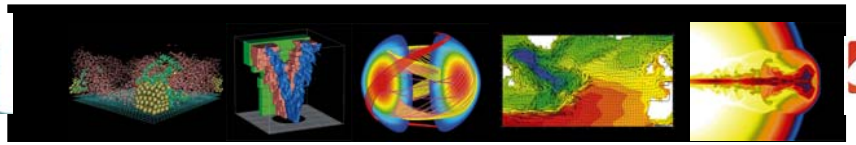
SDCA Roles

Technology development

- Implementing the science-driven architecture development
- Dialogue with vendor partner

National facility operations

- Establish close connections and strategic collaborations with computer science programs and facilities funded by federal agencies and with universities



SCDA Result Example: LBNL Blue Planet

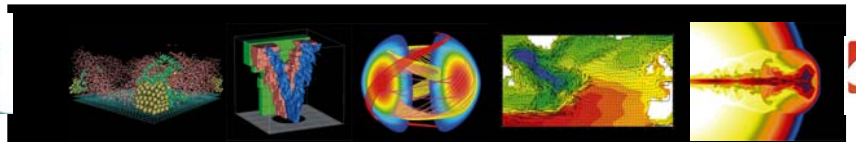
2002: Berkeley Lab launches science-driven architecture process

2003: Multiple design discussions, reviews with scientists and computer architects at Berkeley Lab, LLNL, IBM

2004: IBM incorporates Blue Planet node design and enhanced interconnect in product roadmap and will deliver first implementation to ASCI program

"The Blue Planet node conceived by NERSC and IBM [...] features high internal bandwidth essential for successful scientific computing. LLNL elected early in 2004 to modify its contract with IBM to use this node as the building block of its 100 TF Purple system. This represents a singular benefit to LLNL and the ASC program, and LLNL is indebted to LBNL for this effort."

– Dona Crawford, Associate Director for Computation, LLNL



Science-Driven Computer Architecture Impact

IBM is committed to science-driven design process

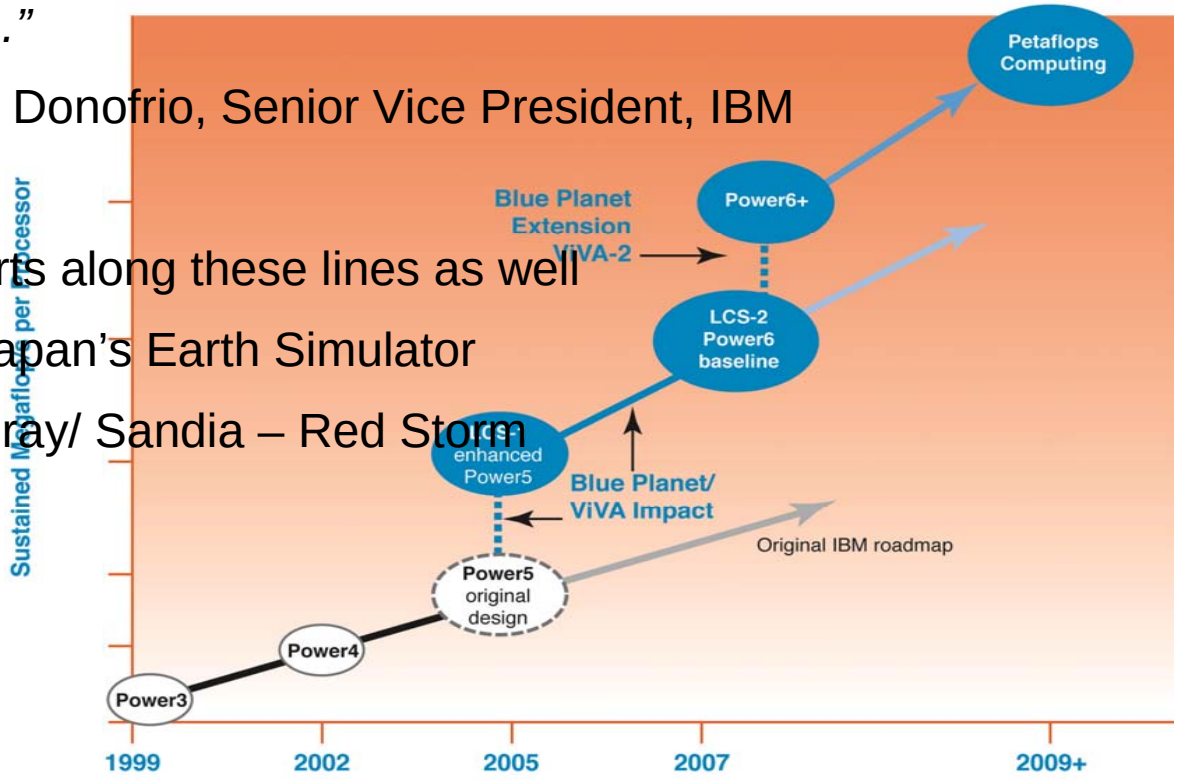
"IBM has already adopted the concepts of 'Science Driven Architecture Design' in redesigning the Power 5/6 node. We will continue the Science Driven Design approach..."

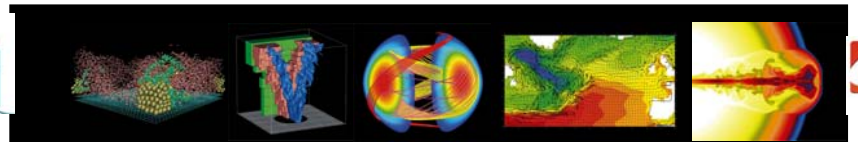
– Nicholas Donofrio, Senior Vice President, IBM

Other Efforts along these lines as well

Japan's Earth Simulator

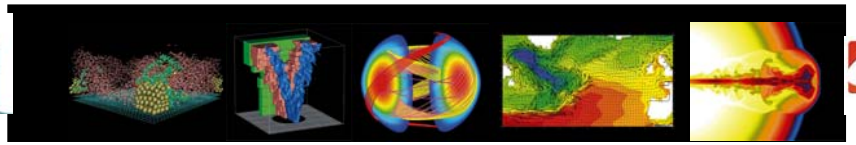
Cray/ Sandia – Red Storm





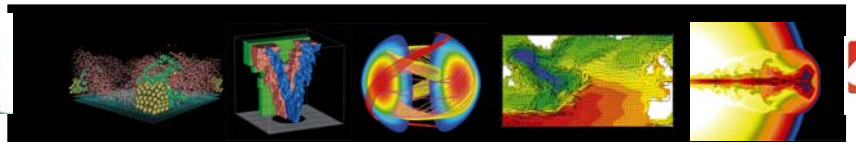
The Path to Petascale Computing

- Collaboration between scientists and computer vendors on science-driven architecture is the path to continued improvement in application performance.
- NERSC staff and users are representing the scientific application community to continue the Science-Driven Computer Architecture process.
- Initial objectives:
  - ViVA-2 architecture development – Power6 scientific application accelerator
  - Additional investigation with other architectures
  - BlueGene and other collaboration and evaluation
  - Lower interconnect latency and large spanning
  - Involvement with HPCS and other efforts
- Long-term objective: integrate the lessons of the large scale systems, such as the Blue Gene/L and HPCS experiments with other technologies into a hybrid system for petascale computing.



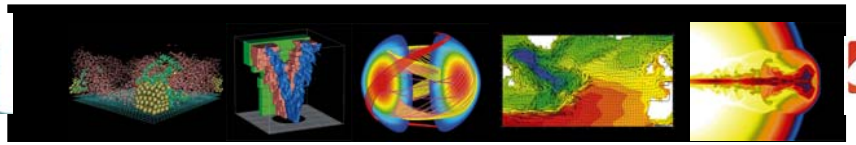
NERSC Efforts for SCDA

- **Workload Analysis**
  - **Difficult to do for diverse science – which is what NERSC runs and vendors are interested in**
  - **Explicit scaling analysis -**
- **Performance Collection**
  - **poe+ data collection**
    - **Expands IBM HW performance collection**
    - **Being converted to generic code**
    - **Required in proposals for allocations**
- **System modeling**



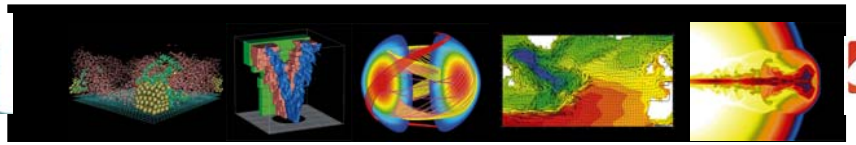
NERSC Expertise is Critical to SDCA Process

- **Samples of Algorithms and Computational Science Successes**
  - **Numerical and System Libraries – SuperLU, Scalapack, MPI-2, parallel NETcdf**
  - **Applications and Tools - ACTs Toolkit**
  - **Programming Languages - UPC**
  - **System Software - Linux Checkpoint/Restart, VIA Protocol**
  - **APDEC/NASA Computational Technology Center**
- **Performance evaluation and analysis**
  - **LBNL staff includes authors of widely used evaluation tools: NAS Parallel Benchmarks (NPB), Sustained System Performance (SSP) benchmark, Effective System Performance (ESP) benchmark**
  - **The "Performance Evaluation Research Center" (PERC), a multi-institution SciDAC project funded by DOE.**
  - **Tuning and analysis of dozens of applications on NERSC scalar and vector systems.**
- **Architecture evaluation and design**
  - **Multi-application study of the Earth Simulator**
  - **Other studies of Cray X1; NEC SX-6, IBM BlueGene/L; Cray Red Storm; Tera MTA, Sun Wildfire, etc.**
  - **Collaborators on architecture design projects, including Blue Planet.**
  - **Clusters - UCB Millennium and NOW, Processor in Memory (PIM)**
  - **RAID**
  - **HPSS**



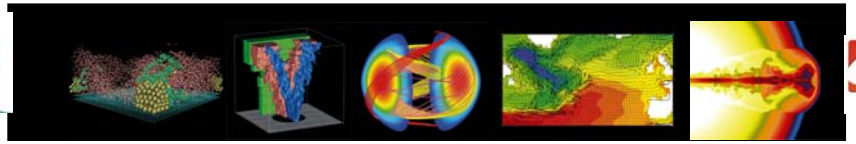
The PERVU Methodology

- A “holistic” evaluation methodology for large systems
- Performance
  - Many ways to determine this – some better than others
    - Application benchmarks
    - Linpack, NPBs, etc
    - Sustained System Performance (SSP) tests
- Effectiveness
  - Effective System Performance Test
- Reliability
  - Looking for new ways to proactively assess
  - Use for design tradeoffs
  - Apply to software
- Variation
  - CoV and other methods
- Usability
  - A relative metric – rather than absolute

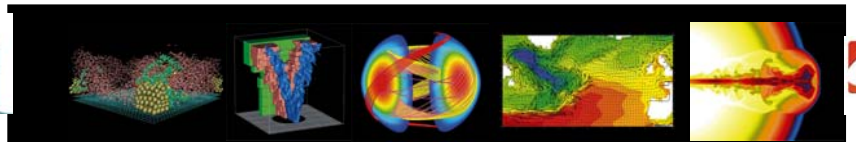


Summary

The impact of the Science-Driven Computer Architecture process will have lasting impact – leading to production Petaflop/s computing.

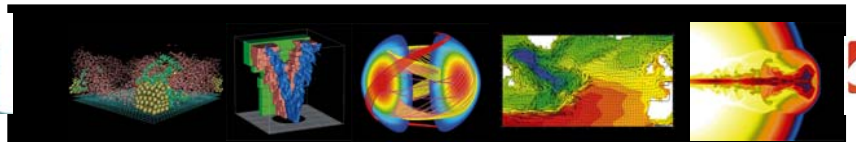


Backup



References

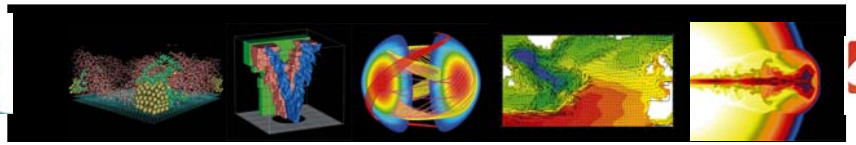
- National Facility for Advanced Computational Science: A Sustainable Path to Scientific Discovery (PDF)
- Creating Science-Driven Architecture: A New Path to Scientific Leadership (PDF)
- IBM SP Parallel Scaling Overview (PDF)
- ESP: A System Utilization Benchmark (PDF)



DOE Data Grows Dramatically

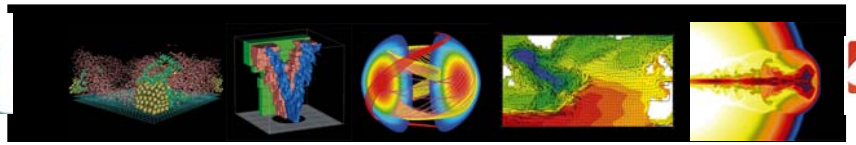
	1995-1999	2002-2004	2007-2009
Climate	5 TB/year	20 TB/year	3 PB/year
Fusion Energy	2 TB/year	20 TB/year	1 PB/year
Hadron Structure	50 TB/year	300 TB/year	2-3 PB/year
Quark-Gluon Plasma	50 TB/year	600 TB/year	5 PB/year
Material Science - Neutrons			200 TB/year
Material Science - Photons	3 TB/year	30 TB/year	150 TB/year
Chemistry – Combustion	100 GB/year	4 TB/year	100s of TB/year
Chemistry - Environmental	250 GB/year	100 TB/year	2 PB/year
Genomes to Life	10 GB/year	400 TB/year	10's of PB/year
Particle Physics	70 TB/year	500 TB/year	10-15 PB/year
Universe Asymmetry		200 TB/year	1 PB/year

gate.hep.anl.gov/may/ScienceNetworkingWorkshop/



NERSC System Ratios

System	Year	Peak Tera-ops	Ratio Bytes/Ops	Usable Shared Storage (TB)	Comment
NERSC-1 – C-90	1993	.16			Vector
NERSC-2 - T3E	1997	.63	4.4	3	Limited IO bandwidth
NERSC-2 – SV-1's (three)	1997	.12	12	1	Vector
NERSC 3	1999/2002	10	5	48	Undersized for shared storage
PDSF	2000	.3?	10	30	Limited I/O bandwidth
PDSF	2004	1.5?	60	100	
NCS	2004	3.3	7.5	25	
Visualization	2004	.03	100	3	
NCSb	2005	6-8 if a cluster	7.5	60	Projection
NERSC-5	2006	30-40 if a cluster	20	800	Projection



Large Jobs
Run Regularly – AY 2003