National Leadership Computing Facility - *Bringing Capability Computing to Science*

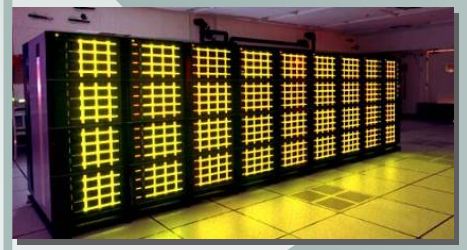
Jack Wells
Computer Science and Mathematics Division,
Oak Ridge National Laboratory

Jeff Nichols, Director
Computer Science and Mathematics Division,
National Center for Computational Sciences
Oak Ridge National Laboratory

PSIF Workshop, Oak Ridge
March 21, 2005

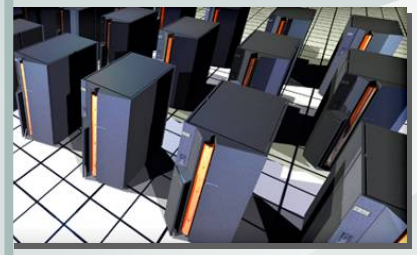
Center for Computational Sciences performs three inter-related activities for DOE

- Deliver National Leadership Computing Facility for science
  - **Focused on grand challenge science and engineering applications**
- Principal computing resource for many DOE programs
  - **Specialized services to the scientific community: biology, climate, nanoscale science, fusion**
- Evaluate new hardware for science
  - **Develop/evaluate emerging and unproven systems and experimental computers**



Intel Paragon: World's fastest computer

IBM Power3: DOE-SC's first terascale system



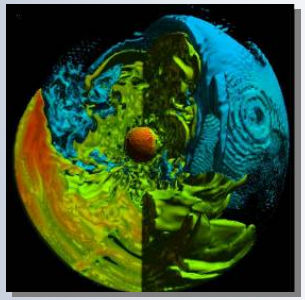
IBM Power4: 8th in the world (2001)

Cray X1: Capability computer for science

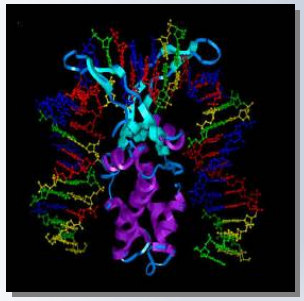
1995 2000 2001 2003

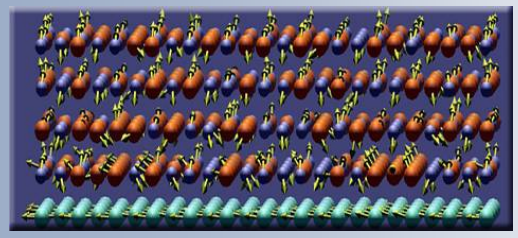
Focused on grand challenge scientific applications

SciDAC Astrophysics

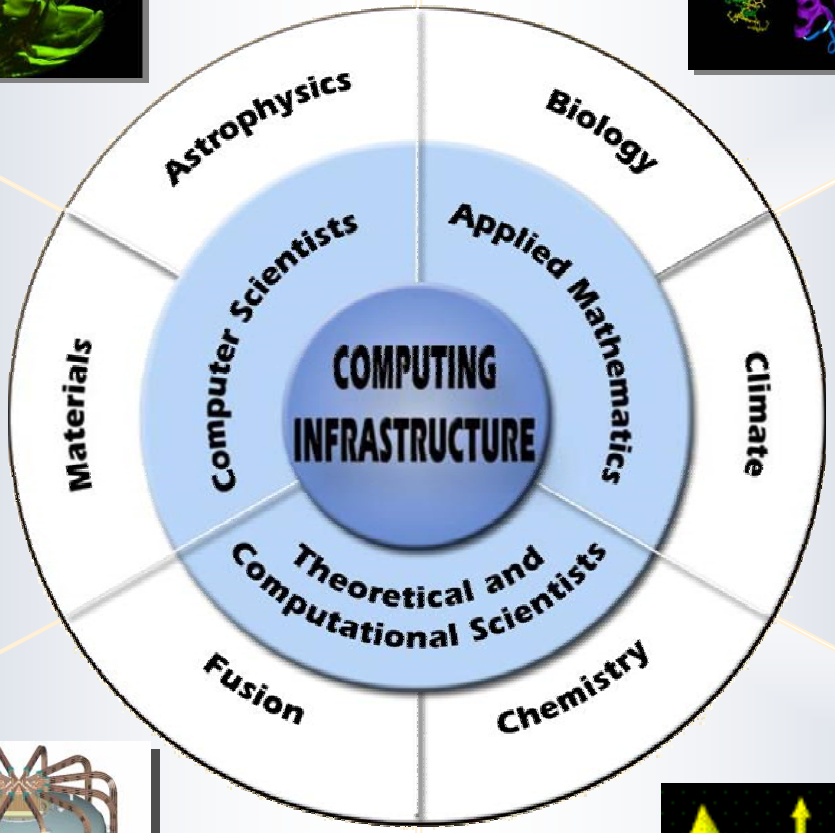


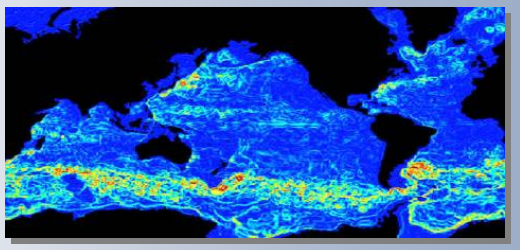
Genomes to Life





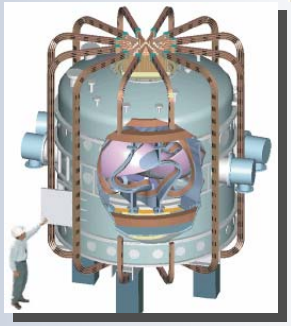
Nanophase Materials Prototype Topical Center

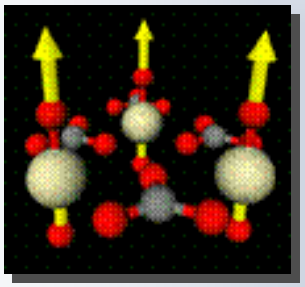




SciDAC Climate Prototype Topical Center

SciDAC Fusion Prototype Topical Center





SciDAC Chemistry

Center for Computational Sciences: New facilities

- $72M private sector investment in support of leadership computing
- Space and power:
  - **40,000 ft² computer center with 36-in. raised floor, 18 ft. deck-to-deck**
  - **8 MW of power (expandable) @ 5c/kWhr**
- High-ceiling area for visualization lab (Cave, Powerwall, Access Grid, etc.)
- Separate lab areas for computer science and network research

CCS Terascale Computer Systems

Cray X1 – Phoenix (6.4 TF)
Largest X1 in the world and first in DOE
Scalable Vector Architecture
512 Multi-Streaming Processors 400 MHz

IBM Power4 – Cheetah (4.5 TF)
First Power4 system in DOE - #8 on Top500
Cluster using IBM Federation Interconnect
864 IBM Power4 processors 1.3 GHz

SGI Altix – Ram (1.5 TF)
Large Globally Addressable Memory System
256 Intel Itanium2 Processors 1.5 GHz
Linux with single operating system image

IBM Power3 – Eagle (1 TF)
First Terascale machine in DOE-SC
Cluster of SMP Nodes
736 IBM Power3 Processors 375 MHz

February, 2004: Office of Science notice to SC Laboratories

Leadership-Class Computing Capability for Science

"The focus of the proposed effort should be on capability computing in support of high-end science – rather than on enhanced computing capacity for general science users."

"The proposed effort must be a user facility providing leadership class computing capability to scientists and engineers nationwide independent of their institutional affiliation or source of funding."

"Proposals must include specific information on architecture or architectures to be provided over the life of this project as well as a list of targeted applications."

Our plan of action to deliver leadership computing for DOE

- **Rapidly field most powerful open capability computing resource for scientific community**
  - **Providing clear upgrade path to at least 100 teraflop/s (TF) by 2006 and 250 TF by 2007/2008**
- **Deliver outstanding access and service to research community**
  - **Utilizing most powerful networking capability extant coupled with secure and highly cost-effective operation by proven team**
- **Deliver much higher sustained performance for major scientific applications than currently achievable**
  - **Developing next generation models and tools**
  - **Engaging computer vendors on hardware needs for scientific applications**
- **Engage research communities in climate, fusion, biology, materials, chemistry, and other areas critical to DOE-SC and other federal agencies**
  - **Enabling high likelihood of breakthroughs on key problems**
- **Conduct in-depth exploration of most promising technologies for next-generation leadership-class computers**
  - **Providing pathways to petaflop/s (PF) computing within decade**

National Leadership Computing Facility

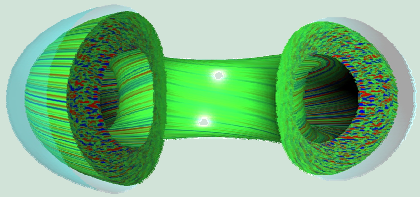
May 12, 2004

Phoenix – The CCS Cray X1

- Largest Cray X1 in the world – 6.4TF
- 2 TB globally addressable memory
- 512 processors
  - 400 MHz, 800 MHz vector units
- 32 TB of disk
- Most powerful processing node
  - 12.8 GF CPU, 2-5x commodity processors
- Highest bandwidth communication with main memory
  - 34.1 GB/sec

- Highly scalable hardware and software
- High sustained performance on real applications

GYRO: A 5-D gyrokinetic-Maxwell solver



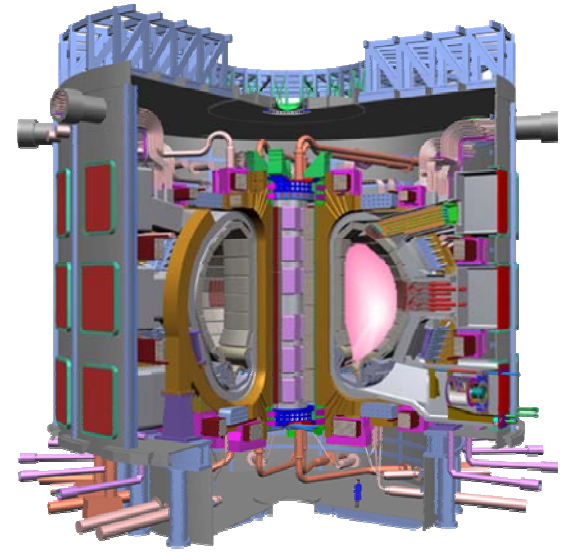
$$\frac{\partial f}{\partial t} = L_a f + L_b \Phi + \{f, \Phi\} \text{ where } F\Phi = \iint dv_1 dv_2 f$$

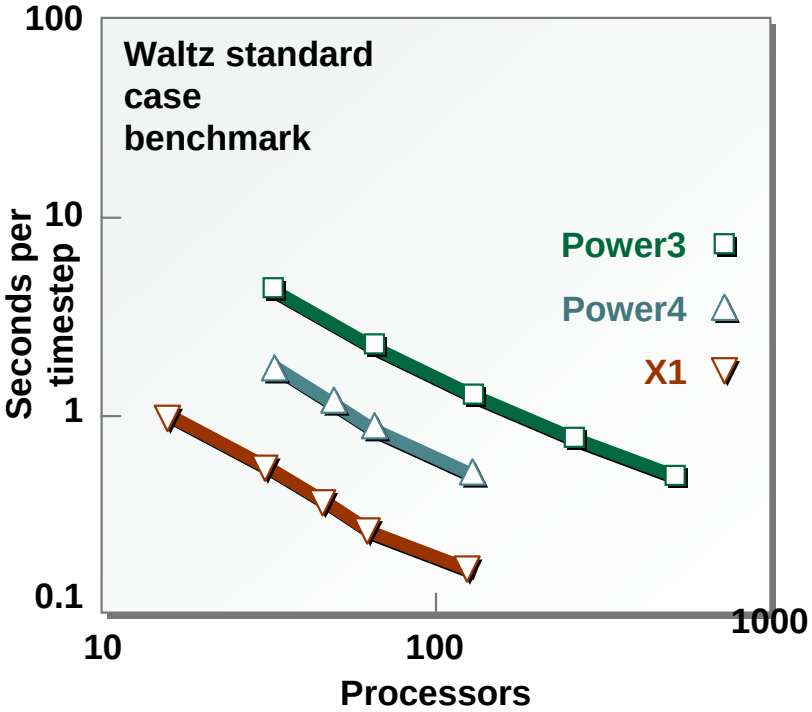
GYRO employs Implicit-Explicit Runge-Kutta scheme discretized on Eulerian grid

For typical simulations, GYRO performance is at least 4x faster on Cray X1 than SGI Altix or IBM Power4;

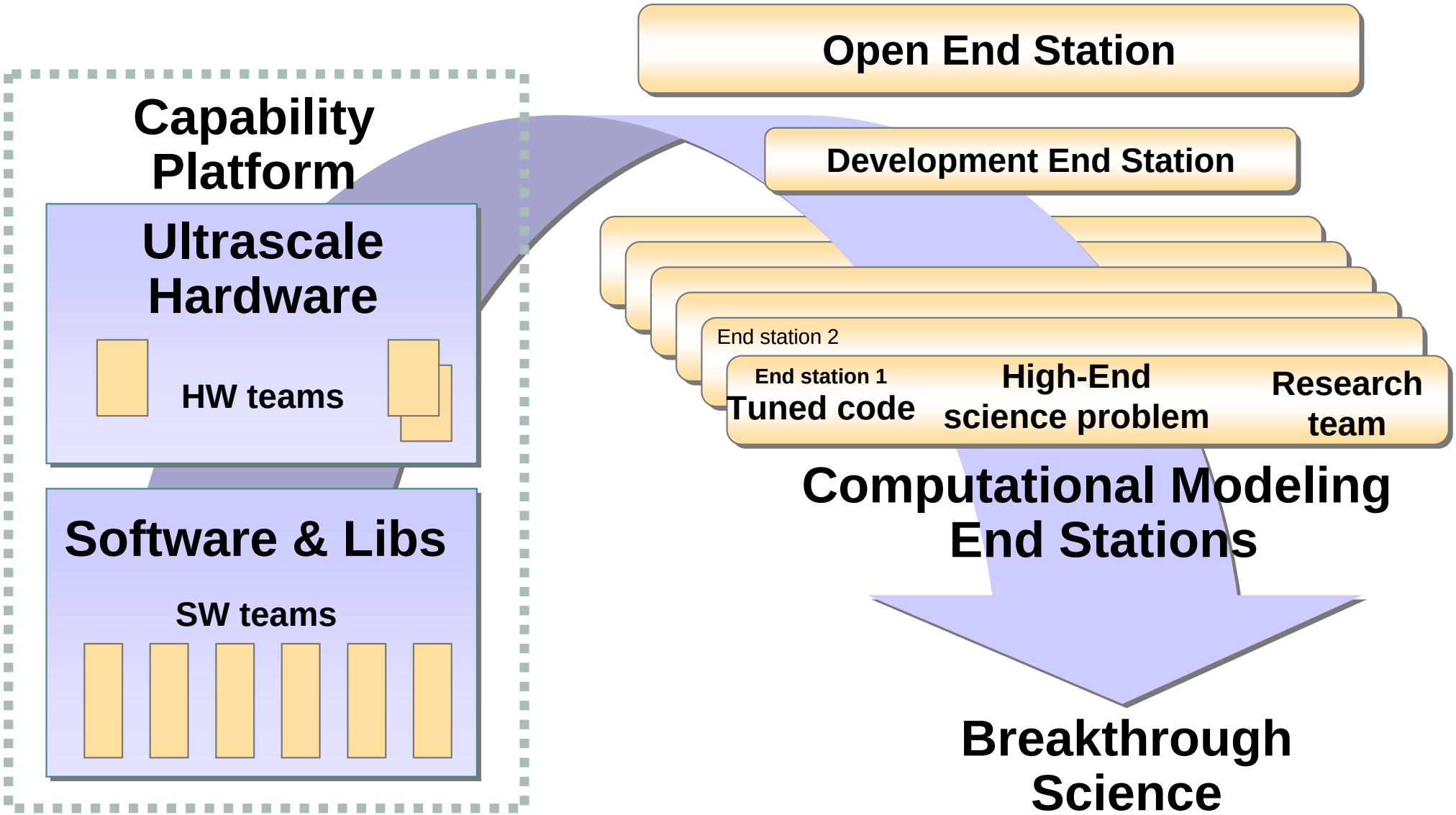
X1 is significantly faster for larger problems.

Recent accomplishments include: completed the most comprehensive tokamak turbulence simulation, showed transport is smooth across an s=0 (minimum-q surface) due to the appearance of gap modes, found that χ_i matched the "Cyclone value" at small ρ_* resolving an earlier conflicting report while simultaneously finding there is a long transient period for which χ_i exceeds the statistical average in GK simulations.





Active engagement of a few science teams enabled through "End Stations"



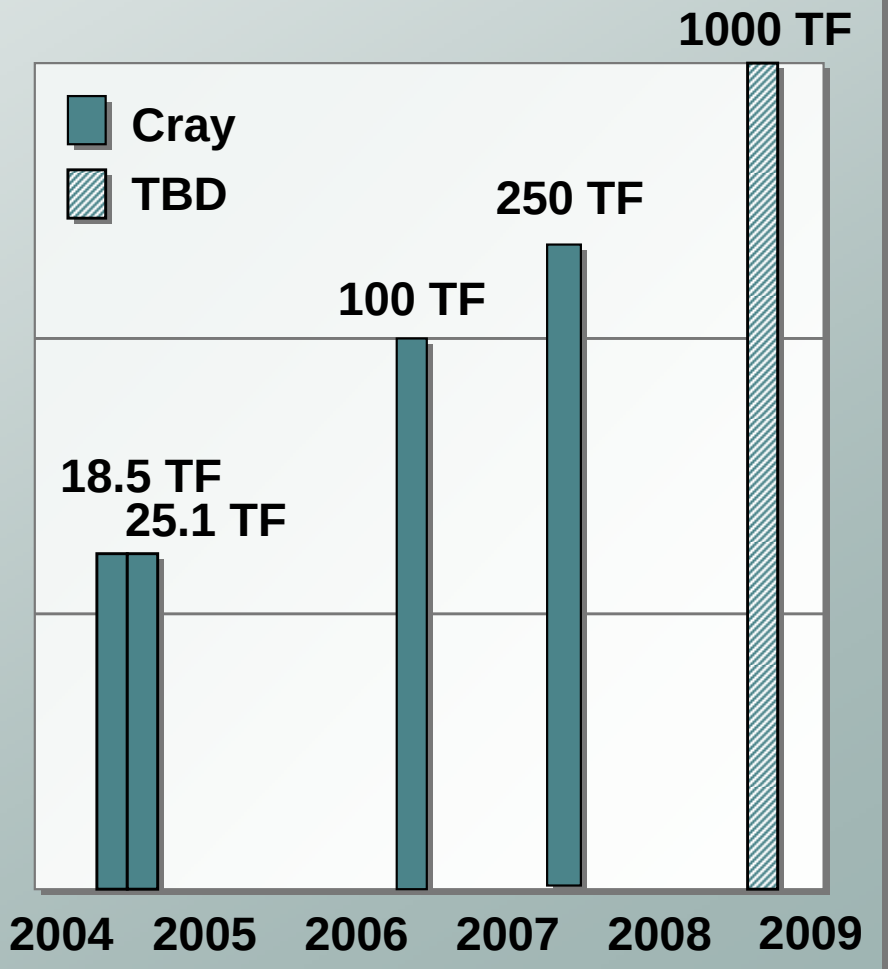
NLCF hardware roadmap

In 2005,

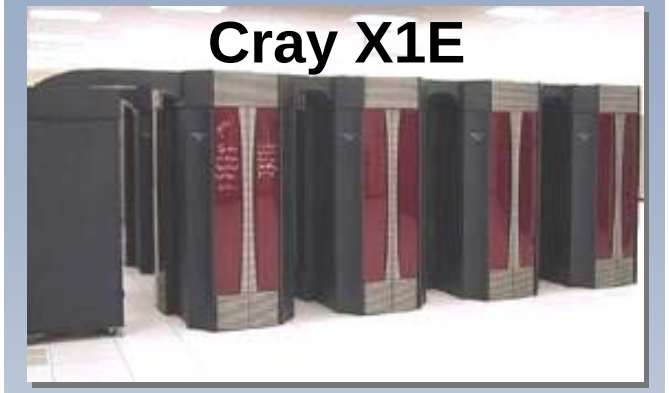
- **Deploy 18.5TF Cray X1E and 25.1TF Cray XT3 systems**
- **Cray forms and supports "Supercomputing *Center of Excellence*"**
- **Develop/deploy complementary software environment**
- **Full operational support of NLCF as a capability computing center**
- **Deliver computationally intensive projects of large scale and high scientific impact through competitive peer review process**

In 2006, deploy 100TF Cray XT3

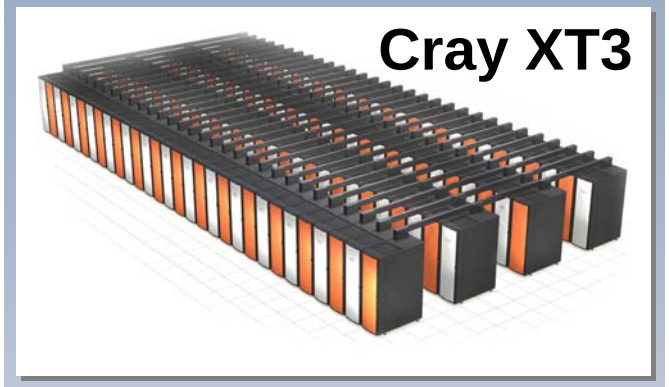
In 2007-8, deploy 250TF Cray Rainier



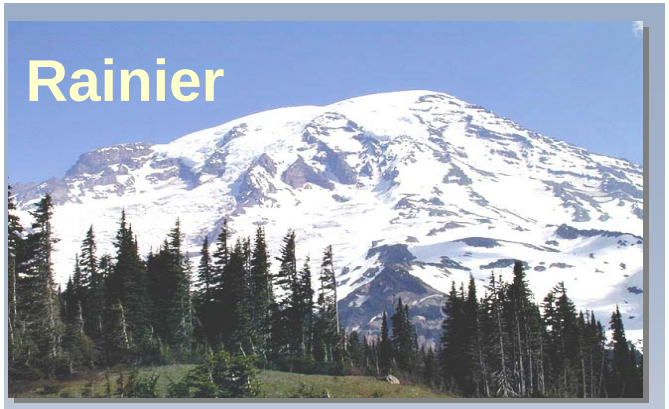
Three purpose-built architectures optimized for applications



- Proven architecture for performance and reliability
- Most powerful processors and interconnect
- Scalable, globally addressable memory and bandwidth
- Offers capability computing for key applications



- Extremely low latency, high bandwidth interconnect
- Efficient scalar processors, balanced interconnect
- Known system architecture – based on ASCI Red
- Linux operating system on service processors with microkernel on compute processors



- Unified product including vector, scalar, and potentially reconfigurable and multi-threaded processors
- Single solution for diverse workloads
- Single user interface and environment
- Improved performance by matching processor to job
- Leverage all engineering resources

Cray X1E Supercomputer

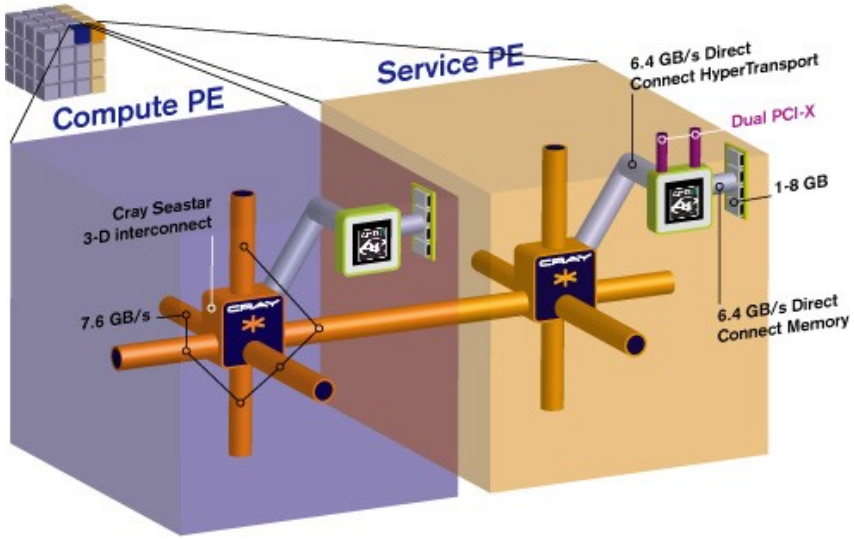
- **Re-implement X1 in 0.13μm technology**
  - Double processor density
  - 41% performance increase per processor
  - Cache scales in bandwidth with processor
  - Significantly reduces cost per processor
- **Approximately triples X1 performance per cabinet**
- **Upgradeable from X1 by processor swap**
- **Low-risk upgrade**
  - Software that runs on Cray X1E is faster, but *not* different
    - Users do not need to recompile
    - Supported features identical between X1 and X1E
    - Uses one code base for both products

Cray XT3 Architecture

Design Goals

- High-performance commodity microprocessor
- Surround with balanced or “bandwidth rich” environment
- Eliminate “barriers” to scalability
  - SMPs do not help here
  - Eliminate Operating System Interference (OS Jitter)
  - Reliability must be designed in
  - Resiliency is key
  - System Management
  - I/O

Cray XT3 Scalable Architecture



- XT3 is 3rd generation Cray MPP
- System implements many T3E architectural concepts in current best-of-class technologies
- Scales to many 1000s of processors
  - Scalable single-PE architecture
  - Scalable reliability and system management
  - Scalable SW environment
  - Scalable I/O subsystem

Jaguar – The CCS XT3 system

Cabinets	Performance	Processors	Memory	Disk Space	I/O Bandwidth
56	25 TF	5,304	10.5 TB	120 TB	15 GB/s
120	54.6 TF	11,374	23 TB	240 TB	30 GB/s
120	109 TF	22,748	46 TB	480 TB	60 GB/s

"Jaguar" first 11 cabinets

National Leadership Computing Facility
2005 Call for Proposals

Eligibility

- **Principal investigators engaged in scientific research with the intent to publish results in open peer-reviewed literature are eligible**
- **Program specifically encourages proposals from universities, other research institutions and industry**

Applications will be accepted electronically

- **Proposals accepted from May 1, 2005 until midnight EDT Wednesday, June 1, 2005**
- **Awards expected to be announced by July 30, 2005**
  - **Access established to NLCF facilities for awardees October 1, 2005**
  - **Remains in effect until September 30, 2006**

National Leadership Computing Facility
2005 Call for Proposals

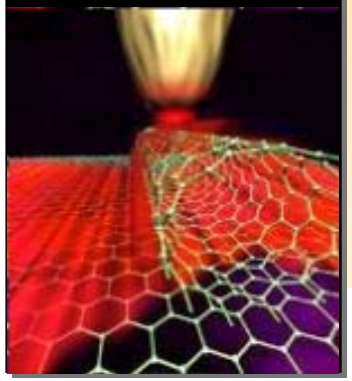
- **Largest, most computationally intensive scientific endeavors to be conducted on high capability Leadership systems of DOE's National Leadership Computing Facility (NLCF)**
- **Seeks high priority, challenging, high-payoff, and heretofore intractable, computationally-intensive experiments, where capability of NLCF systems can enable new breakthroughs in science**
- **Selection of limited set of scientific applications (perhaps 10 per year) and given substantial access to leadership system**
- **Additional allocation of smaller "development projects" for those teams which are anticipated to progress to leadership-class status.**

Leadership Computing for Science

Critical for success in key national priorities

Office of Science research priorities

Manipulating the Nanoworld



Computational design of innovative nanomaterials

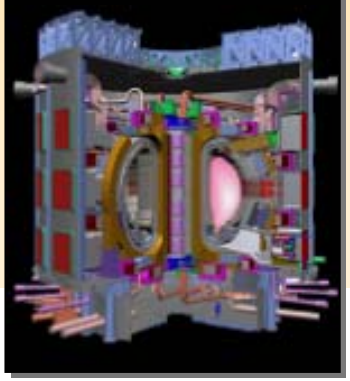
Taming the Microbial World

Predictive understanding of microbial molecular and cellular systems

Environment and Health

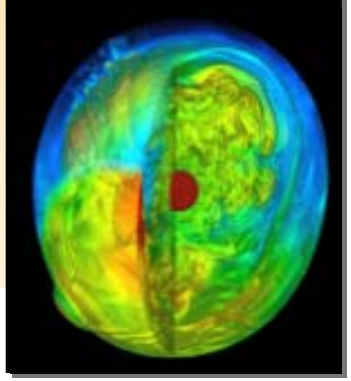
Full carbon cycle in climate prediction, IPCC

ITER for Fusion Energy



Simulation of burning plasma, Fusion Simulation Project

Search for the Beginning



Terascale Supernovae Simulation

Theory, Mathematics, Computer Science

National Leadership-Class Computing Facility for Science