High-Performance Computing Research at Oak Ridge National Laboratory

Dr. Christian Engelmann

Computer Science and Mathematics Division
Oak Ridge National Laboratory, USA

Largest Multipurpose Science Laboratory within the U.S. Department of Energy



- Privately managed for US DOE
- $1.1 billion budget
- 4200 employees total
- 3,000 research guests annually
- 30,000 visitors each year
- Total land area 58mi^2 (150km^2)

- Nation's largest energy laboratory
- Nation's largest science facility:
  - The $1.4 billion Spallation Neutron Source
- Nation's largest concentration of open source materials research
- Nation's largest open scientific computing facility

ORNL East Campus: Site of World Leading Computing and Computational Sciences



National Center for Computational Sciences



- 40,000 ft^2 (3700 m^2) computer center:
  - 36-in (~1m) raised floor, 18 ft (5.5 m) deck-to-deck
  - 12 MW of power with 4,800 t of redundant cooling
  - High-ceiling area for visualization lab:
    - 35 MPixel PowerWall, Access Grid, etc.

- 4 systems in the Top 500 List of Supercomputer Sites:
  - Jaguar: 2. Cray XT5, MPP with 37544 quad-core Processors ⇨ 1 381 TFlop
  - Jaguar: 8. Cray XT4, MPP with 7744 quad-core Processors ⇨ 260 TFlop
  - Kraken: 15. Cray XT4, MPP with 4489 quad-core Processors ⇨ 165 TFlop
  - Eugene: 130. IBM Blue Gene/P, MPP with 2048 quad-core Processors ⇨ 27 TFlop



Jaguar: World's most powerful computer— Designed for science from the ground up



Peak performance	**1.645 PF**
System memory	362 TB
Disk space	10.7 PB
Disk bandwidth	240+ GB/s
Interconnect bandwidth	532 TB/s

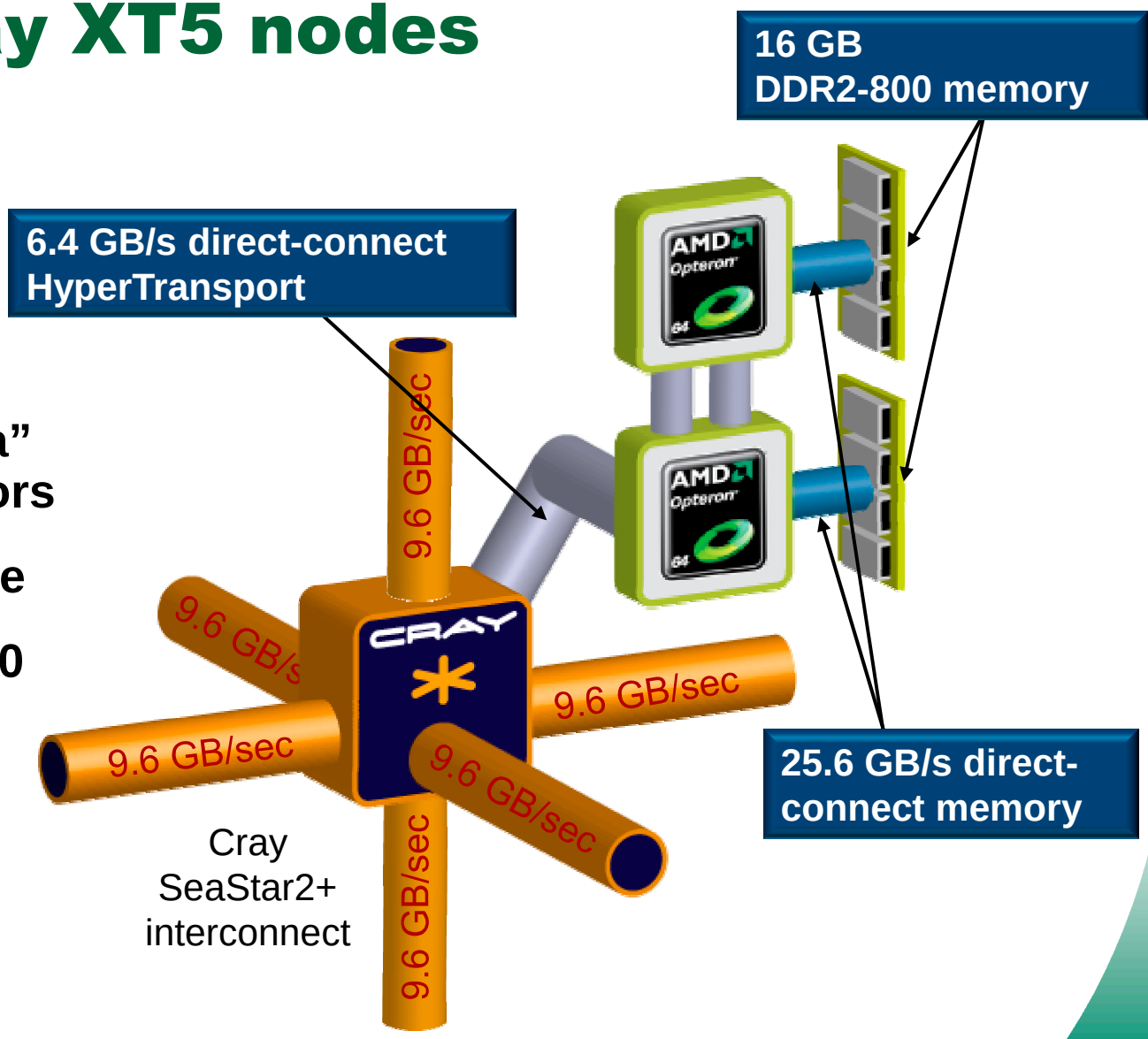
What does the system look like?

Jaguar combines the existing 263 TF Cray XT4 system at ORNL's NCCS with a new 1,382 TF Cray XT5 to create a 1.64 PF system

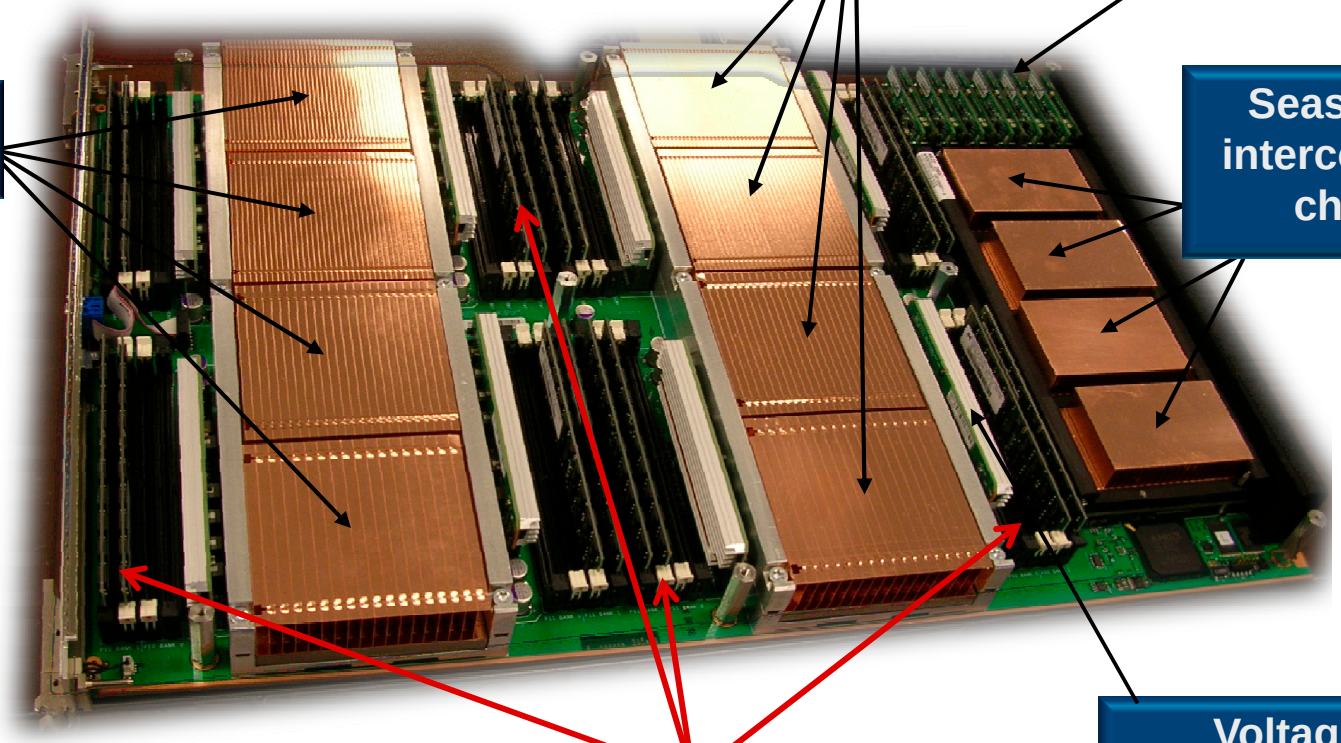
System attribute	XT5	XT4
Quad-core AMD Opteron processors	37,544	7,832
Node architecture	Dual socket SMP	Single Socket
Memory per core/node (GB)	2/16	2/8
Total system memory (TB)	300	62
Disk capacity (TB)	10,000	750
Disk bandwidth (GB/s)	240	44
Interconnect	SeaStar2+ 3D torus	SeaStar2+ 3D torus

Jaguar's Cray XT5 nodes

- Eight-core SMP
- Two 2.3 GHz AMD Opteron "Barcelona" quad-core processors
- 73.6 Gflops per node
- Four 4 GB DDR2-800 DIMMS with four empty DIMM slots per node
- OpenMP Support



Jaguar's Cray XT5 blades



- 4 nodes per blade
- 294.4 GF
- 64 GB
- 4 network connections

24 Cray XT5 blades make a Jaguar cabinet



One Jaguar XT5 cabinet:

- 7 TF
- 192 Opteron processors
- 776 Opteron cores
- 1,536 GB memory
- 96 nodes
- 1•4•24 section of 3D torus

200 cabinets complete the XT5 side of the system



Compute partition

- Nodes: 18,772
- Opteron processors: 37,544
- Processor cores: 150,176
- Peak TF: 1.38
- Memory: 300 TB
- Interconnect: 25•32•24 3D torus

Service and I/O partition

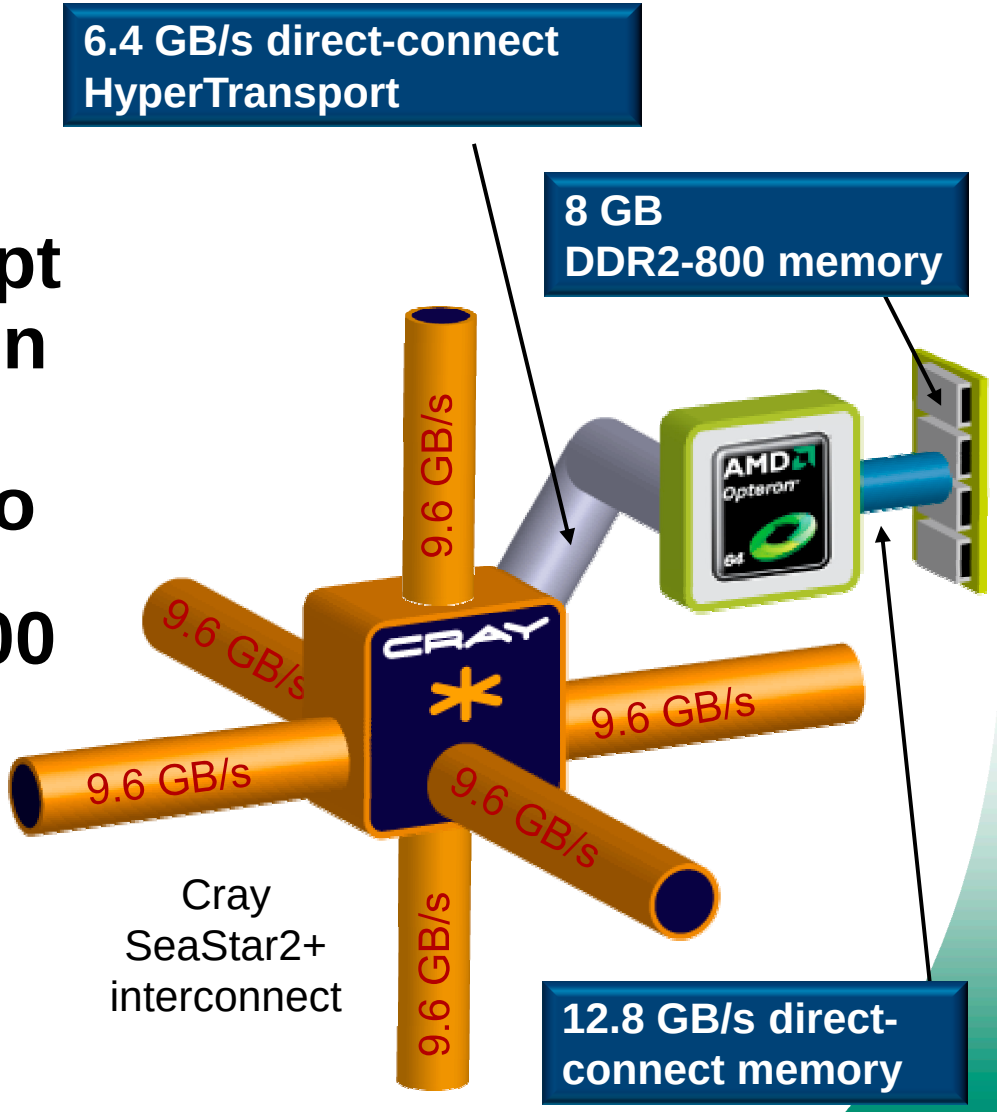
- Nodes: 214
- InfiniBand cards: 192 DDR
- Memory per SIO node: 8 GB
- 10 Gigabit Ethernet: 10

Power and cooling

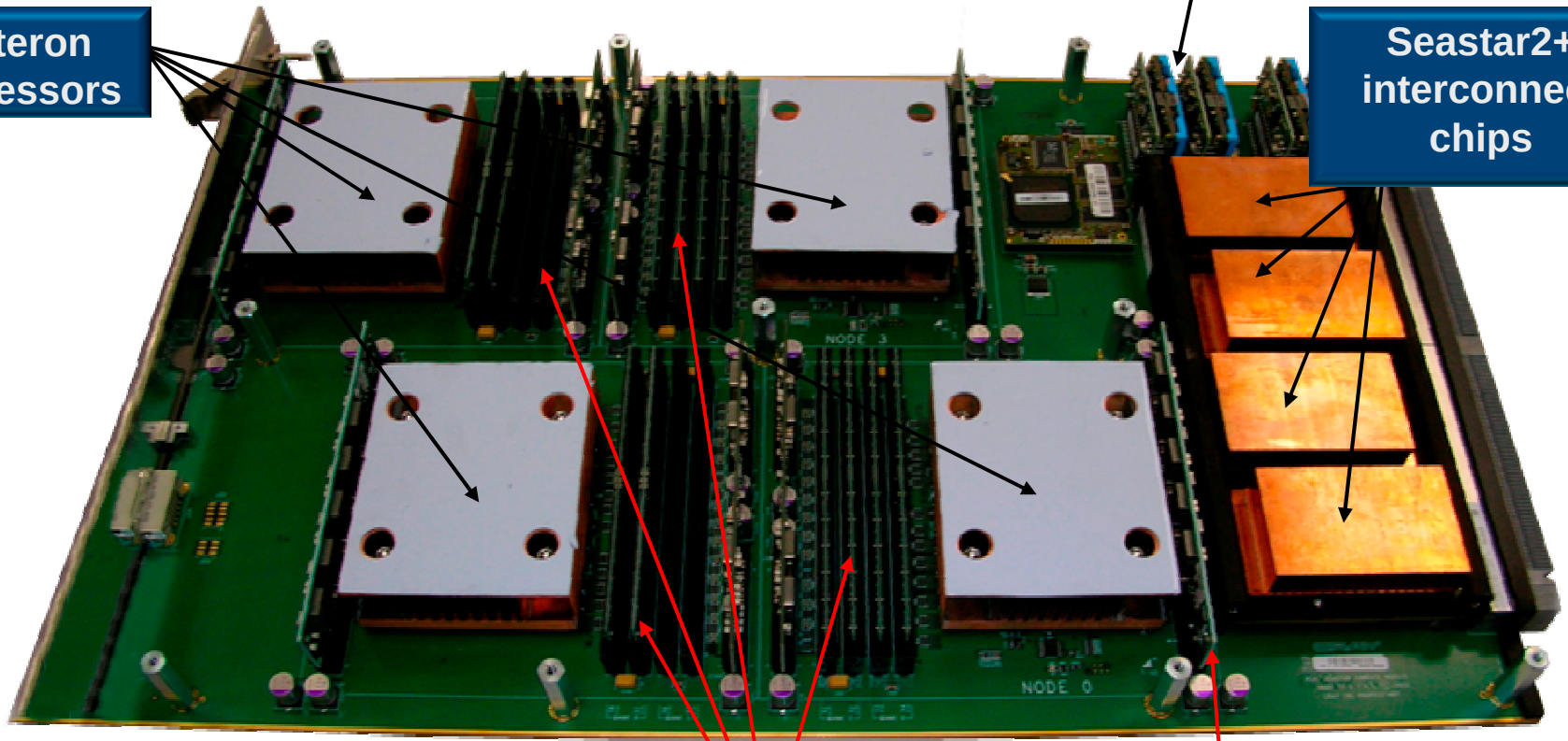
- 480 V power supplies
- R134a refrigerant
- Single large fan per cabinet

How is the XT4 side different?

- **Basic node design is the same as XT5, except it has one AMD Opteron quad-core processor per node instead of two**
- **Same 2 GB of DDR2-800 memory per core**
- **Same Seastar2+ interconnect**

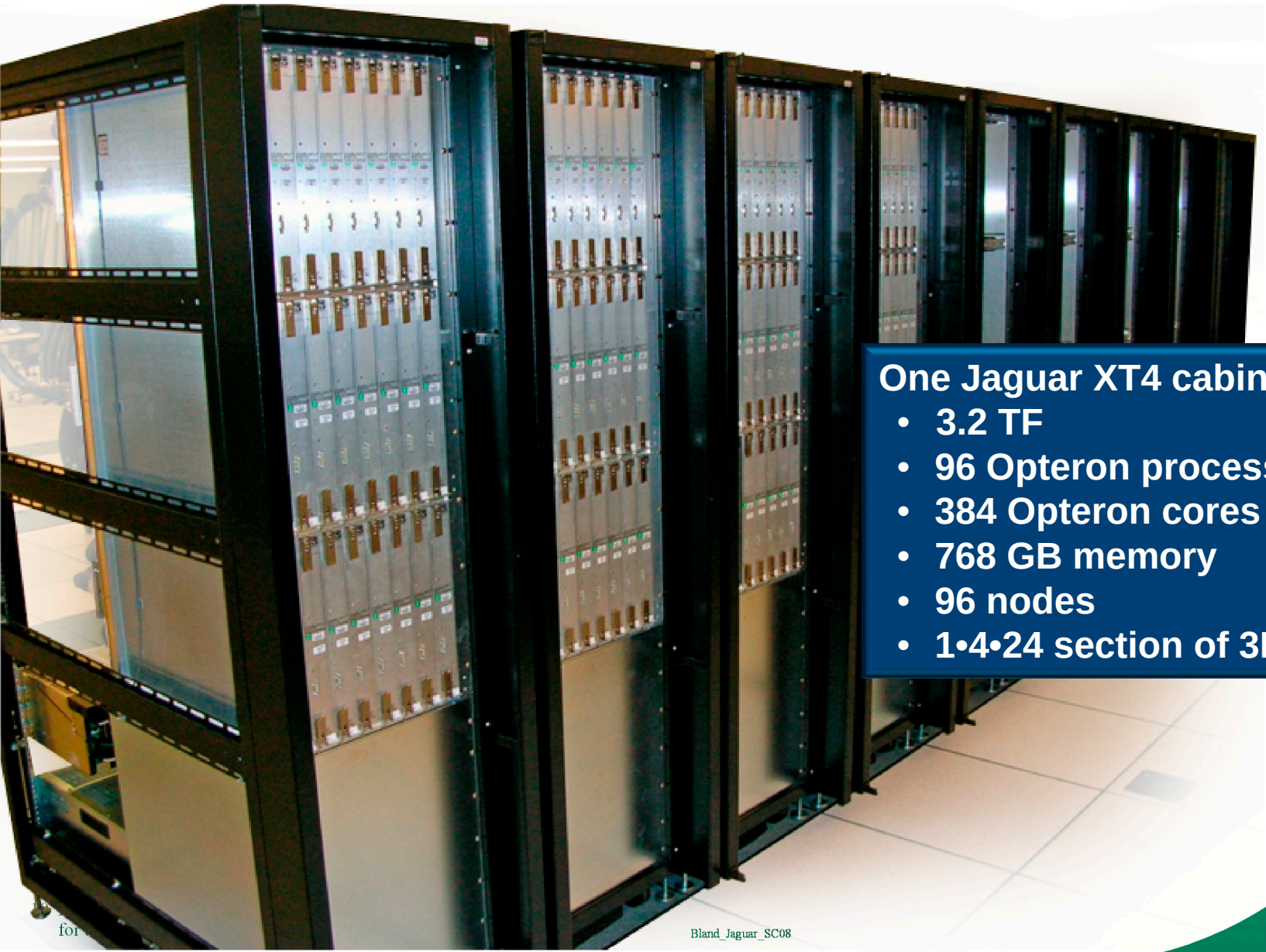


The XT4 board has four nodes



- 4 nodes per blade
- 134.4 GF
- 32 GB
- 4 network connections

And 24 Cray XT4 blades make a Jaguar cabinet



One Jaguar XT4 cabinet:

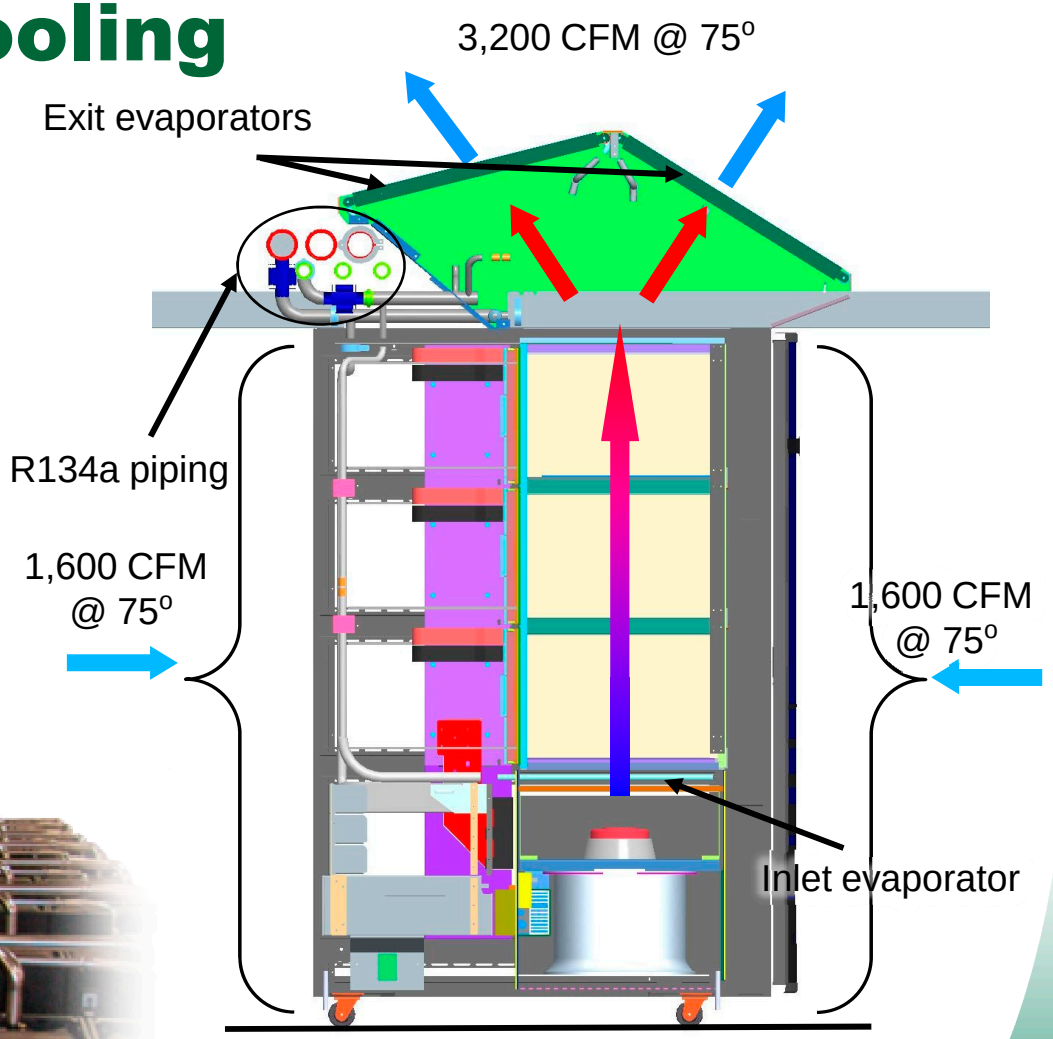
- 3.2 TF
- 96 Opteron processors
- 384 Opteron cores
- 768 GB memory
- 96 nodes
- 1•4•24 section of 3D torus

OAK RIDGE National Laboratory

ECOphlex liquid cooling

- **Liquid-cooled design exhausts heat to R134a before it leaves the cabinet. Replaces 100 CRAC units!**
- **Saves about 900 KW of power in air movement alone**





- **Phase change of liquid to gas removes heat much more efficiently**

OAK RIDGE National Laboratory

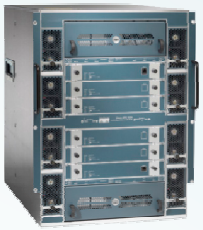
480 V power saves money

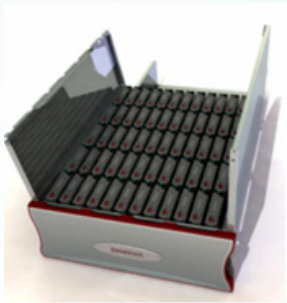


- Power supply efficiency
  - Worked with Cray to select highest-efficiency power supplies
- 480 V power to the cabinet saves the inefficiency of converting from 480 V to 208 V to 48 V
- Keeping the voltage high saved $1M in installation and $500K in electrical costs

Centerwide file system







- "Spider" is being installed to provide a shared, parallel file system for all systems
  - Based on Lustre file system
- Bandwidth of more than 240 GB/s
- More than 10 PB of RAID6 Capacity
  - 13,440 1 TB SATA drives
- 192 storage servers
  - 3 TB of memory
  - 14 TF
- Available from all systems via our high-performance scalable I/O network
  - More than 3,000 InfiniBand ports
  - More than 3 miles of cables
  - Scales as storage grows
- Engineered for high availability

Jaguar combines a new 1.38 PF Cray XT5 with the existing 263 TF Cray XT4





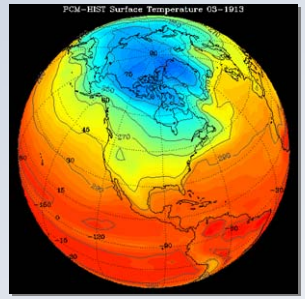


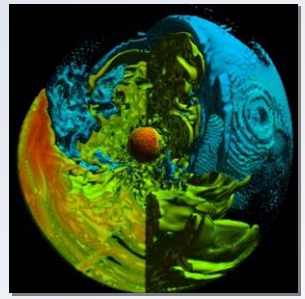


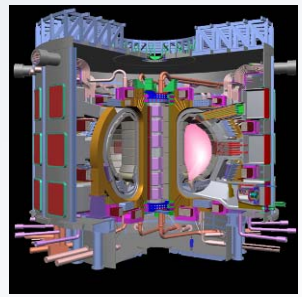
System components are linked by 4×-DDR InfiniBand (IB) using three Cisco 7024D switches

- **XT5 has 192 IB links**
- **XT4 has 48 IB links**
- **Spider has 192 IB links**

At Forefront in Scientific Computing and Simulation









- Leading partnership in developing the National Leadership Computing Facility
  - Leadership-class scientific computing capability
  - 1 PFlop/s in 2008 (upgrade recently finished)
  - 500 TFlop/s in 2009 (installation in progress)
  - 1 PFlop/s in 2009 (commitment made)
- Attacking key computational challenges
  - Climate change
  - Nuclear astrophysics
  - Fusion energy
  - Materials sciences
  - Biology
- Providing access to computational resources through high-speed networking

ORNL provides leadership computing to 2008 INCITE program

- The NCCS is providing leadership computing to 30 projects in 2008 under DOE's Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program
- Leading researchers from government, industry, and the academic world will use more than 75 million processor hours on the center's Cray leadership computers
- The center's Cray XT4 (Jaguar) and Cray X1E (Phoenix) systems provide more than 75% of the computing power allocated for the INCITE program



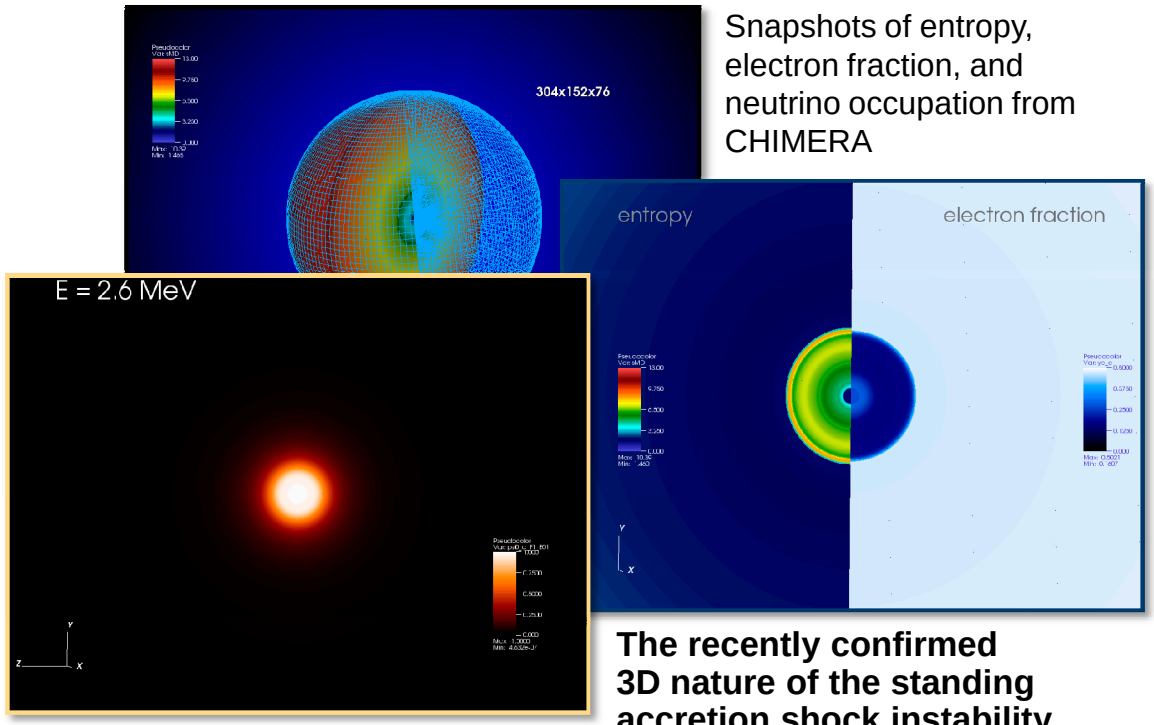
Project allocations: 145.3 million hours

Industrial allocations: 11.9 million hours

Astrophysics

Three-dimensional supernova explosions simulations with high fidelity physics

- The world's first 3D supernova simulations incorporating multifrequency neutrino transport are currently under way on Jaguar
- 304×152×76 spatial mesh running on 11,552 CPUs
- Initial results are promising, but full evolution of the explosion epoch will require millions of hours



Snapshots of entropy, electron fraction, and neutrino occupation from CHIMERA

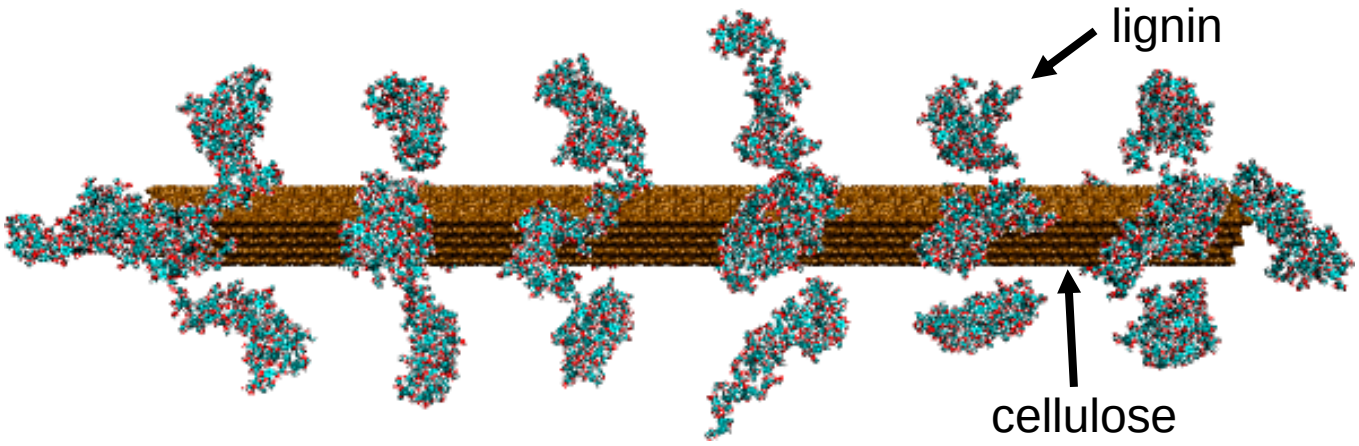
The recently confirmed 3D nature of the standing accretion shock instability and the known 3D nature of convective overturn make performing supernova simulations in 3D essential

LCF liaison contributions

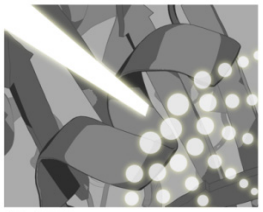
- Implementing efficient, collective I/O
- New visualization tools
- Improved software management

Biology

Cellulosic ethanol: Physical basis of recalcitrance to hydrolysis of lignocellulosic biomass



Simulations provide physical insight into physical origins of biomass recalcitrance and cellulosome organization and assembly



Center for Molecular Biophysics

ORNL

Loukas Petridis, Benjamin Lindner, and Jeremy C. Smith

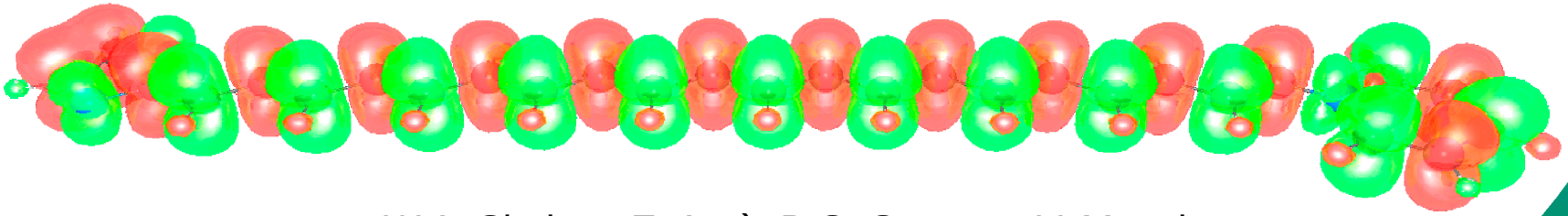
See L. Petridis and J.C. Smith, *J. Comp. Chem*. In press.

- **Understanding biomass structure is key to overcoming recalcitrance**
- **Large-scale molecular dynamics simulation (1−3M atoms)**
  - **Currently using NAMD code with 1.9M atoms for lignocellulose (on 6072 cores)**
- **Results to be used to interpret biophysical experiments**
  - **Lignin-cellulose simulations will be used to calculate small-angle neutron scattering intensities, which will be compared with experimental data obtained at the Spallation Neutron Source**

Chemistry
Organic molecular magnets from zwitterionic molecules

PI: Robert Harrison, ORNL

- **Hybrid DFT calculations of the electronic structure of a new class of zwitterionic molecules (betaine derivatives) reveals an unexpected magnetic ground state**
- **Nanochains with a maximum length of 10.4 nm and over 1700 basis functions. Over 500 Jaguar processors used. Scaling with respect to system size is $O(N^2)$**
- **Origin of the magnetism is quite complex and not well understood as compared to more conventional magnetic systems**
- **Magnetic materials that are purely organic are rather rare, and the betaine derivatives promise to offer a new class of molecular magnetic materials with diverse potential applications including organic spintronics, information storage, and nanoscale sensors**



W.A. Shelton, E. Aprà, B.G. Sumpter, V. Meunier

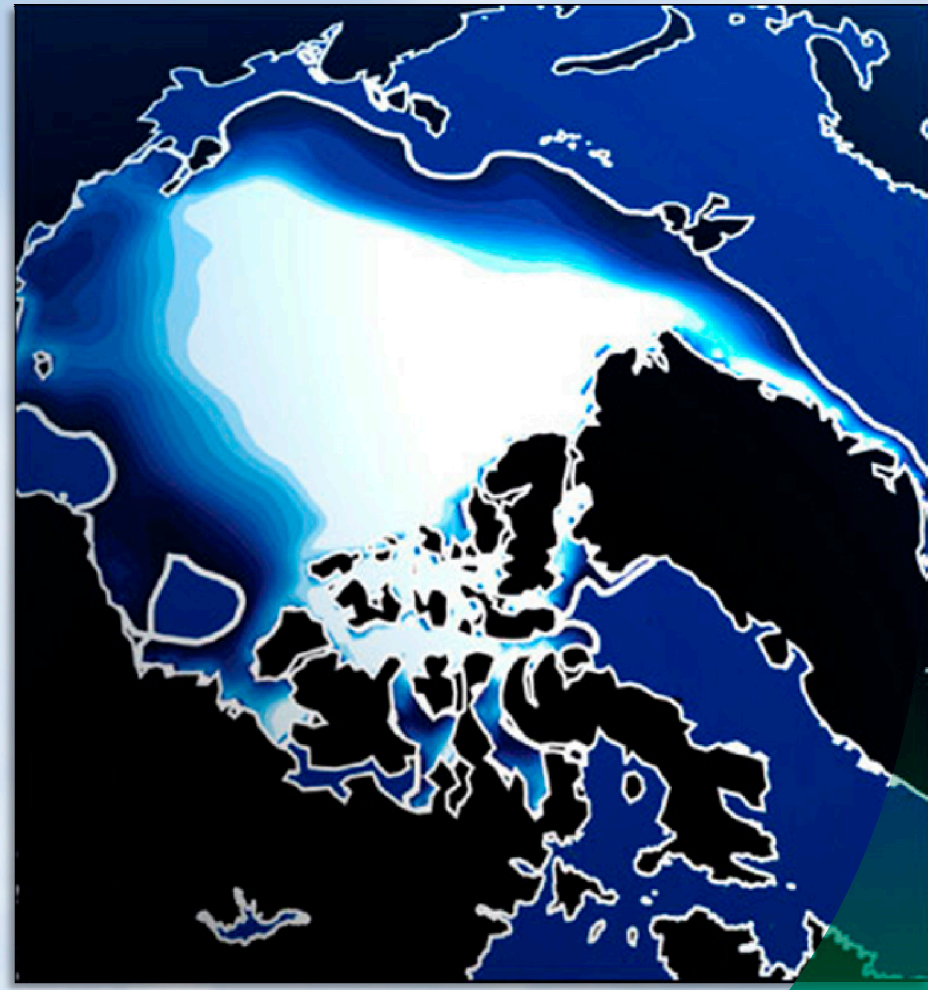
Climate

Modeling effects and agents of abrupt climate change

PI: Zhengyu Liu, University of Wisconsin

This project is addressing two fundamental questions on future climate change

- **What is the sensitivity of the climate system to the change of greenhouse gases, notably CO_2?**
- **How does the climate system exhibit abrupt changes on decadal-centennial timescales?**



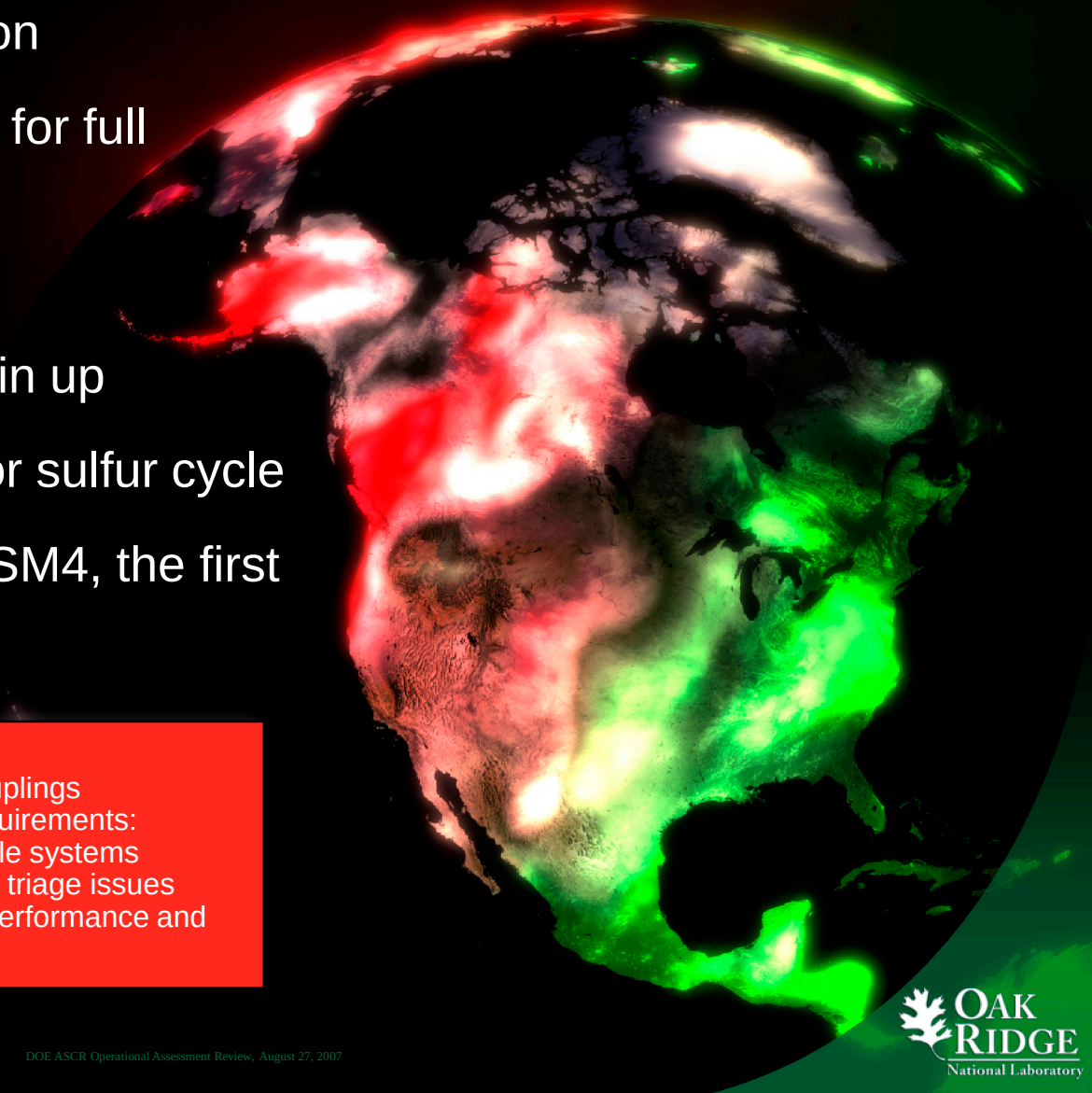
CCSM simulation of the 21,000-year last deglaciation

PI: Warren Washington, NCAR

Climate
Modeling the full Earth system

- New standard and low-emission climate scenarios at higher resolution
- Final stage of ocean spinup for full Earth System Model
- 100 years of ocean spin up
- 20 years of carbon-cycle spin up
- Approaching steady state for sulfur cycle
- Approaching release of CCSM4, the first Earth System Model

Simulated time evolution of the atmospheric CO_2 concentration originating from the land's surface



LCF liaison contributions
- Advanced visualizations of new climate couplings
- Tuning of LCF infrastructure for climate requirements: special queues, reservations, dedicated file systems
- Weekly conference calls with developers to triage issues
- Development of timing libraries to assess performance and balance components

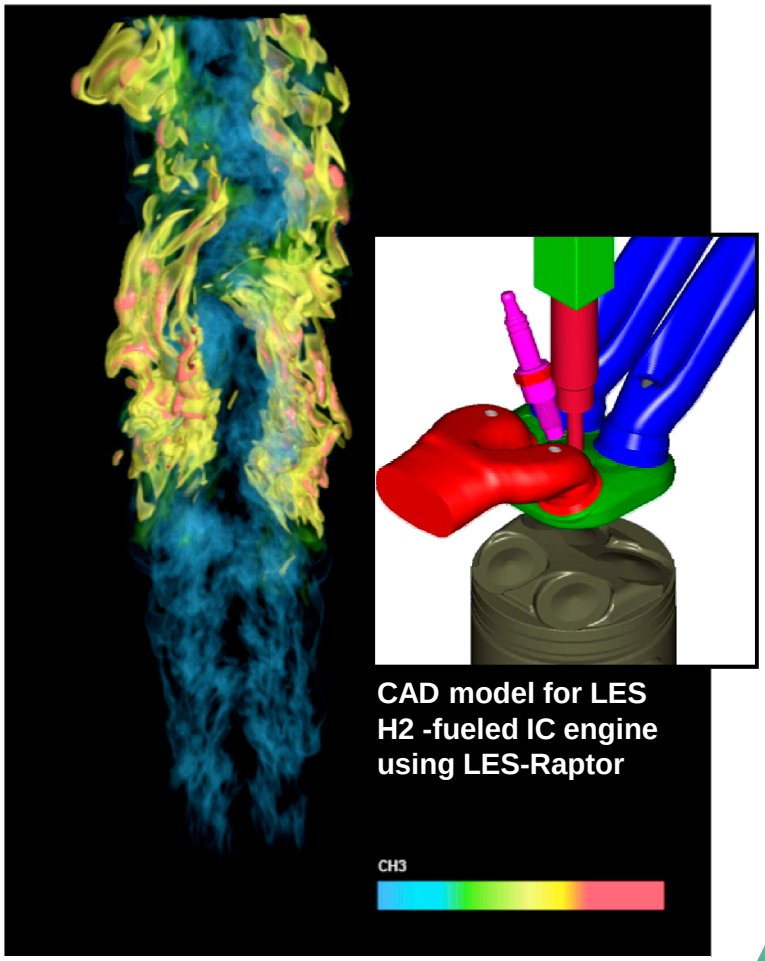
Engineering

Flame stabilization in a lifted hydrocarbon flame

- Lifted flames occur in diesel engines and gas turbine combustors
  - Flame stabilized against fuel jet and recirculating hot gases
- Source of flame stabilization was revealed in a simulation of a hydrogen-air lifted flame
  - Upstream autoignition
  - Vorticity generation at flame base due to baroclinic torque
- Direct numerical simulation of a 3D lifted autoigniting ethylene-air jet flame
  - Reynolds number = 10,000
  - 1.3B grid points and 22 chemical species
  - 50 TB of data generated
- Thirteen publications/presentations in just last quarter!

LCF liaison contributions

- Distributed Lagrangian particle-tracking algorithm and S/W
- Ethylene chemistry performance improvement

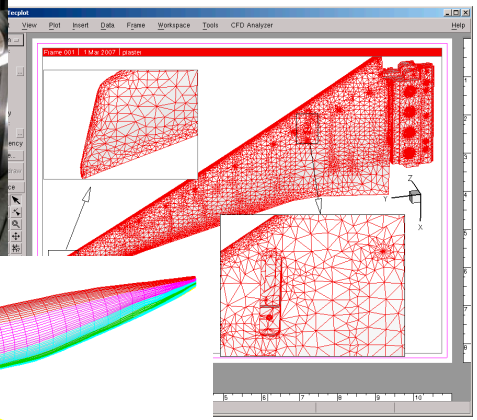


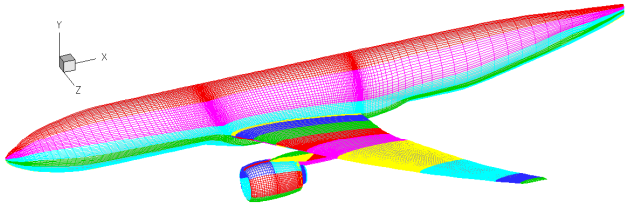
Instantaneous concentration of methyl radical. Low concentrations of methyl upstream of the high-temperature flame base serve as a marker of ignition induction chemistry

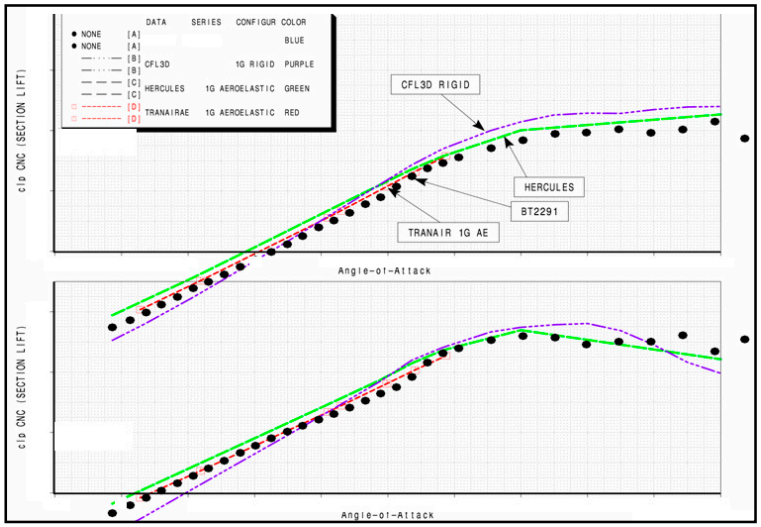
Engineering

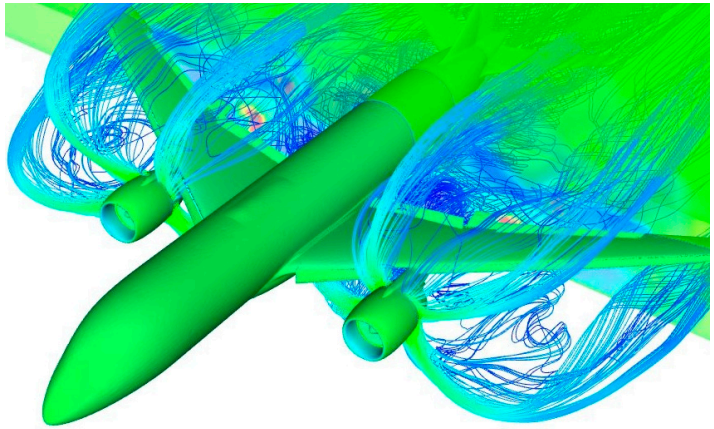
Development and correlations of large-scale computational tools for flight vehicles











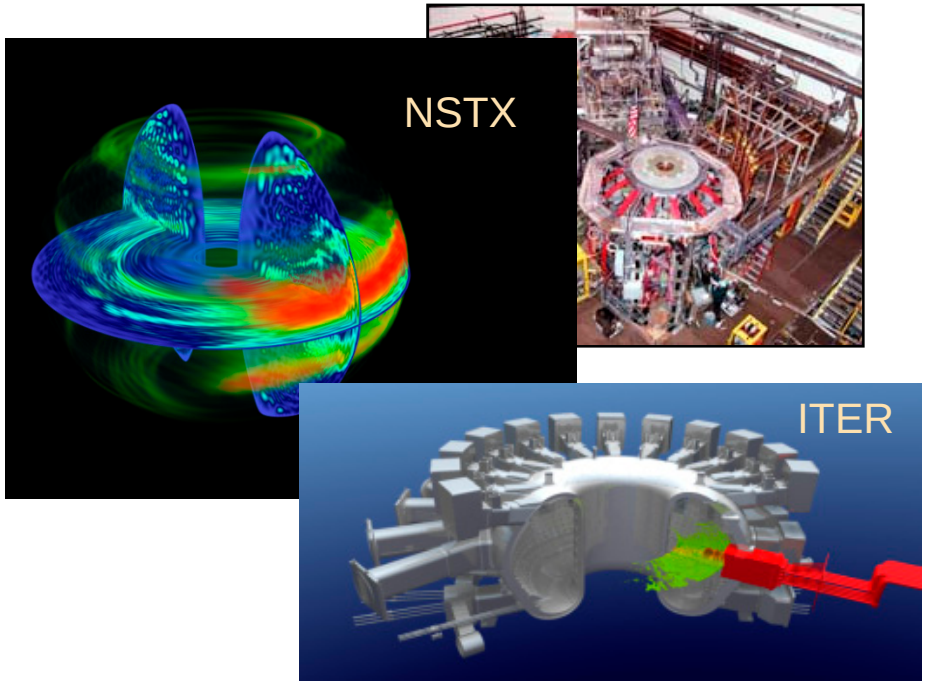
- Rare validation of unsteady Navier-Stokes tools for “simple” unsteady pitch oscillation motion
- Demonstration of the applicability and predictive accuracy of a CFD-based aeroelastic analysis process
- Confidence level improvement for rigid CFD results (toward its use in product downstream processes)
- Testing and development of new methods and methodologies (blade-out analysis process)
- Investigations of unsteady methods for better predictions (e.g. unsteady detached eddy simulations/DES)
- Evaluation of wind-tunnel effects

Fusion

Producing new insights for RF heating of ITER plasmas

Three-dimensional simulations of RF heating in the ITER fusion reactor as well as in present tokamaks (NSTX) shed new light on the behavior of superheated ionic gas

- When deuterium (D) or helium-3 (He^3) are used to damp the launched waves, they are accelerated to high energies, forming supra-thermal tails that significantly affect the wave propagation and absorption
- Energetic minority D ions enhance the fusion reaction rate
- Energetic He^3 ion tails form on both the tritium and He^3 distributions



- **3D simulations reveal new insights**
  - **"Hot spots" near antenna surface**
  - **"Parasitic" draining of heat to the plasma surface in smaller reactors**
- **Work pushing the boundaries of the system (28,900 cores, 154 TF) and demonstrating**
  - **Radial wave propagation and absorption**
  - **Efficient plasma heating**
- **AORSA's predictive capability can be coupled with Jaguar power to enhance fusion-reactor design and operation for an unlimited clean energy, source**

LCF liaison contributions

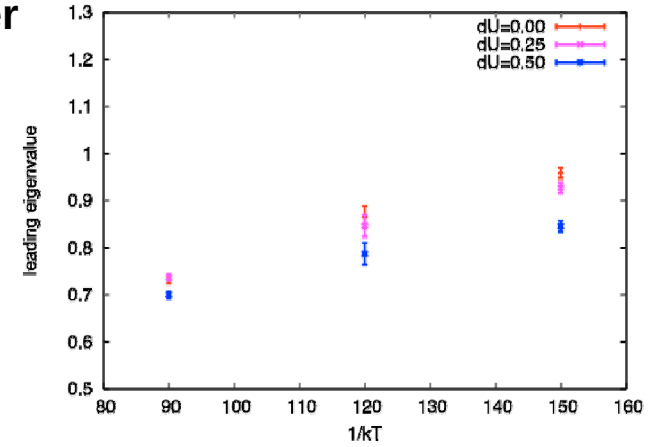
- **Converted HPL from double real to double complex and replaced ScaLAPACK**
- **Acquired new version of BLAS from TACC**
- **Net performance gain of a factor of two**

Materials and Nanoscience

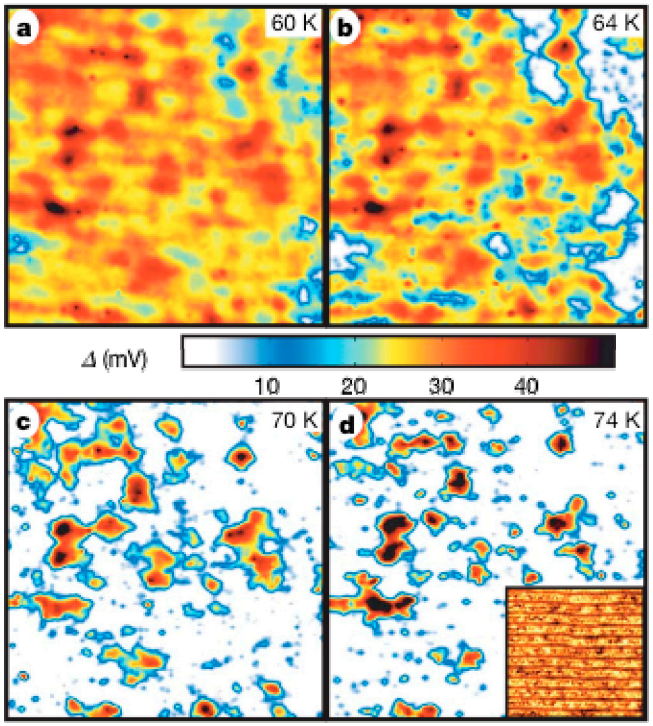
Role of nanoscale inhomogeneities in high-temperature superconductors

PI: Thomas Schulthess, ORNL

- Investigate role of inhomogeneities on pairing mechanisms and transition temperature—petascale computing problem
- Develop materials-specific extensions to the Hubbard model to understand variations of transition temperatures—develop materials design tool
- Simulations of 2D Hubbard model with disorder in local Coulomb interaction parameter (Ui = 1 +/- dU) indicate *Tc* is indeed suppressed due to disorder



- Further analysis of the role of inhomogeneities (substitutions, vacancies, etc.) is now possible in the simulations with 16- and 24-site clusters



Temperature evolution of the superconducting gap taken on a 300 A area of a cuprate with T_c = 65K [reproduced from Gomez et al., *Nature* **447**, 569−572 (2007)]. The gap varies spatially on a scale of 1−3 nm and persists in some regions to temperatures well above T_c, as can be seen from panels c and d

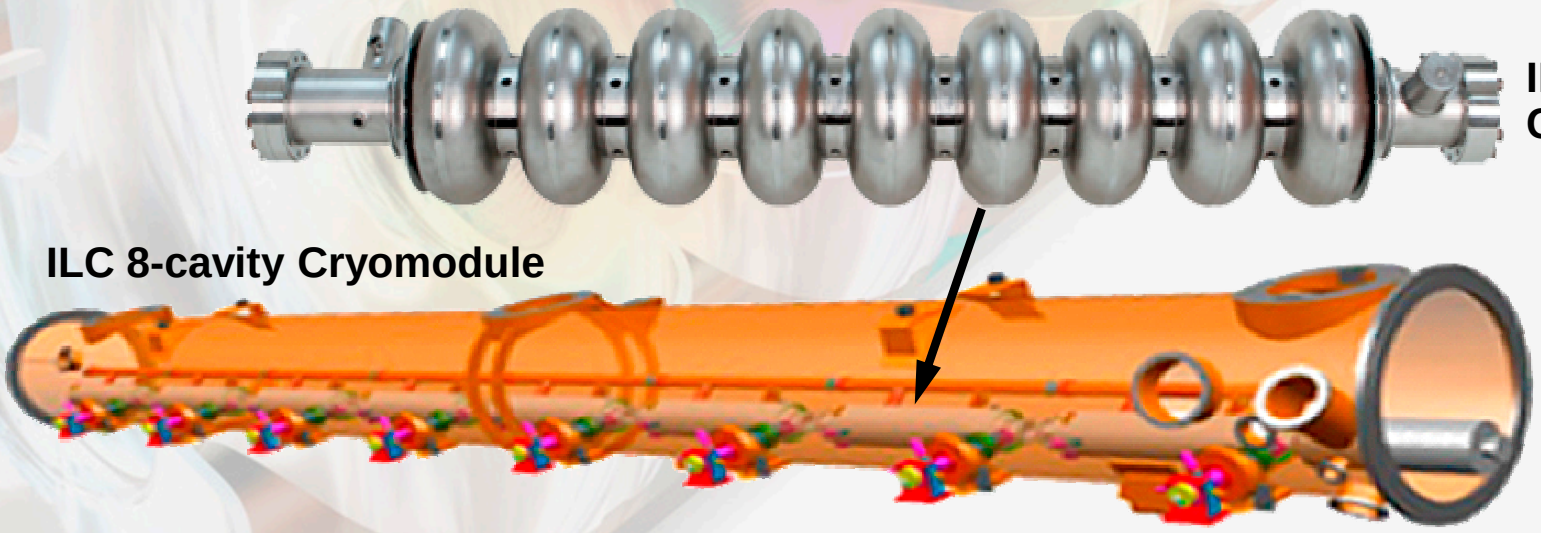
Simulations at 31K cores on Jaguar perform >50% of peak (Gordon Bell submission)

Petascale computing for terascale particle accelerator

International linear collider design and modeling

PI: Lie-Quan Lee, Stanford Linear Accelerator Center

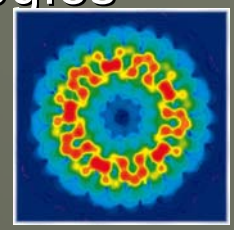
- The ILC is a particle accelerator aimed to address many of the most compelling physics questions about dark matter; dark energy; and the fundamental nature of matter, energy, space, and time
- Using the most advanced simulation tools (Omega3P, T3P), the computationally challenging problems associated with the ILC accelerator design will be tackled for improved machine performance, increased reliability, and reduced cost

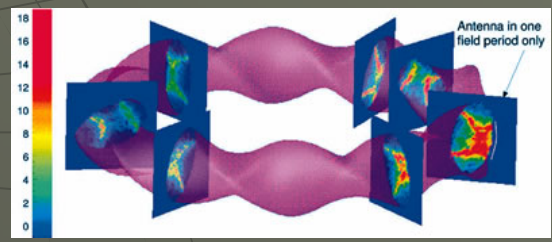


Computer Science Research

- Computer Science and Mathematics (CSM) Division.
  - Applied research focused on computational sciences, intelligent systems, and information technologies
- CSM Research Groups:
  - Climate Dynamics
  - Complex Systems
  - Computational Chemical Sciences
  - Computational Materials Science
  - Future Technologies
  - Statistics and Data Science
  - Computational Mathematics
  - *Computer Science Research*
    (~20 researchers, 3 postdocs/postmasters, students, …)





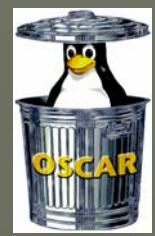


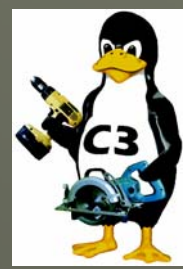
Computer Science Research Projects

- Parallel Virtual Machine (PVM)
- MPI Specification, FT-MPI and Open MPI
- Common Component Architecture (CCA)
- Open Source Cluster Application Resources (OSCAR)
- Scalable cluster tools (C3)
- Scalable Systems Software (SSS)
- Fault-tolerant metacomputing (HARNESS)
- High availability for HPC (FAST-OS MOLAR)
- System-level virtualization (VSE)
- Harness HPC workbench
- Resilience for HPC (FAST-OS RAS)



FT-MPI





System Research Team

- SRT team
  - Stephen L. Scott
  - Christian Engelmann
  - Hong Ong
  - Geoffroy Vallée
  - Thomas Naughton
  - Anand Tikotekar
- R&D topics
  - High availability
  - Fault tolerance
  - Operating systems
  - Cluster computing
  - Virtualization technologies
  - Resource management
  - Tools