

# Leadership Computing at the NCCS Opportunities and Challenges

**NATIONAL CENTER  
FOR COMPUTATIONAL SCIENCES**



*presented by*

Arthur S. Bland

Director of Operations, National Center for Computational Sciences  
Oak Ridge National Laboratory

Cray Users Group  
Lugano, Switzerland  
May 8, 2006

Oak Ridge National Laboratory  
U.S. Department of Energy

# Overview

- Why Leadership Computing?
- What is Leadership Computing?
- Current System in the NCCS
- Roadmap for future system
- Infrastructure Requirements
- Application Challenges

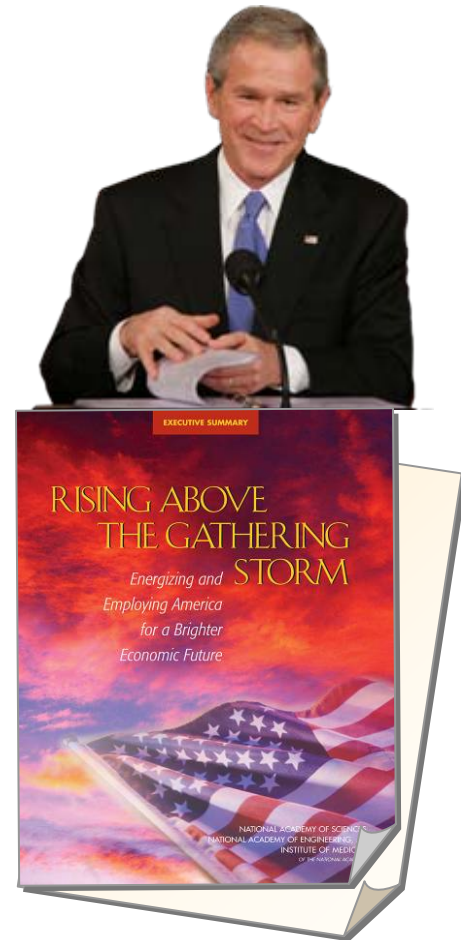
# American Competitiveness Initiative

**In the President's State of the Union Address on January 31, 2006, President Bush stated:**

*"I propose to double the federal commitment to the most critical basic research programs in the physical sciences over the next ten years. This funding will support the work of America's most creative minds as they explore promising areas such as nanotechnology, **supercomputing**, and alternative energy sources."*

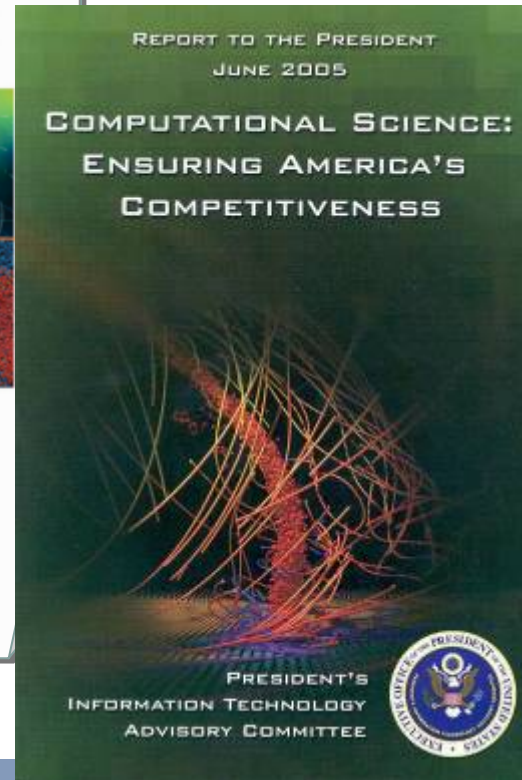
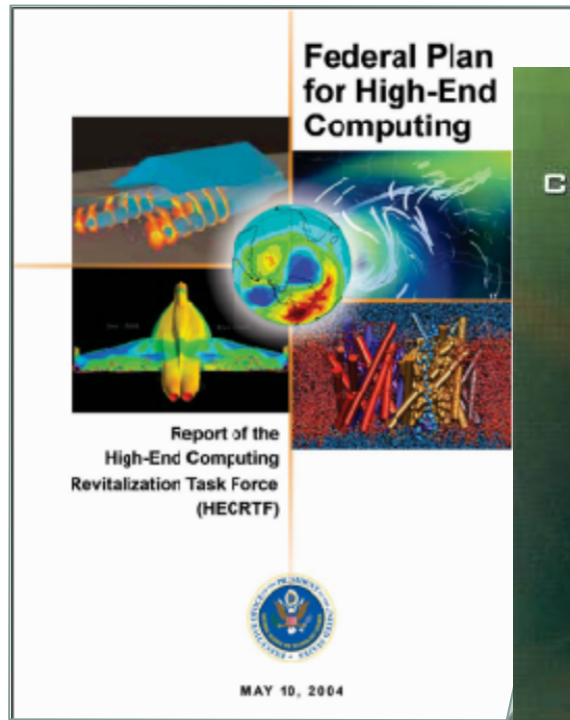
*Secretary of Energy Samuel Bodman:*

*"Developing revolutionary, science-driven technology is at the heart of the Department of Energy's mission. To ensure that America remains at the forefront in an increasingly competitive world, our Department is pursuing transformational new technologies in the cutting-edge scientific fields of the 21st century – areas like nanotechnology, material science, biotechnology, and **high-speed computing**."*



# Leadership Computing is a National Priority

*“The goal of such [leadership] systems is to provide computational capability that is at least **100 times greater** than what is currently available.”*



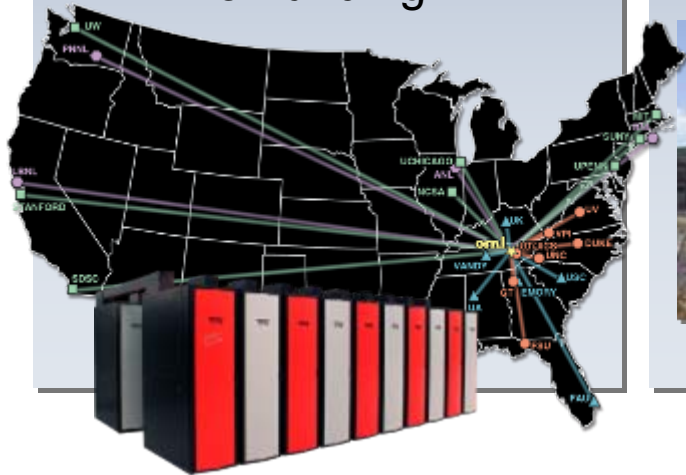
*“High-end system deployments should be viewed not as an interagency competition but as a shared strategic need that requires coordinated agency responses.”*

**In 2004 ORNL's NCCS was selected as the National Leadership Computing Facility**

# NCCS Mission to Enable Science Success

## World leader in scientific computing

“User facility providing leadership-class computing capability to scientists and engineers nationwide independent of their institutional affiliation or source of funding”



## Intellectual center in computational science

Create an interdisciplinary environment where science and technology leaders converge to offer solutions to tomorrow's challenges



## Transform scientific discovery through advanced computing

“Deliver major research breakthroughs, significant technological innovations, medical and health advances, enhanced economic competitiveness, and improved quality of life for the American people”

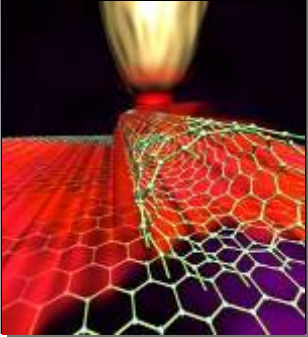


– Secretary Abraham

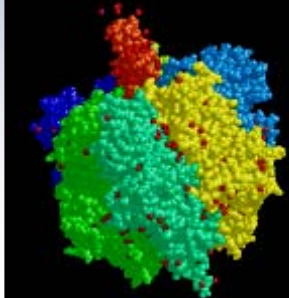


# Key National Science Priorities

**Manipulating  
the Nanoworld**



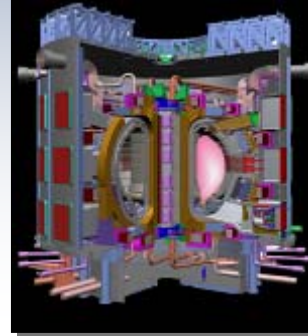
**Taming the  
Microbial  
World**



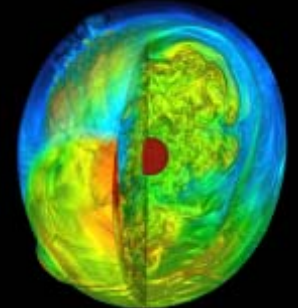
**Environment  
and  
Health**



**ITER for  
Fusion  
Energy**



**Search  
for the  
Beginning**



**Recent NCCS research includes:**

**Largest simulation of plasma behavior in a tokomak**

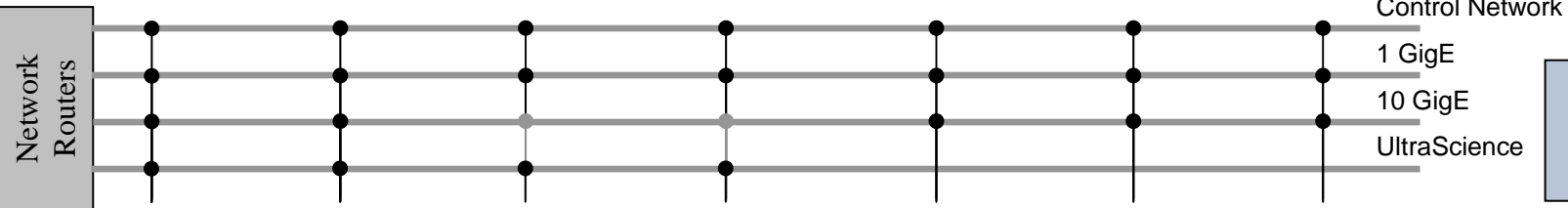
**Resolution of theoretical disputes in superconducting research**

**Identification of shock wave instability in supernovae collapse**

**Seeing interplay of complex chemistry in combustion**

# NCCS Resources

May 2006  
Summary



**7 Systems**

**Supercomputers**  
7,622 CPUs  
16TB Memory  
45 TFlops

**Total Shared Disk**  
238.5 TB

**5 PB**

**CRAY XT3 JAGUAR**



(5,294) 2.4GHz  
11TB Memory

120TB

**CRAY X1E PHOENIX**



(1,024) 0.5GHz  
2 TB Memory

32TB


**SGI ALTIX RAM**



(256) 1.5GHz  
2TB Memory

36TB


**IBM SP4 CHEETAH**



(864) 1.3GHz  
1.1TB Memory

32TB


**IBM LINUX NSTG**



(56) 3GHz  
76GB Memory

4.5TB


**VISUALIZATION CLUSTER**



(128) 2.2GHz  
128GB Memory

9TB

**IBM HPSS**



Many Storage Devices Supported

5TB

**Scientific Visualization Lab**

27 projector,  
35 megapixel  
Powerwall

**Test Systems**

- 1 Cabinet Cray XT3
- 32 processor Cray X1E\*
- 16 Processor SGI Altix

**Evaluation Platforms**

- 144 processor Cray XD1 with FPGAs
- SRC Mapstation
- Clearspeed
- BlueGene (at ANL)

**Backup Storage**

5PB



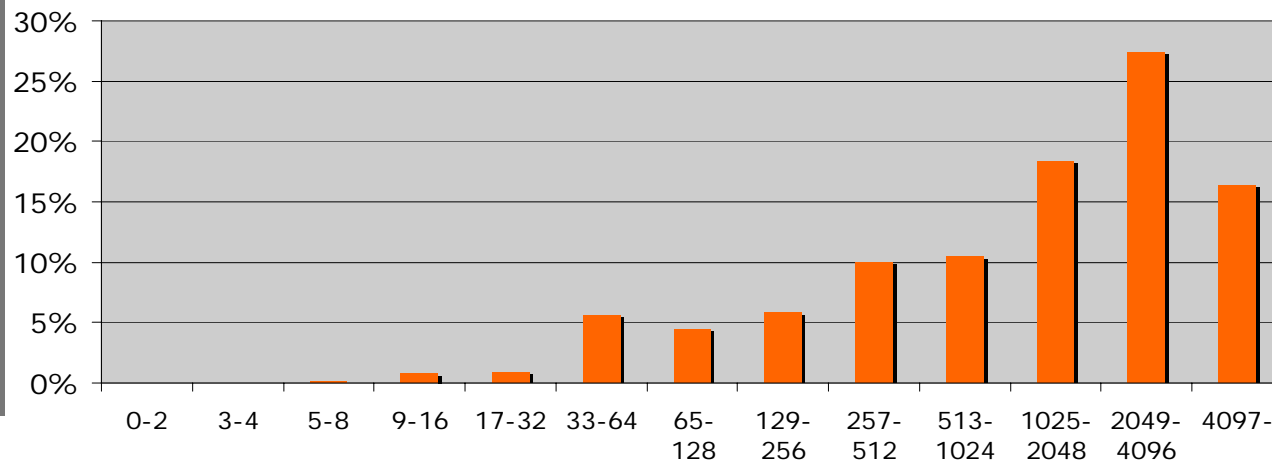
# Jaguar

5,294 processors and 11 TB of memory



*Accepted in 2005 and routinely running applications requiring 4,000 to 5,000 processors*

- 43% of time used by jobs using 40% of system or more
- 61% of time used by jobs requiring 1000+ processors



Machine Usage by Number of Processors



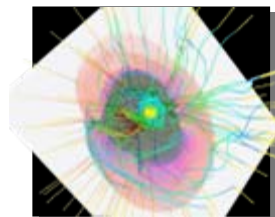
# Phoenix

## 1,024 processors and 2 TB of memory



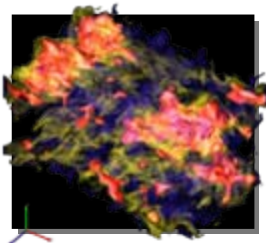
***Highly scalable  
hardware and software***

***High sustained  
performance on real  
applications***



### Astrophysics

Simulations have uncovered a new instability of the shock wave and a resultant spin-up of the stellar core beneath it, which may explain key observables such as neutron star “kicks” and the spin of newly-born pulsars.

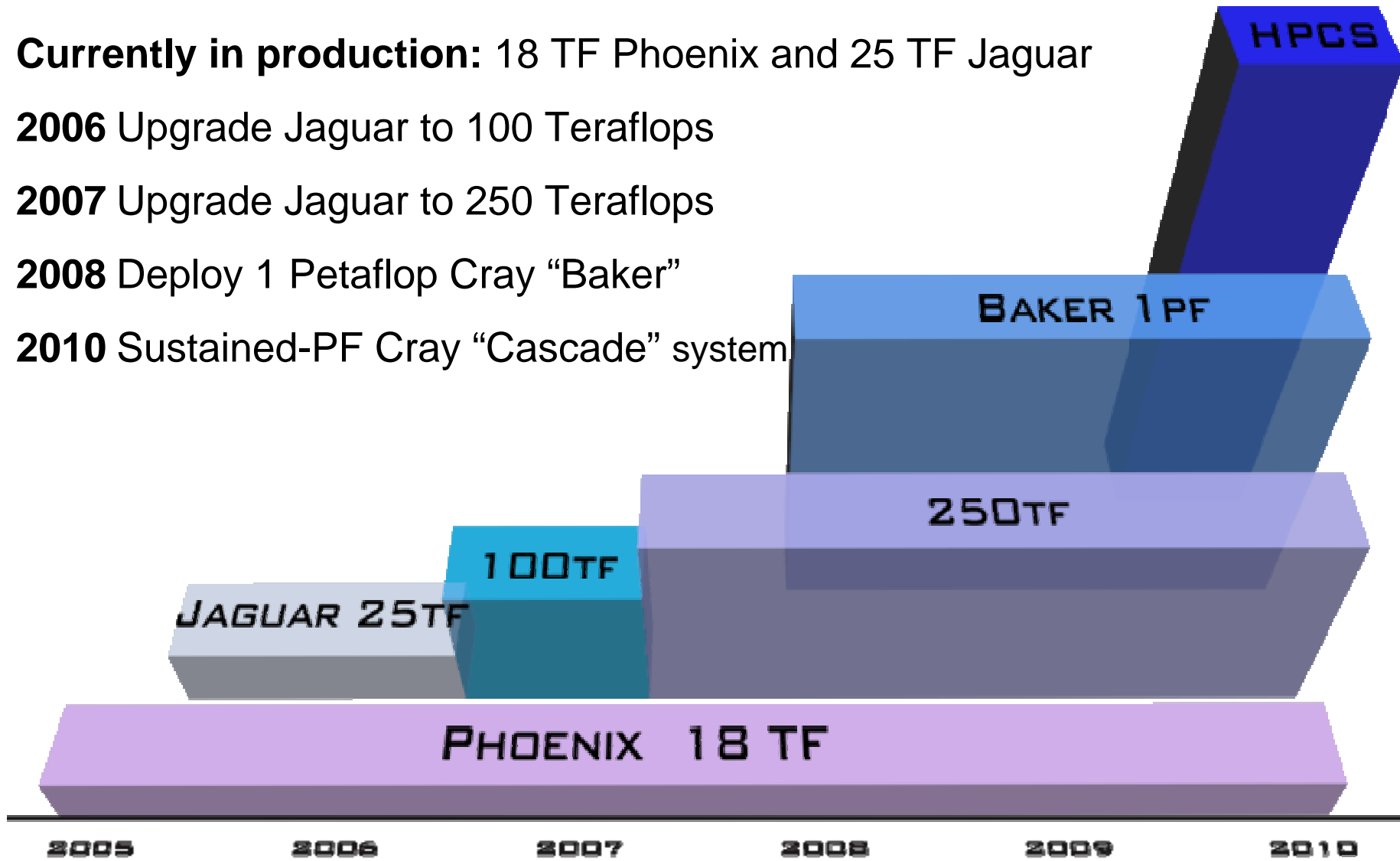


### Combustion

Calculations show the importance of the interplay of diffusion and reaction, particularly where strong finite-rate chemistry effects are involved.

# Hardware Roadmap

- **Currently in production:** 18 TF Phoenix and 25 TF Jaguar
- **2006** Upgrade Jaguar to 100 Teraflops
- **2007** Upgrade Jaguar to 250 Teraflops
- **2008** Deploy 1 Petaflop Cray “Baker”
- **2010** Sustained-PF Cray “Cascade” system



# Jaguar's Path to 250 TF

## Jaguar 2006 upgrade

- Upgrade single-core to dual-core (2.6 GHz)
- Upgrade memory to maintain 2 GB per core
- Add 68 cabinets (total of 124)
- 11,508 dual-core compute sockets
- 119 TF peak
- 46 TB memory
- 900+ TB disk storage
- 55 GB/s disk bandwidth

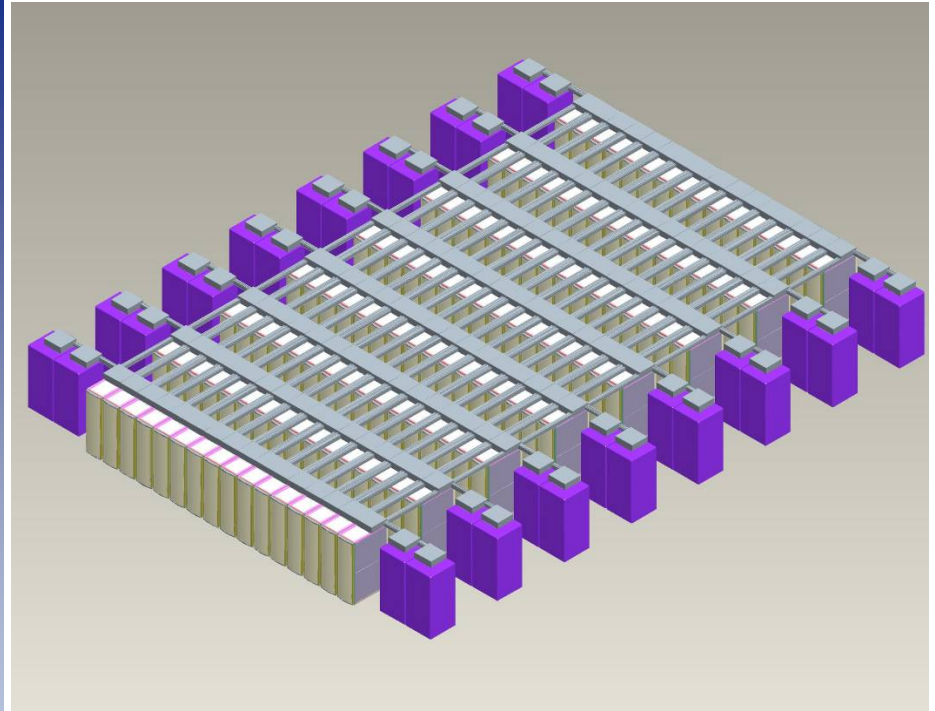
## Jaguar 2007 upgrade

- Upgrade 68 cabinets to multi-core
- O(5000) dual-core compute sockets
- O(6000) multi-core compute sockets
- Double memory in upgraded nodes
- 250+ TF peak compute partition
- 50+ TF data analysis partition
- *70 TB memory*

# 1000 TF Cray "Baker" system at ORNL-2008

## System Configuration

- 1 PF peak
- $O(100,000)$  threads of execution
- 200-400 TB memory
- 4-10 PB file system
- 128 cabinets
- R-134a heat exchange units
- 7-8 MW power
- 480 V power supplies



*1 PF Cray System*

# Petascale Computers Require Extreme Power

- New 70 MW substation will be operational January 2007
- Easily upgradeable to 140 MW
- Redundant 161 kV supply lines, each with over 10 year MTTI
- Computer center power upgrading from 8 MW today to over 30 MW in phases over the next three years



New Power Substation Under Construction



13,800 volt transformers for computers and infrastructure



# Removing the Heat

- Petascale system will be over 45 KW per cabinet, *over 3000 watts per foot<sup>2</sup>!* (*280 watts per meter<sup>2</sup>*)
- Simple forced air cooling will not be enough
- ORNL is installing two 30" diameter (0.76 meter) chilled water lines to supply additional cooling to the building
- Cooling system must be on generator power to prevent damage to computers from residual heat in the event of power loss



Cooling today for 8 MW with hot-spare



30+ MW Central Chiller Plant

# NCCS Infrastructure Systems



## High Performance Storage System

**Multi-Petabyte data archive used by HPC centers around the world**  
**Developed by ORNL, LLNL, LANL, SNL, LBNL, and IBM**



Data Analysis and Visualization Clusters		
128 AMD Opteron	160 Intel Xeon	256 SGI Altix
Quadrics	Infiniband	NUMalink 3



## Visualization Facility

**30' x 8' display wall with 35 megapixel resolution**  
**Stereo Immersadesk, 23 megapixel LCD powerwall**  
**Chromium, DMX, VisIt, EnSight, Paraview, AVS**

# NCCS Software Infrastructure

## Operating Systems

**Linux on all new systems**  
**Unix variants (Unicos/MP, Unicos/LC)**  
**Batch systems: Loadleveler, PBSpro, MOAB**

## File Systems Strategy

**Unified home directories on NFS (moving to Lustre)**  
**Local scratch file systems**  
**High speed parallel file systems (Lustre)**  
**HPSS Archival storage**

## Programming Environment

**Fortran, C, C++, Co-array Fortran, UPC**  
**MPI, OpenMP, shmem**  
**Totalview debugger, variety of performance tools**

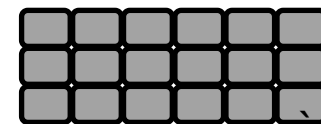
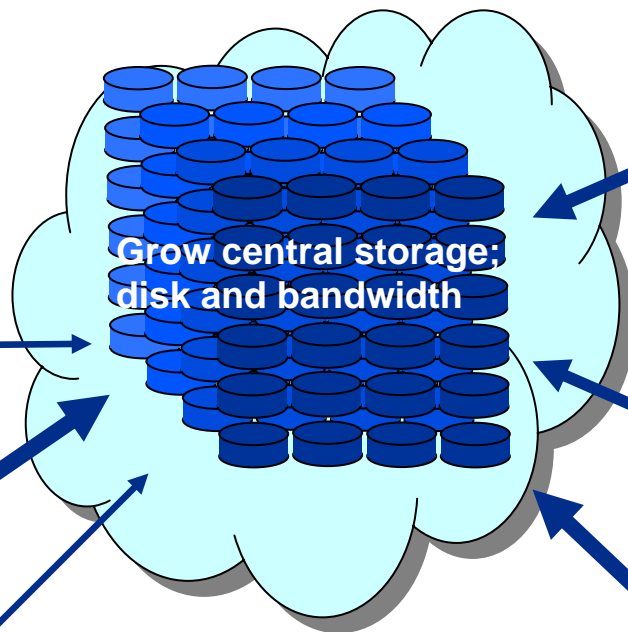
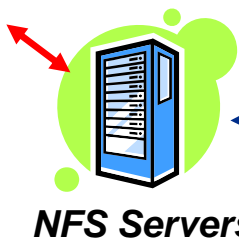
## Libraries

**BLAS, LAPACK, ScaLAPACK,**  
**ESSL, PSSL, scilib, TOPS**

# Maintain Infrastructure Balance – Decouple File System From Computer Systems



**Phoenix  
Cray X1E**



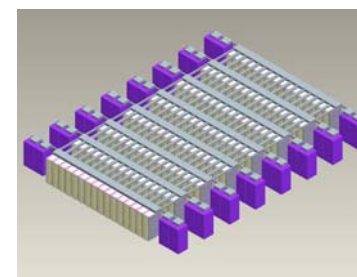
**Data Analysis  
& Visualization**



**Jaguar  
Cray XT3**



**HPSS**



**Baker**

## Late 2006

- 100 TB
- 10 GB/s (aggregate)

## 2008

- 1-10 PB
- 300-750 GB/s (aggregate)

Shane Canon will discuss  
This on Thursday Morning

Increase HPSS  
bandwidth

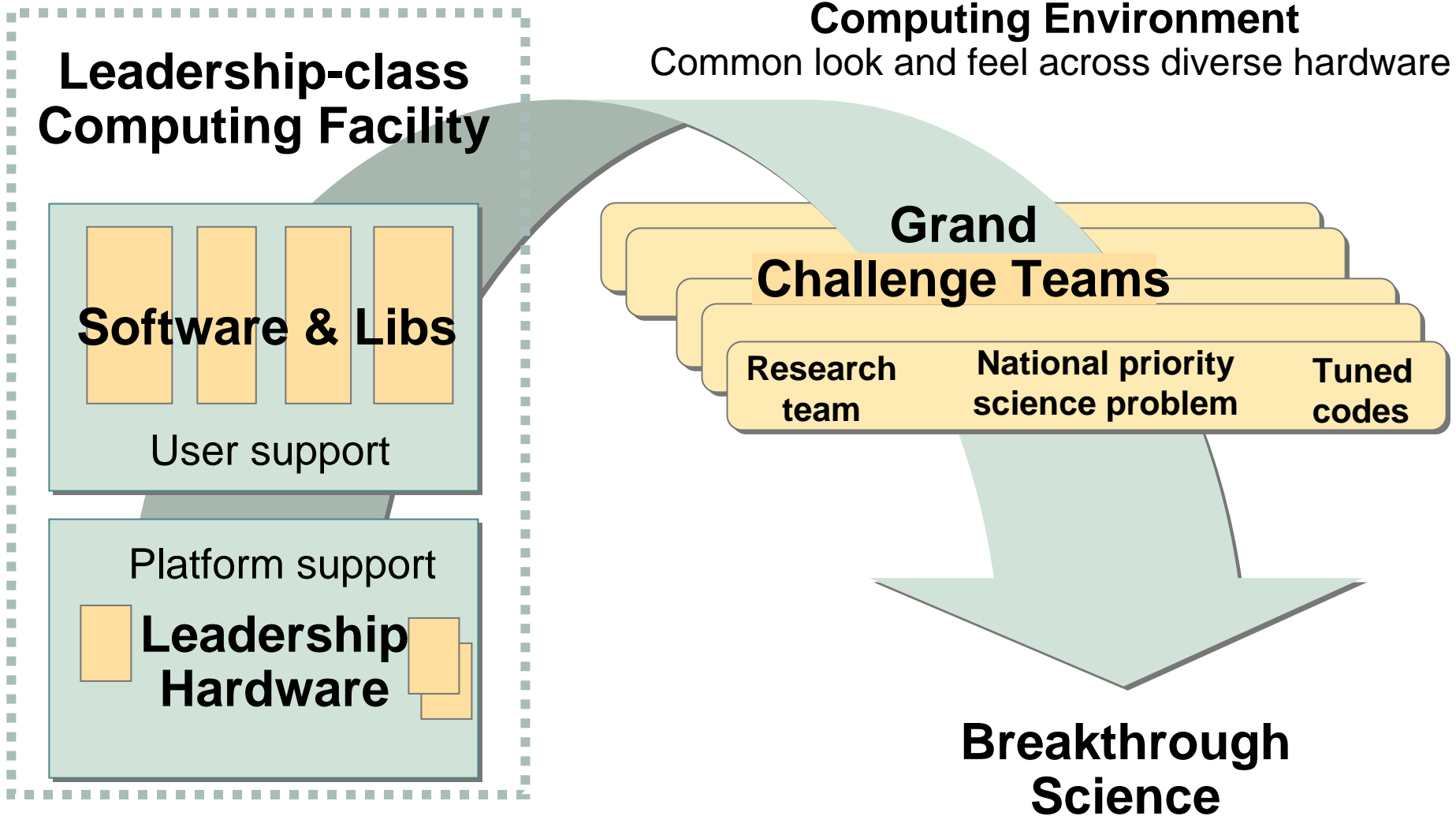
Increase WAN  
bandwidth

*ESnet, USN,  
TeraGrid,  
Internet2, NLR*

# Application Challenges

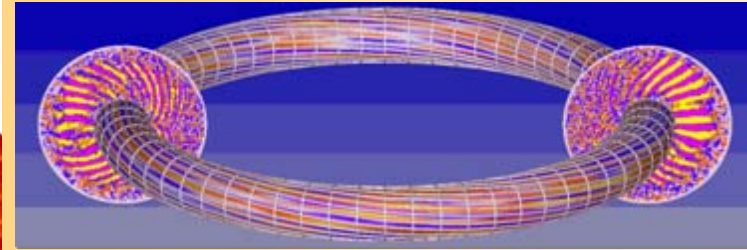
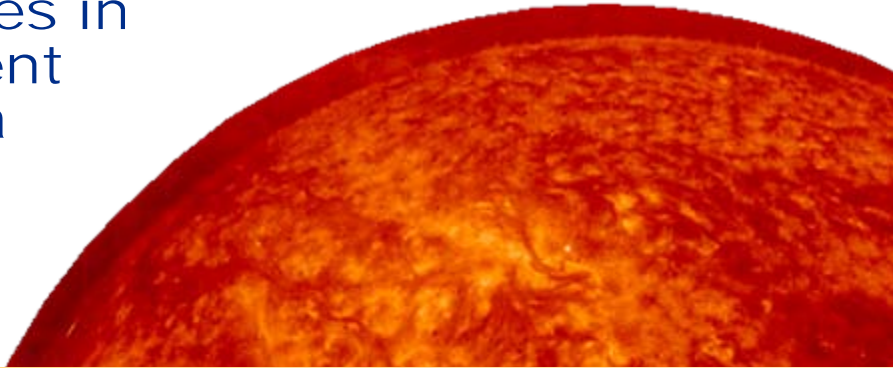
- Scaling from 25 TF to 1000 TF in three years
- Applying multi-core processors effectively
- Converting codes to use scalable, parallel I/O techniques
- Analysis of 100+ Terabyte Datasets
- Data movement (disk, archive, visualization, etc.)
- Larger cache line sizes require effective cache and memory blocking to achieve high memory bandwidth
  - DDR chips give 2 bits per clock
  - DDR2 chips give 4 bits per clock
  - DDR3 chips give 8 bits per clock
  - **Use them or lose them!**

# Addressing the Challenges





# Fusion Simulation: Particles in turbulent plasma



*A twisted mesh structure is used in the GTC simulation.*

## Principal Investigators

William Tang and Stephane Ethier  
Princeton Plasma Physics Laboratory

- **The Problem**

Ultimately, fusion power plants will harness the same process that fuels the sun. Understanding the physics of plasma behavior is essential to designing reactors to harness clean, secure, sustainable fusion energy.

- **The Research**

These simulations will determine how plasma turbulence develops. Controlling turbulence is essential because it causes plasma to lose the heat that drives fusion. Realistic simulations determine which reactor scenarios promote stable plasma flow.

- **The Goal**

The NLCF simulations will be the highest-resolution Gyrokinetic Toroidal Code (GTC) models ever attempted of the flow of charged particles in fusion plasmas to show how turbulence evolves.

- **Impact of Achievement**

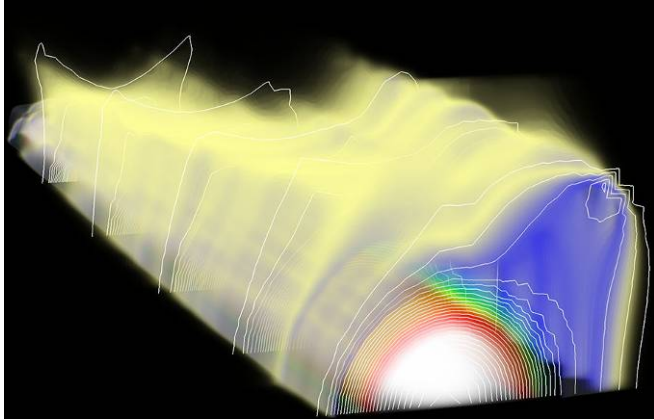
High-resolution computer simulations are needed for preliminary data to set up experiments that make good use of limited and expensive reactor time. Engineers will use the resulting data to design equipment that creates scenarios favorable to efficient reactor operation.

- **Why NLCF**

The fusion simulations involve four billion particles. The Cray X1E's vector processors can process these data 10 times faster than non-vector machines, achieving the high resolution needed within weeks rather than years.

# Largest ever AORSA Simulation

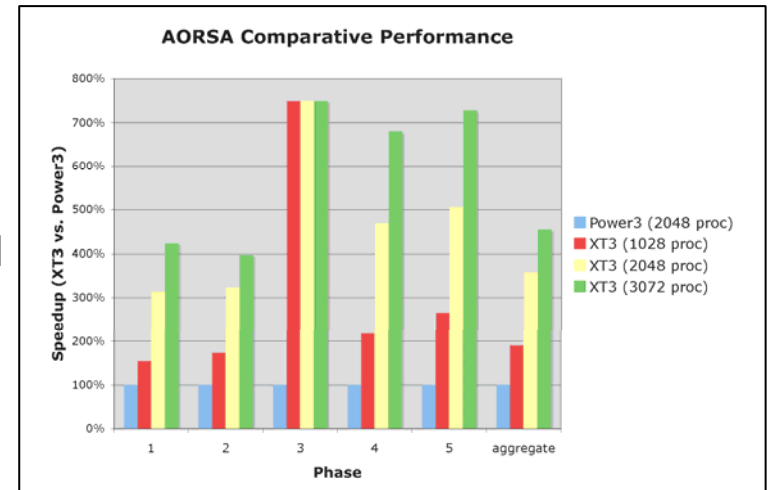
## 3072 processors of NCCS Cray XT3



In August 2005, just weeks after the delivery of the final cabinets of the Cray XT3, researchers at the National Center for Computational Sciences ran the largest ever simulation of plasma behavior in a tokamak, the core of the multinational fusion reactor, ITER.

*Velocity distribution function for ions heated by radio frequency (RF) waves in a tokamak plasma.*

The code, AORSA, solves Maxwell's equations – describing behavior of electric and magnetic fields and interaction with matter – for hot plasma in tokamak geometry. The largest run by Oak Ridge National Laboratory researcher Fred Jaeger utilized 3072 processors: roughly 60% of the entire Cray XT3.



*AORSA on the Cray XT3 “Jaguar” system compared with Seaborg, an IBM Power3. The columns represent execution phases of the code: Aggregate is the total wall time, with Jaguar showing more than a factor of 3 improvement over Seaborg.*

# ORNL Talks at CUG 2006

1. Portable Performance with Vectorization, Mark Fahey and James B. White III
2. Leadership Computing at the NCCS, Arthur Bland
3. A User Perspective on the High Productivity Computer Systems (HPCS) Languages, Wael R. Elwasif, David E. Bernholdt,
4. Moab Workload Manager on Cray XT3, Don Maxwell
5. Performance Evaluations of User Applications on NCCS's Cray XT3 and X1E, Arnold Tharrington
6. Evaluation of the Cray XT3 at ORNL: A Status Report, Richard Barrett, Mark Fahey, Bronson Messer, Philip Roth
7. FV-CAM Performance on the XT3 and X1E, Patrick Worley
8. HPCC Update and Analysis, Jeff Kuehn
9. Evaluation of UPC on the Cray X1E, Richard Barrett
10. Comparing Optimizations of GTC for the Cray X1E and XT3, James B. White III
11. High Level Synthesis of Scientific Algorithms for the Cray XD1 System, Philip LoCascio
12. Characterizing Applications on the MTA2 Multithreading Architecture, Richard Barrett, and Philip Roth
13. Resource Allocation and Tracking System (RATS) Deployment on the Cray X1E, XT3, and XD1 Platforms, Robert Whitten
14. Co-Array Fortran Experiences Solving PDE Using Finite Differencing Schemes, Richard Barrett
15. Experiences Harnessing Cray XD1 FPGAs and Comparisons to other FPGA High Performance Computing (HPC) Systems, Olaf Storaasli
16. A Center Wide File System Using Lustre, Shane Canon

# Questions?

Arthur S. Bland

Director of Operations, National Center for Computational Sciences  
Oak Ridge National Laboratory

BLANDAS@ornl.gov

**NATIONAL CENTER**  
FOR COMPUTATIONAL SCIENCES



This work was prepared by UT-Battelle, LLC for the U.S.  
Department of Energy under contract DE-AC05-00OR22725