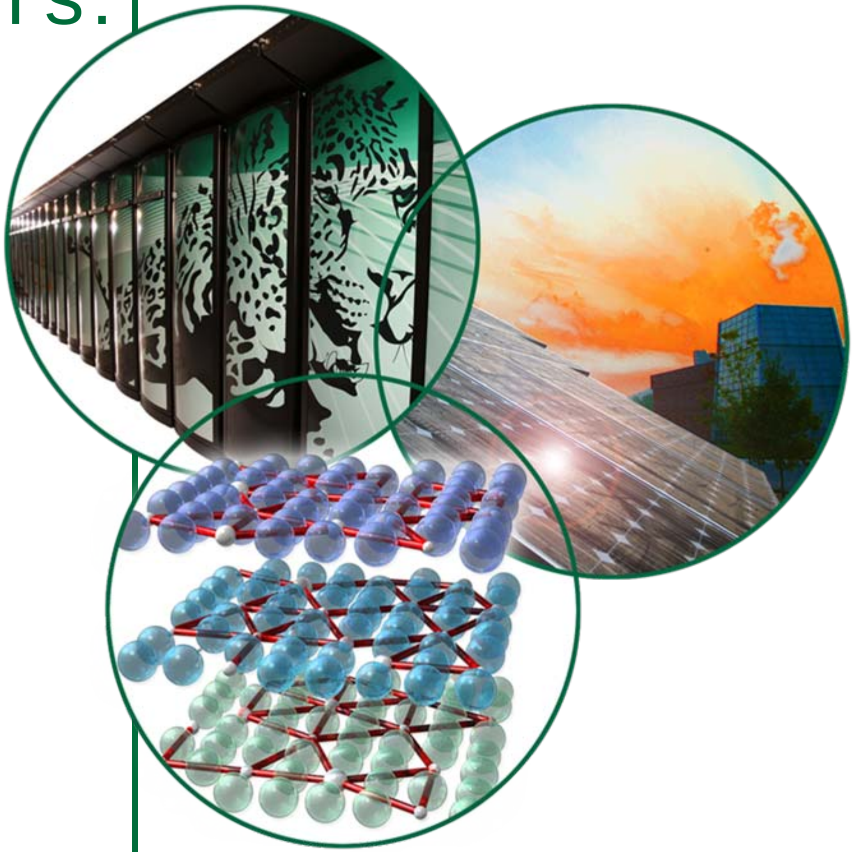
HPC Landscape - Application Accelerators: Deus ex machina?



Presented to the
High Performance Embedded Computing 2009 (HPEC)

Jeffrey Vetter
Future Tech Group Leader, ORNL
Joint Professor, Georgia Institute of Technology

22 September 2009

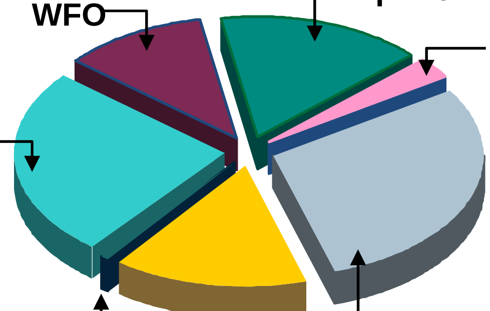
ORNL in the Early Days

- 1943 - Manhattan Project
  - Los Alamos
  - Oak Ridge
  - Hanford
- 1950-60s
  - Nuclear energy, medicine, materials
- 1970s - Divided into two sites
  - Y12 – NNSA
  - X10 – DOE Office of Science

Guards stop cars or trucks coming in or out of the Carbide & Carbon Chemicals Corp. plant.

ORNL Today

- Managed by UT-Battelle since April 2000
- $350M modernization project
- Key capabilities
  - Neutron science
  - Ultrascale computing
  - Systems biology
  - Materials science at the nanoscale
  - Advanced energy technologies
  - National security
- 4,600 staff
- 3,000 guest researchers
- 30,000 visitors
- FY08 budget: $1.4B



Spallation Neutron Source and Center for Nanophase Materials Sciences

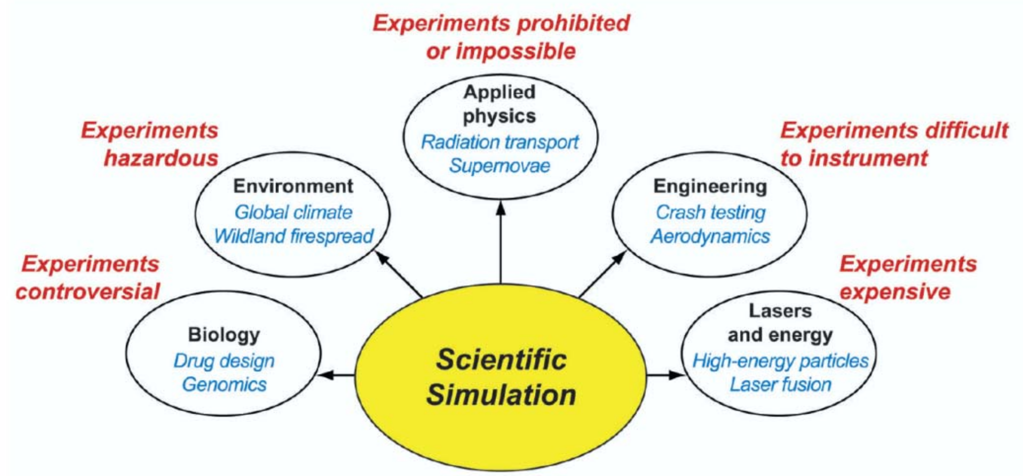
Mission: Conduct basic and applied research and development to create scientific knowledge and technological innovations that enable the solution of compelling national problems

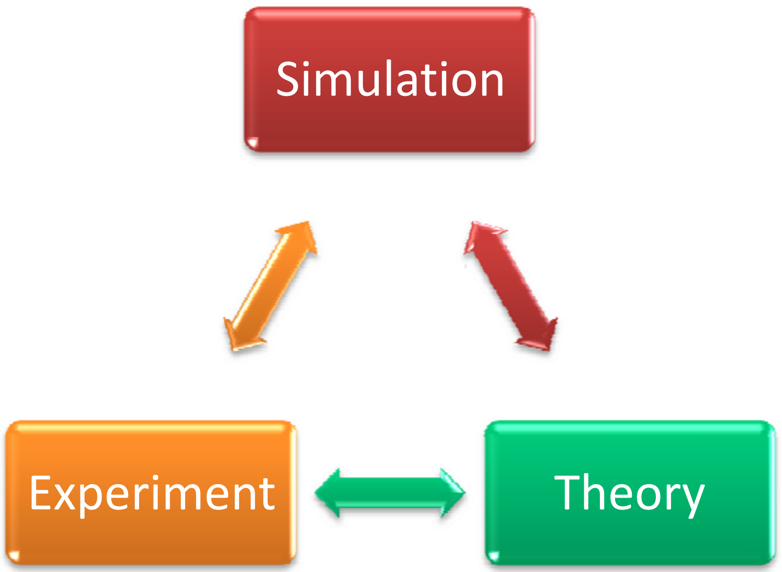
We have a three pronged strategy for sustained leadership and programmatic impact

- Provide the nation's most powerful open resource for capability computing
- Follow a well-defined path for maintaining national leadership in this critical area
- Deliver cutting-edge science relevant to the missions of key federal agencies
- Synergy of requirements and technology
- Unique opportunity for multi-agency collaboration for science

Scientific Simulation

- Simulation is becoming an accepted form of science
- Complex phenomena that
- Examples
  - Political limits
    - Nuclear test ban treaty
  - Experimentation limits
    - Climate modeling
    - Atomistic scale materials, etc

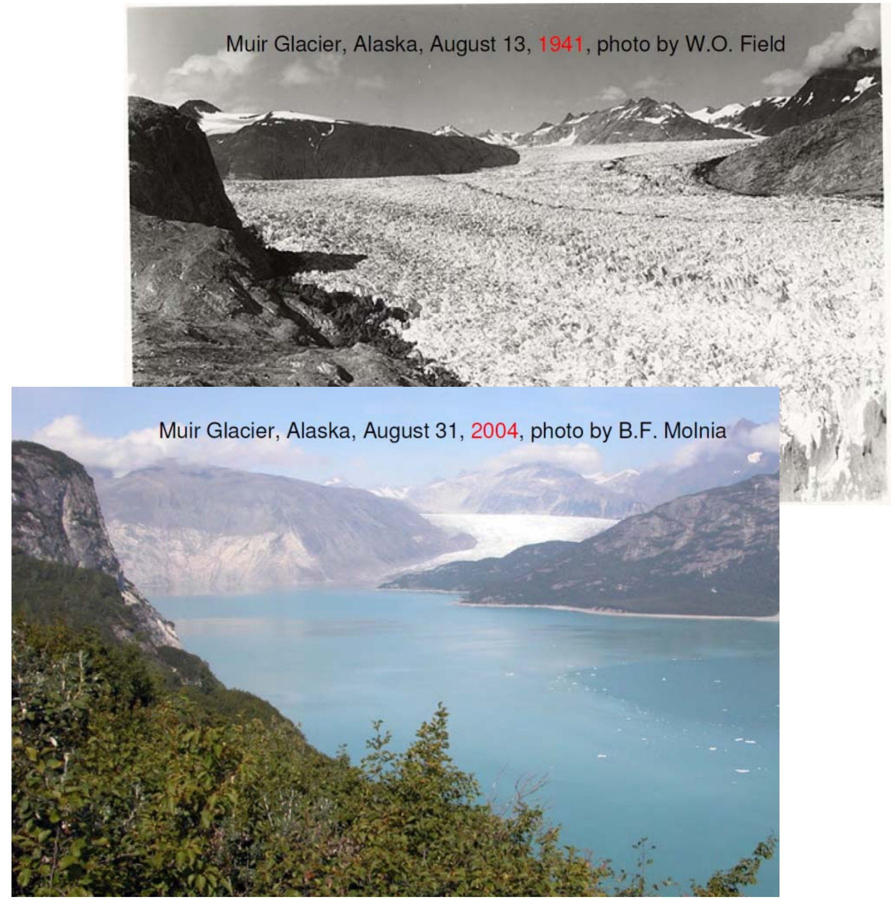




US Department of Energy Office of Science, "A Science-Based Case for Large-Scale Simulation," US Department of Energy Office of Science 2003.

Motivating Example: Climate Modeling

- Intergovernmental Panel On Climate Change
  - Sponsored by UNEP and WMO
  - Thousands of scientists from around the world
  - Won Nobel Prize w/ Gore in 2007
  - Provide policy-neutral scientific information
  - Assess basis, impact, adaptation, vulnerability, mitigation
- UN Conference TODAY

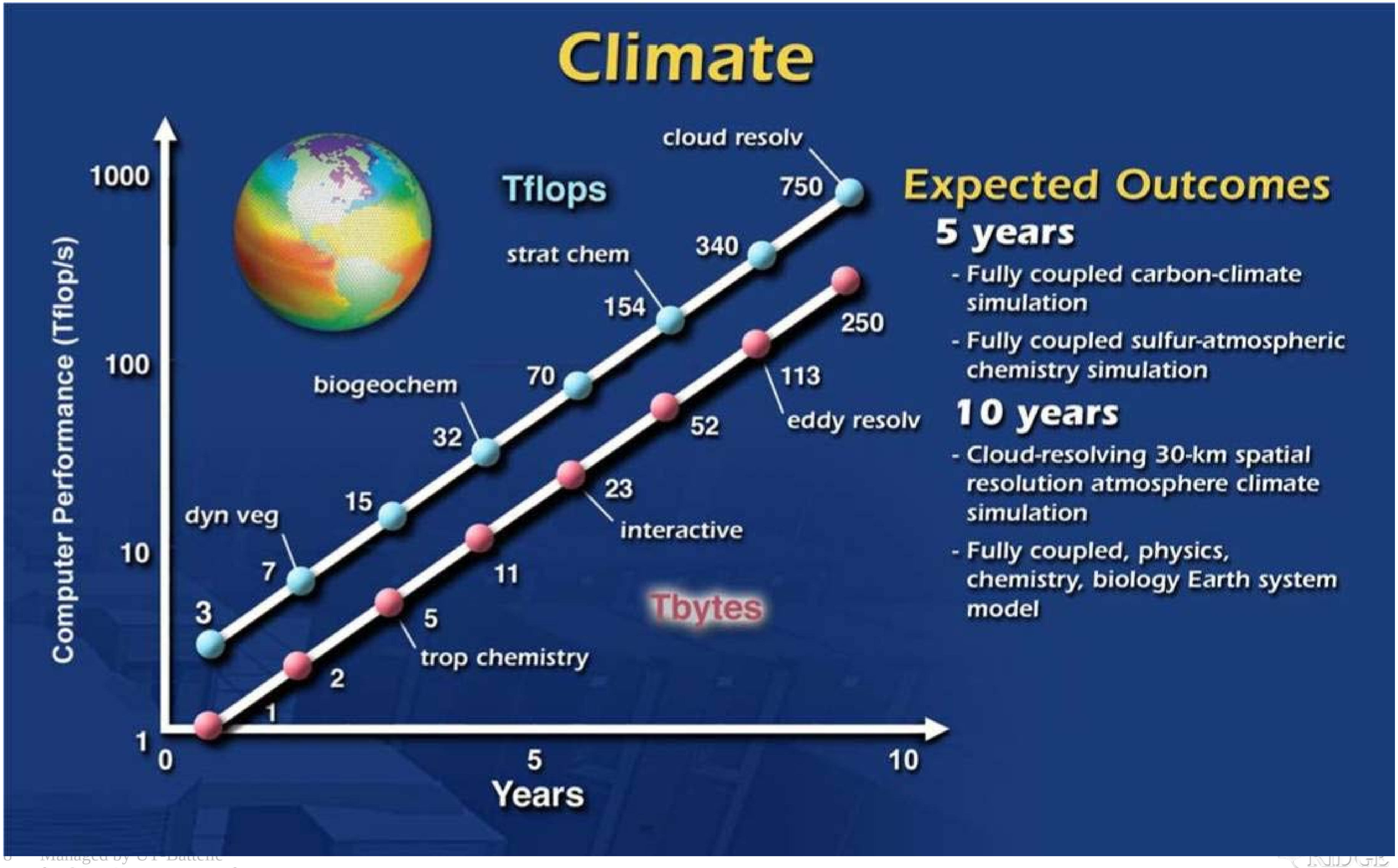


http://www.ipcc.ch

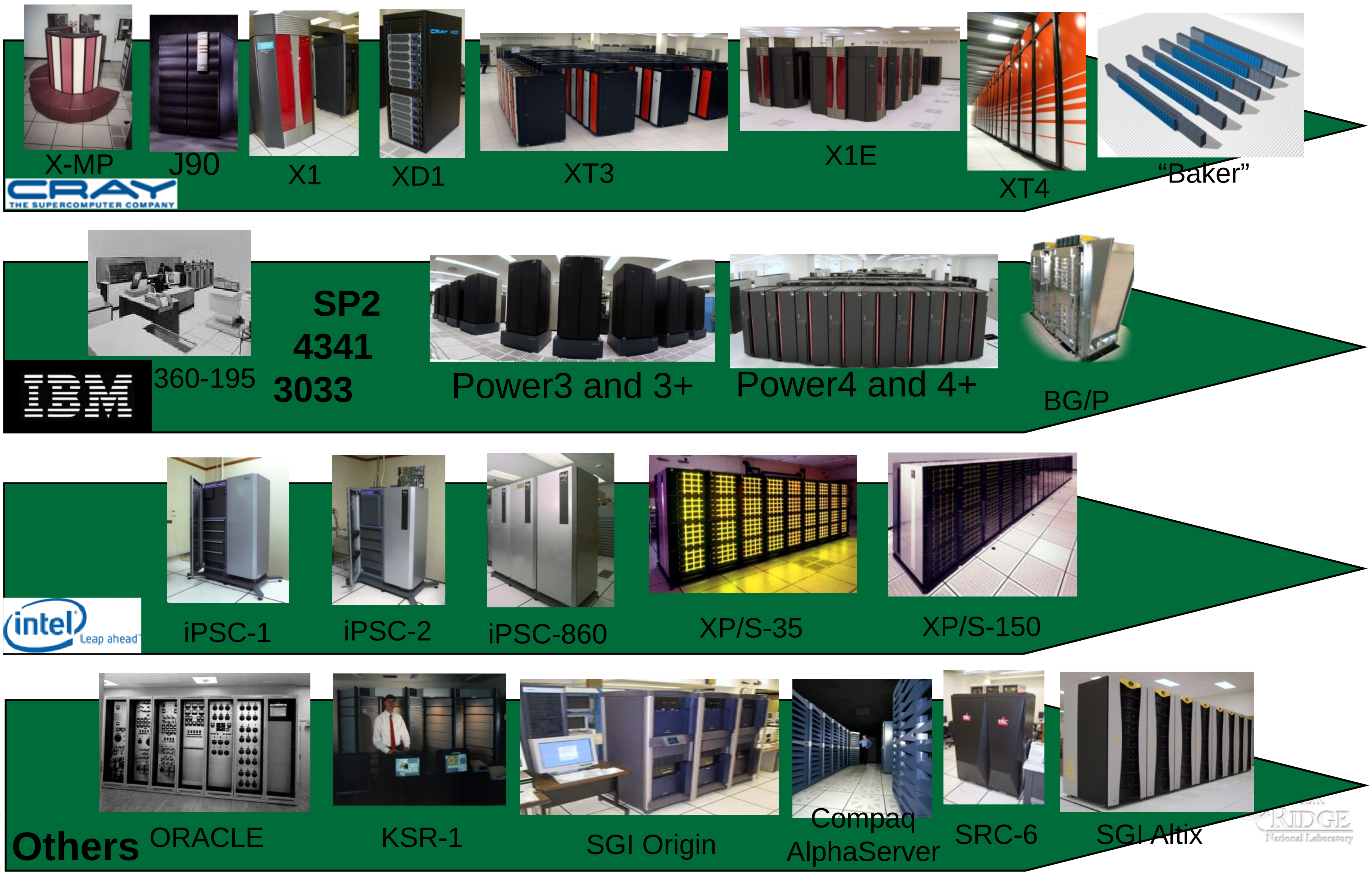
Climate Modeling - Future

- Observations and natural events provide effective tools
- Scientific simulation provides a critical tool to assess future scenarios
- Community Climate System Model is used in these assessments
- Simulation
  - ORNL has provided considerable computing resources over the past five years for IPCC simulations
  - NICS Cray XT5 at ORNL will be dedicated to climate simulations for 3 months
  - NOAA will provide $215M over the next 5 years to sponsor climate research at ORNL

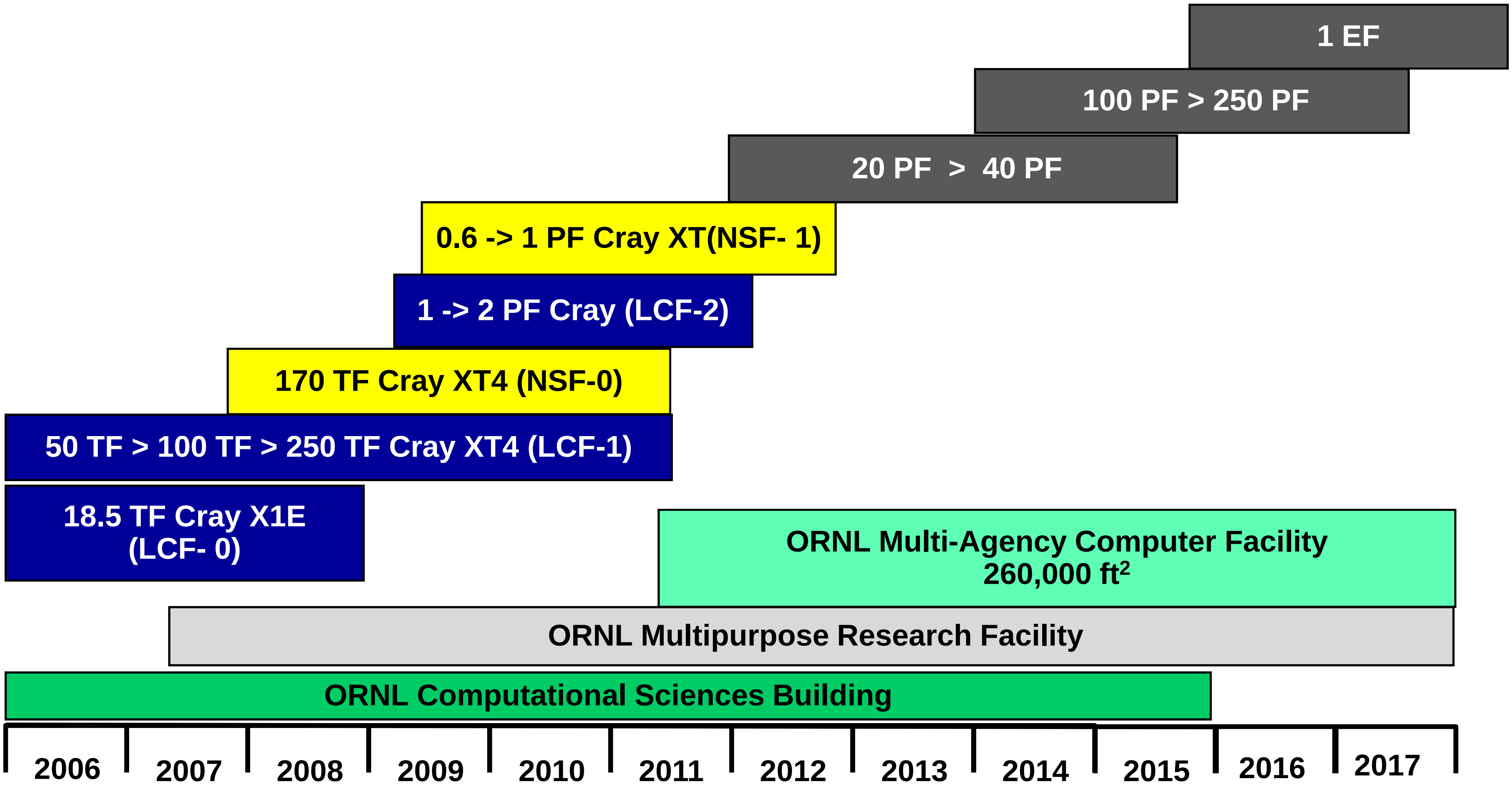
Climate Computational Projected Requirements



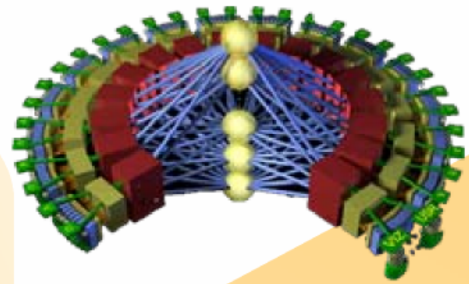
ORNL – Scientific and Technical Computing



ORNL Roadmap to Exascale



1000x increase in computing and data capabilities



2004

Cray X1
3 TF

2005

Cray XT3
Single-core
26 TF

2006

Cray XT3
Dual-core
54 TF

2007

Cray XT4
119 TF

2008

Cray XT4
Quad-core
263 TF
62 TB, 1 PB

Cray XT4
Quad-core
166 TF
18 TB, 0.3 PB

2009

Cray XT5
8/12-core, dual-socket SMP
1-2 PF
300 TB, 10 PB

Cray XT5
8/12-core, dual-socket SMP
~1 PF
100TB, 2.5 PB

2011

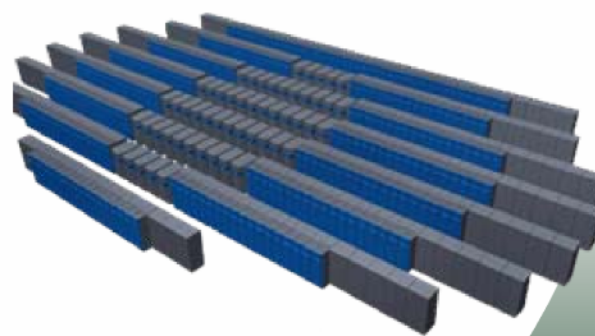
20 PF

2015

Future system
100–250 PF

2018

Future system
1000 PF
(1 EF)

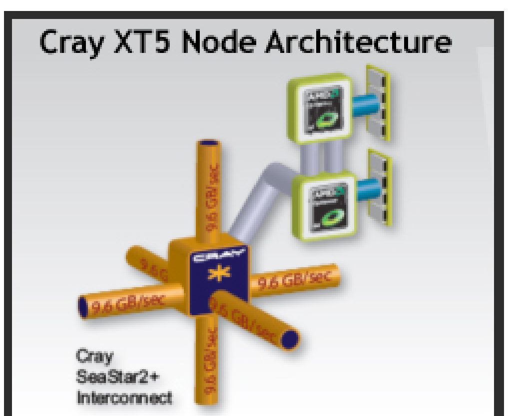


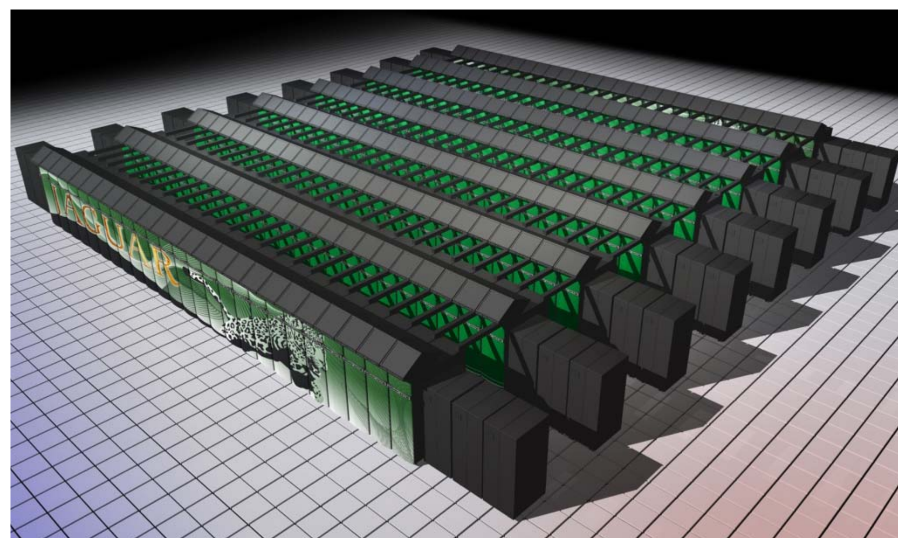
Resources available through
DOE INCITE and NSF TeraGrid

Oak Ridge National Laboratory

Jaguar Today

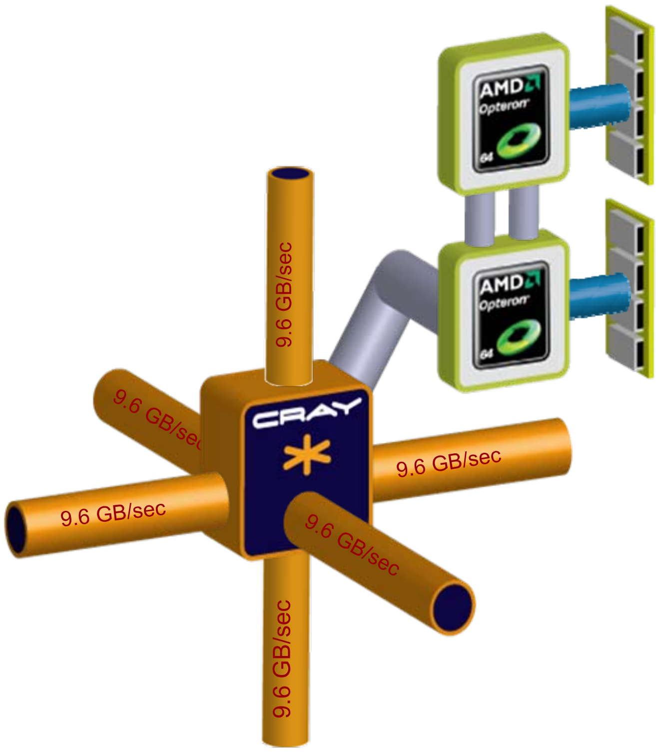
Jaguar Specifications	Total	XT5	XT4
Peak Teraflops	1,645	1,382	263
Quad-Core AMD Opterons	45,376	37,544	7,832
AMD Opteron Cores	181,504	150,176	31,328
Compute Nodes	26,604	18,772	7,832
Memory (TB)	362	300	62
Disk Bandwidth (GB/s)	284	240	44
Disk Space (TB)	10,750	10,000	750
Interconnect Bandwidth (TB/s)	532	374	157
Floor Space ($feet^2$)	5,800	4,400	1,400
Cooling Technology		Liquid	Air





Current Upgrade of XT5 to Istanbul Opteron is Underway – 3/5 of system already complete

- Upgrade of existing XT5 system to 6-core "Istanbul" processor (2H 2009 upgrade)
- Socket replacement of the 4-core 2.3 GHz "Barcelona" Opteron processors with 6-core 2.6 GHz "Istanbul" processors
- Upgrades XT5 from 1.375 PF to 2.3 PF
- Provides compute boost for applications and a platform for scaling applications to 225K cores
- Node peak performance goes from 73.6 GF to 124.8 GF, an increase of 70%
- 21-27% increase in memory bandwidth, maybe more if we can run in unganged mode

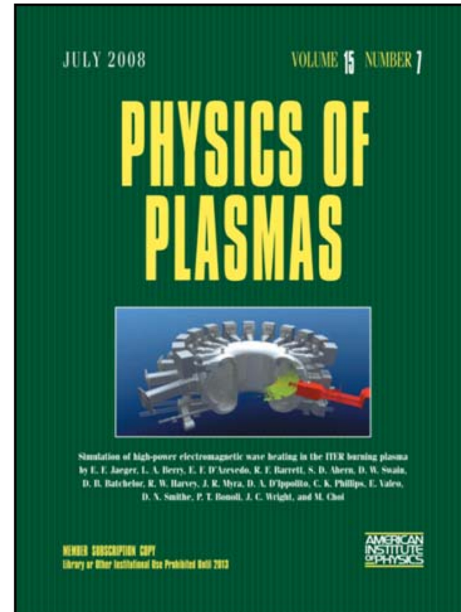


Highly visible science output

Physical Review Letters:
High temperature superconductivity

Combustion and Flame:
3D flame simulation

Nature:
Astrophysics

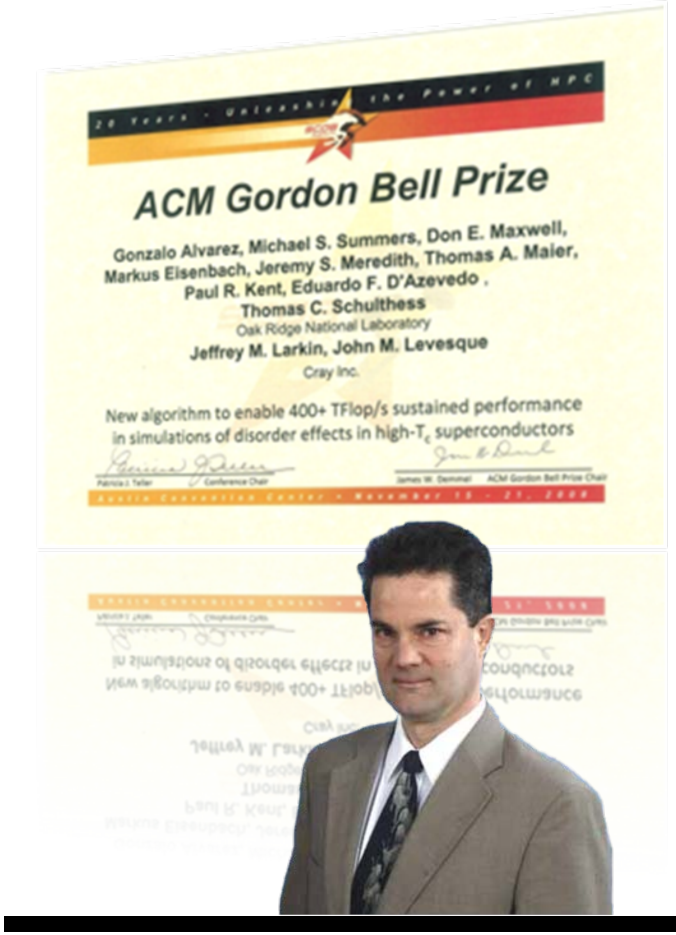


Physics of Plasmas:
ICRF heating in ITER

Transformational science enabled
by advanced scientific computing

2008 Gordon Bell Prize: 3 of 6 finalists ran on Jaguar

- Prize awarded to ORNL team led by Thomas Schulthess for DCA++ simulation of disorder effects in high-temperature superconductors
  - Simulation achieved 1.352 petaflops MP on ORNL's Cray XT Jaguar
  - Modifications of algorithms and software design boosted performance tenfold



Finalists

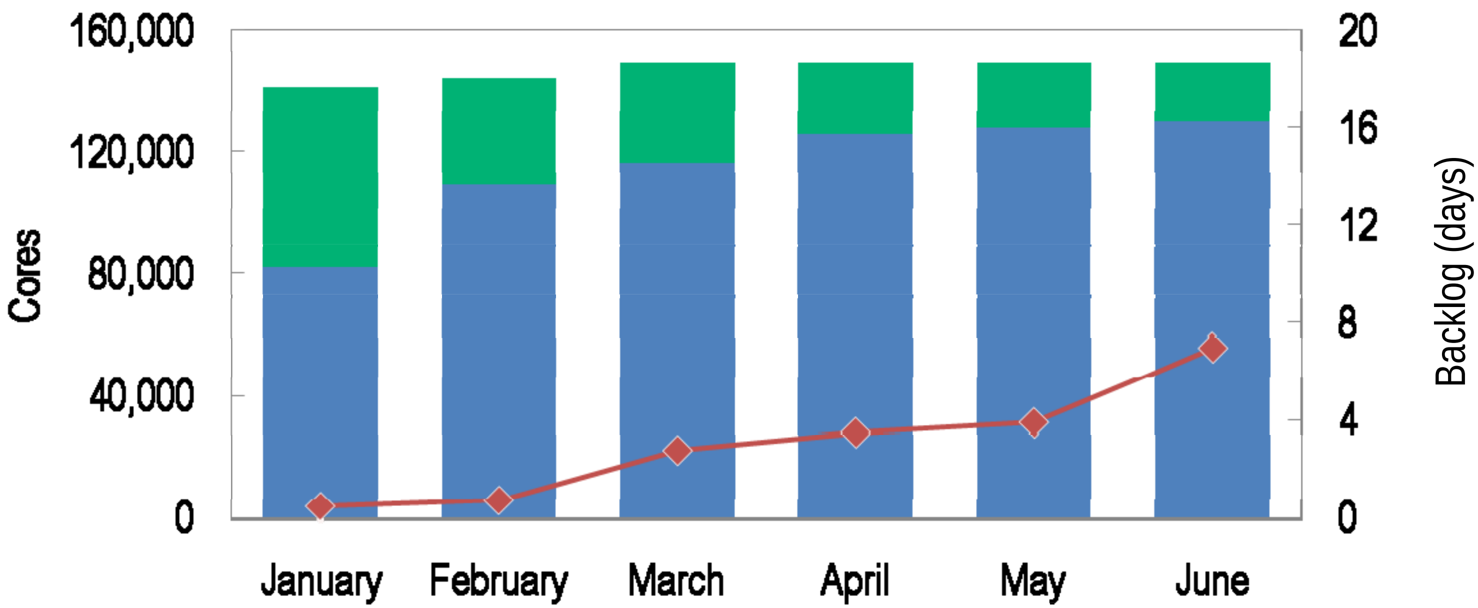
DCA++	ORNL	Jaguar
LS3DF	LBNL	Jaguar
SPECFEM3D	SDSC	Jaguar
RHEA	TACC	Ranger
SPaSM	LANL	Roadrunner
VPIC	LANL	Roadrunner

Science applications are scaling on Jaguar

Science area	Code	Contact	Cores	Performance
Materials	DCA++	Schulthess	150,144	1.3 PF MP
Materials	LSMS	Eisenbach	149,580	1.05 PF
Seismology	SPECFEM3D	Carrington	149,784	165 TF
Weather	WRF	Michalakes	150,000	50 TF
Climate	POP	Jones	18,000	20 simulation years/day
Combustion	S3D	Chen	144,000	83 TF
Fusion	GTC	PPPL	102,000	20 billion particles/second
Materials	LS3DF	Lin-Wang Wang	147,456	442 TF
Chemistry	NWChem	Apra	96,000	480 TF
Chemistry	MADNESS	Harrison	140,000	550+ TF

Jaguar XT5 status, April 2009

Stable	• MTTI: 32 hours • MTTF: 52 hours	• Driver for downtimes: Spider testing • System has been up as long as 10 days
Available	• Scheduled at 94% • Overall availability at 86%	• Driver for downtimes: Spider testing • Target completion date: Mid-July
Usable	• Fully loaded utilization: 84%	• Steadily climbing since start of transition to operations on January 12, 2009
In demand	• Current queue backlog: 3.9 days	• Delivered 73M hours in April 2009



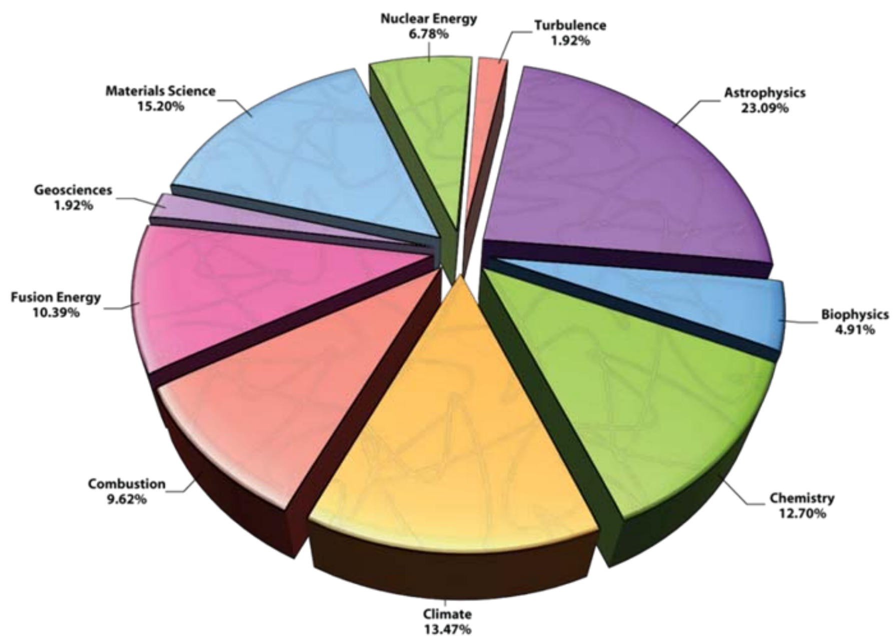
Jaguar XT5 science workload

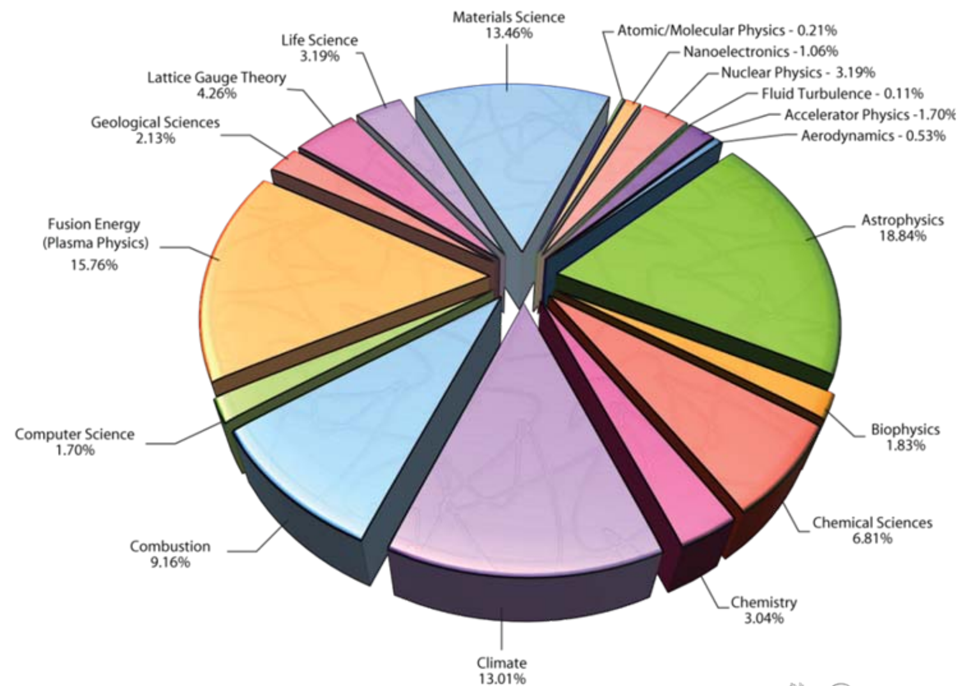
Current: Transition to operations

- 27 early-access projects, January–July
  - Total allocation: 540.8M hours
  - Usage since January: 247M hours (159 users)
- ALCC project: FY09 Joule metric applications
  - Total allocation: 25M hours
  - Usage since January: 17M hours (17 users)

Fall 2009: General availability

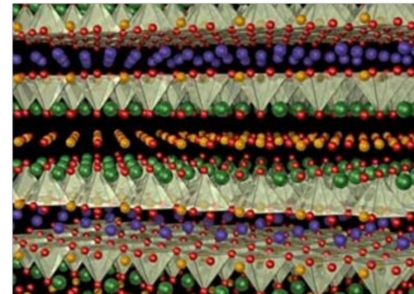
- 38 INCITE projects transitioned to system
  - Total allocation: 469M hours
  - Usage (XT4) since January: 91M hours (448 users)
- Discretionary projects: Climate AR5 production
  - Total allocation: 80M hours
  - 2 projects: NCAR/DOE and NOAA (~50 users)

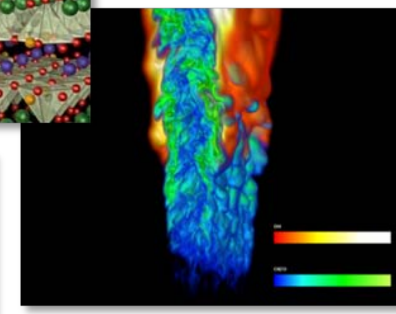


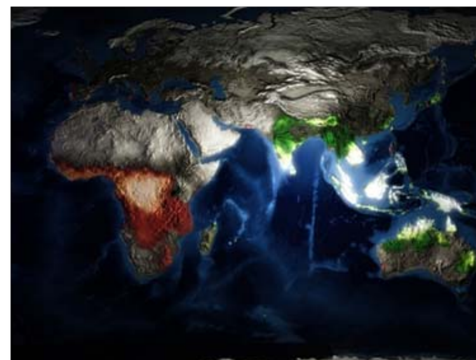


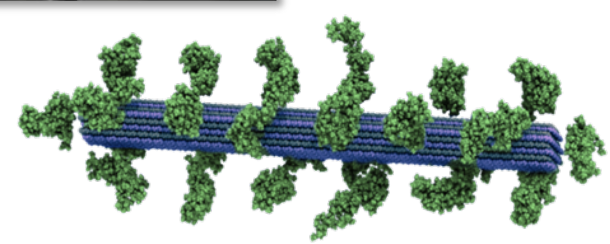
Jaguar XT5 early science projects: Tackling compelling problems

- Energy for environmental sustainability
  - Climate change: Carbon sequestration, weather event impacts on global climate, decadal climate predictive skill in aerosol forcing, global climate at unprecedented resolution, mantle convection
  - Energy storage: Charge storage and transfer in nano-structured supercapacitors
  - Combustion: Stabilizing diesel jet flames for increased efficiency and decreased emissions
  - Bioenergy: Recalcitrance in cellulosic ethanol
  - Solar: Nonequilibrium semiconductor alloys
  - Energy transmission: Role of inhomogeneities in high-temperature superconducting cuprates
  - Fusion: ITER design, optimization, and operation
  - Nuclear energy: Fully resolved reactor core neutron state



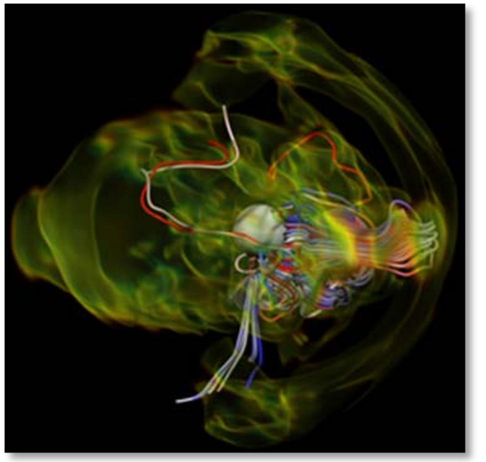


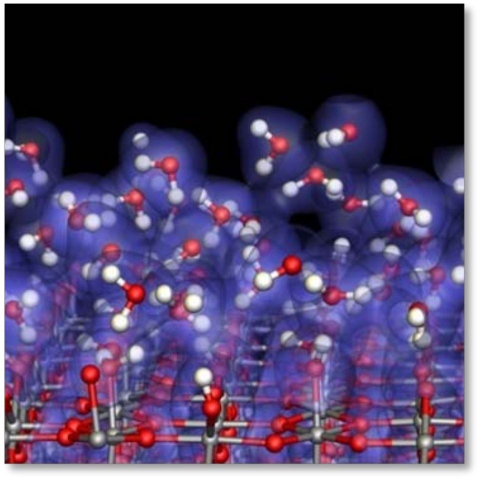




Jaguar XT5 early science projects: Tackling compelling problems

- Materials and nanoscience
  - Structure of nanowires, nanoelectronics, nanorods, and strongly correlated materials (magnets)
- Fundamental science
  - Astrophysics: Decipher core-collapse supernovae and black hole mergers
  - Chemistry: Water structure in biological and aqueous-phase systems; water bonding on graphene and carbon nanotubes
  - Nuclear physics: Probe the anomalously long lifetime of carbon-14
  - High energy physics: Generate lattices at the physical values of the light quark (u, d, s) masses
  - Turbulence: Dispersion relative to air quality modeling and bioterrorism; physics of shock/turbulence interaction
  - Biology: Effect of an embolization on hemodynamics in the entire neurovasculature





Reality: Facilities Designed for Exascale Computing

Open Science Center (40k ft^2)

- Upgrading building power to 15 MW
- 210 MW substation, upgradeable to 280 MW
- Deploying a 6,600 ton chiller plant
- Tripling UPS and generator capability

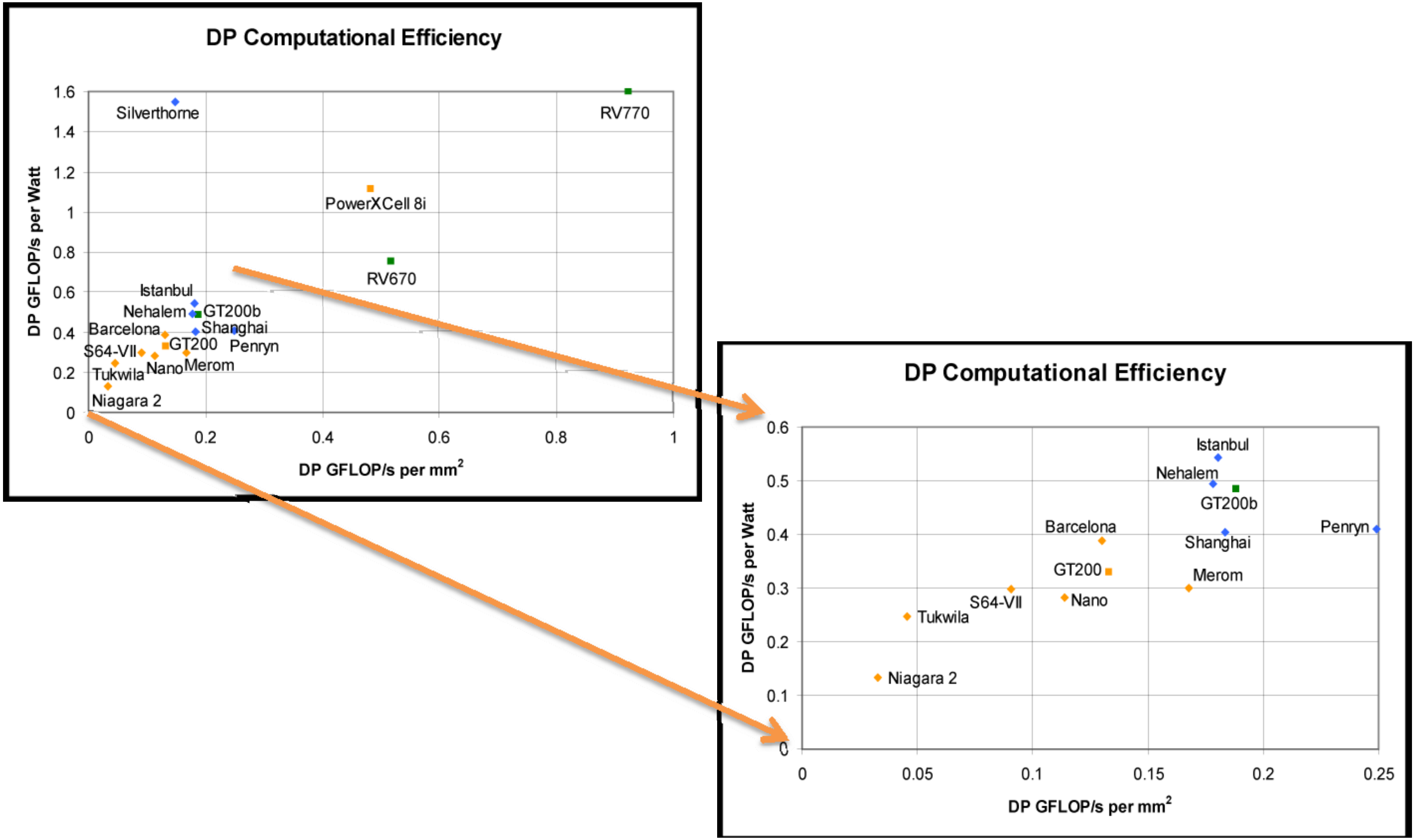
National Security Center (32k ft^2)

- Capability computing for national defence
- 25 MW of power and 8,000+ ton chillers

New Computer Facility (260k ft^2)

- 110 ft^2 raised floor classified; same unclassified
- Shared mechanical and electrical
- Lights out facility

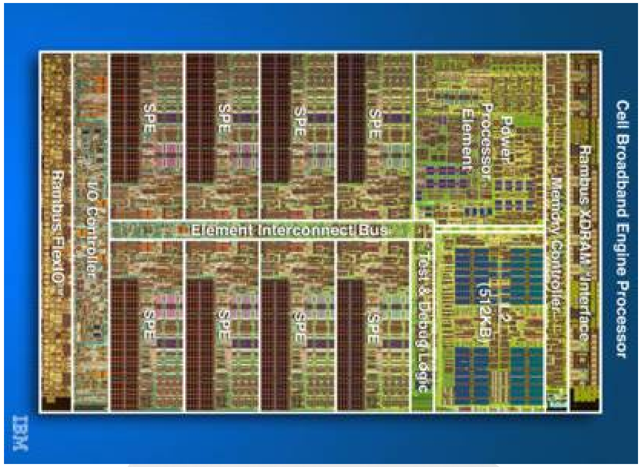
Power Efficiency! See Dr. Kogge's Talk

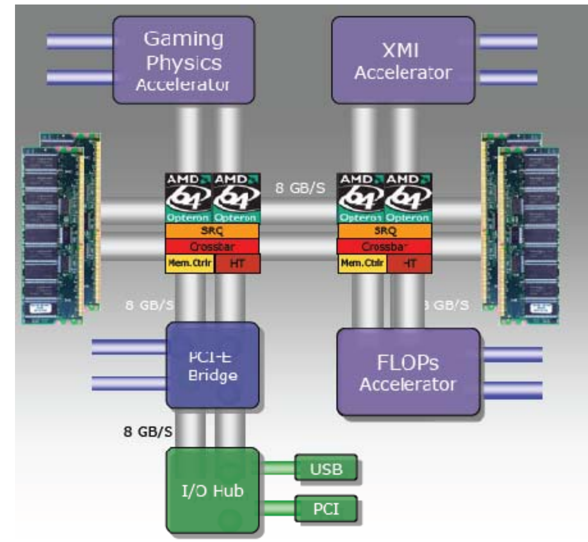


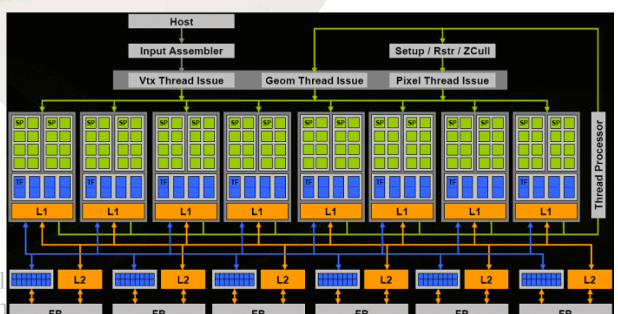
Source: http://www.realworldtech.com

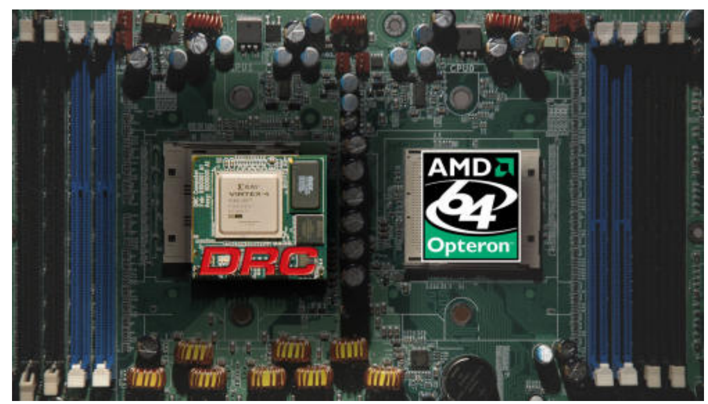
Architectural Complexity – Heterogeneity, Specialization, Multi-paradigm

Architectures target specific workloads: games, graphics, business, encryption, media





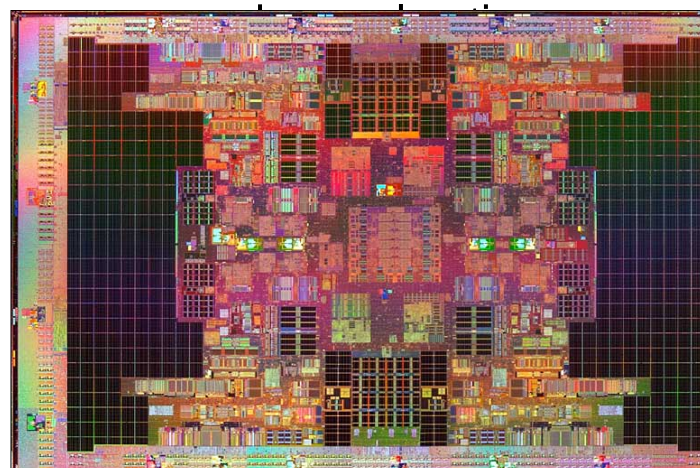




Trends will easily scale performance achieved to 1 TFlop on a node/socket.

Several Factors Motivate Specialization

- Sacrifice generality and compatibility to address specific algorithms
- Computational power
  - Asymmetric, dynamic cores can provide performance advantages [1]
  - Use transistors differently
- Tukwila – First 2 billion transistor chip
  - 80486 had ~1.2M transistors, ~50MHz, 1989
  - Specialization can be free
- Power
  - Combination of these features provide more ops per watt for your targeted application
- Memory
  - Include different cache or memory organization tuned for data access patterns
- Performance of general purpose



ISSCC 2008.

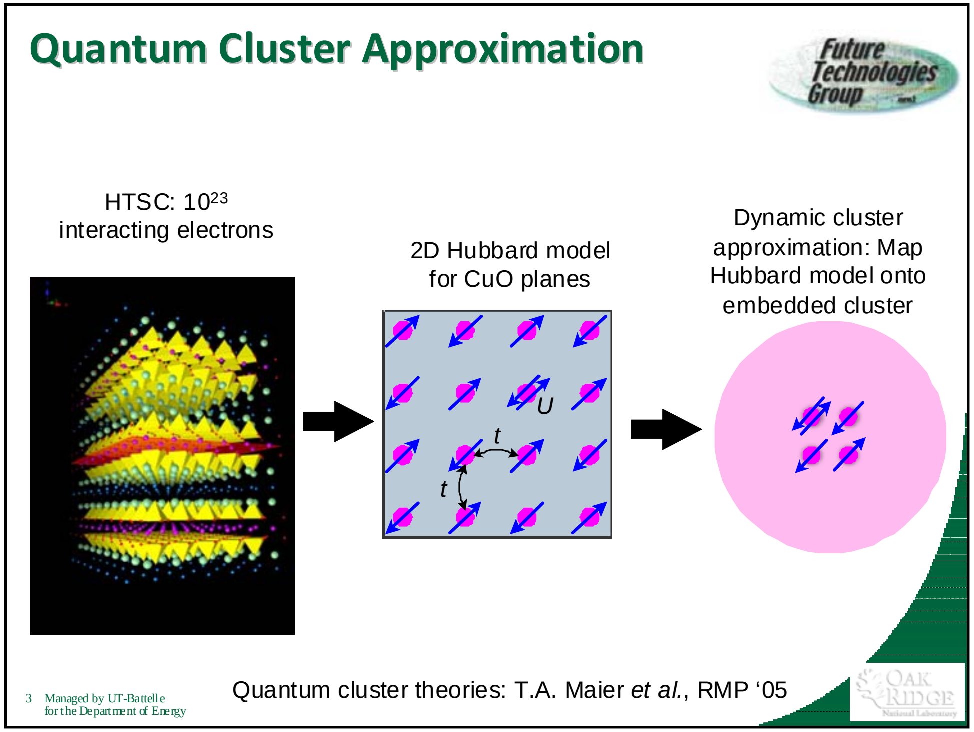
[1] M.D. Hill and M.R. Marty, “Amdahl's Law in the Multicore Era,” *IEEE Computer*, 2008.

Applications

- In depth application analysis of existing applications

		Usability and Availability			Motifs												Models				
Application	Domain	Scalability	Restrictions	Readiness	Structured Grid	Unstructured Grid	Multiresolution Grid	Dense Matrix	Sparse Matrix	Spectral/FFT	MD/N-Body	MapReduce / Monte Carlo	Graphs	PIC	Algo Load Imbalance	Adaptive	C/C++	FORTRAN	1-Sided Model Pref	MPI	OpenMP
MILC	QCD	3	3	3	3				3	3				3			3			3	
NEK5000	Nuclear Physics / CFD	3	3	3				3		3								3		3	
AMG	Multigrid solver	3	3	3					3											3	
IRS	implicit radiation solver	3	3	3	3			3												3	
LAMMPS	classical MD code	3	3	3						3	3						3			3	
Sphot	monte carlo photon transport	3	3	3								3								3	
UMT	radiation transport	3	3	3		3														3	
GTC	Fusion	3	3	3	2	2			3					3			1	3		3	3
MILC	QCD	3	3	3	3				3	3				3			3			3	
DCA++	Materials	3	2	3				3				3								3	
GTC	Fusion	3	2	3	2	2			3					3			1	3		3	3
POP	Climate	2	3	3	3				3					3				3		3	3
S3D	Combustion	3	2	3	3													3		3	
MILC	QCD	3	3	3	3				3	3				3			3			3	
GAMESS	Chemistry	1	3	3				3							1			3	3	2	
HYCOM	Ocean	1	3	3	3															3	
WRF	Weather	3	3	3		3	3									1	3	3		3	1
FLASH	Astrophysics	3	1	3												3				3	
Flash IO	IO Benchmark	3	1	3																3	
NAMD	Bio	1	3	3																3	
QBOX	Materials	3	1	3				3		3				3			3			3	
AMR Solver Kernel	Solver	2	3	3	3		3								3	3				3	
IOR	IO Benchmark	2	3	3													3			3	
PARATEC	Materials	2	3	3	3			3										3		3	
Flash	Astrophysics	3	2	3			3									3				3	
Flash IO	IO Benchmark	3	2	3			3													3	
LSMS	Materials	3	2	3	3			3									2	2		3	
WRF	Weather	3	3	3		3	3									1	3	3		3	

Computational Materials - Case Study



- Joint work with ORNL Computational Materials Group (T. Schulthess)

CPU: single 2.0GHz Opteron

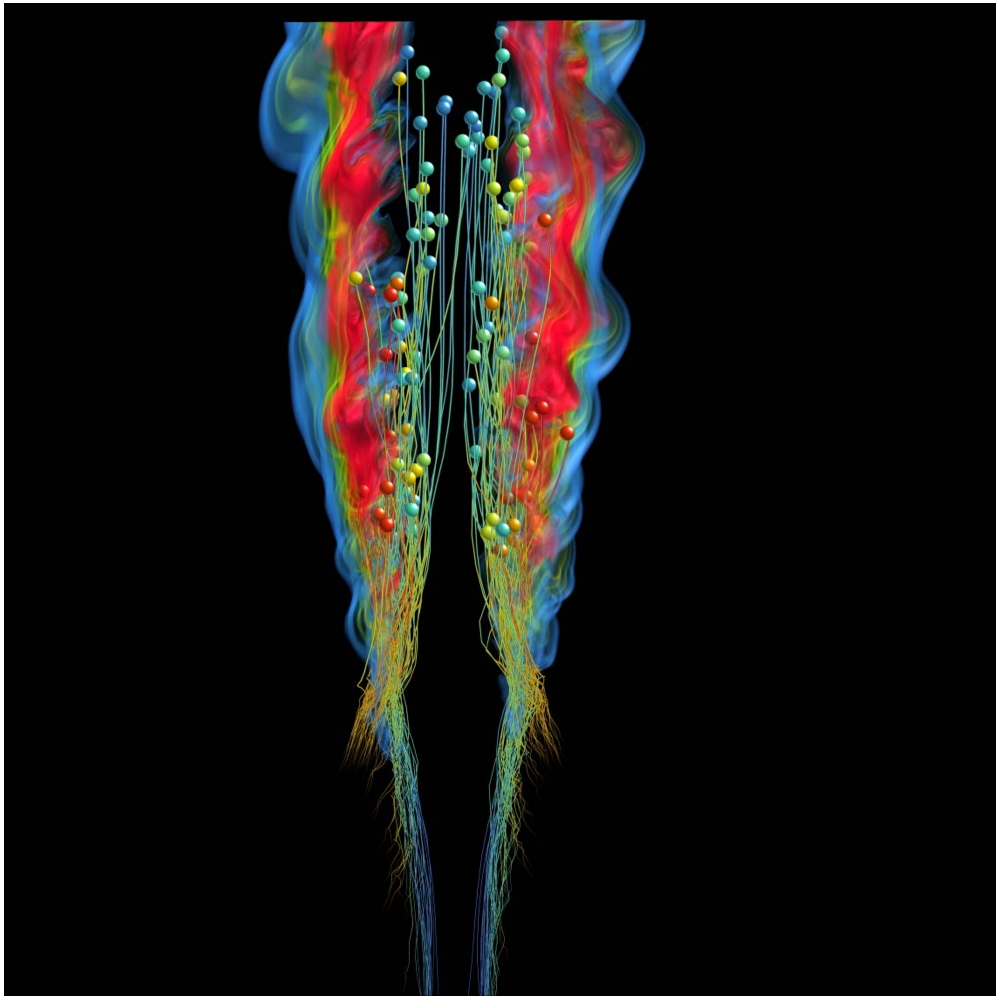
GPU: single 500MHz/96SP G80 (8800GTS 320MB)

- 10x speedup
  - Offload BLAS to GPU, transfer all data
- 13x speedup
  - Offload BLAS, smart data orchestration
- 19x speedup
  - Offload BLAS + additional functions, smart data orchestration

DCA++: J.S. Meredith, G. Alvarez *et al., "Accuracy and performance of graphics processors: A Quantum Monte Carlo application case study," Parallel Comput., 35(3):151-63, 2009.*

S3D – A case study

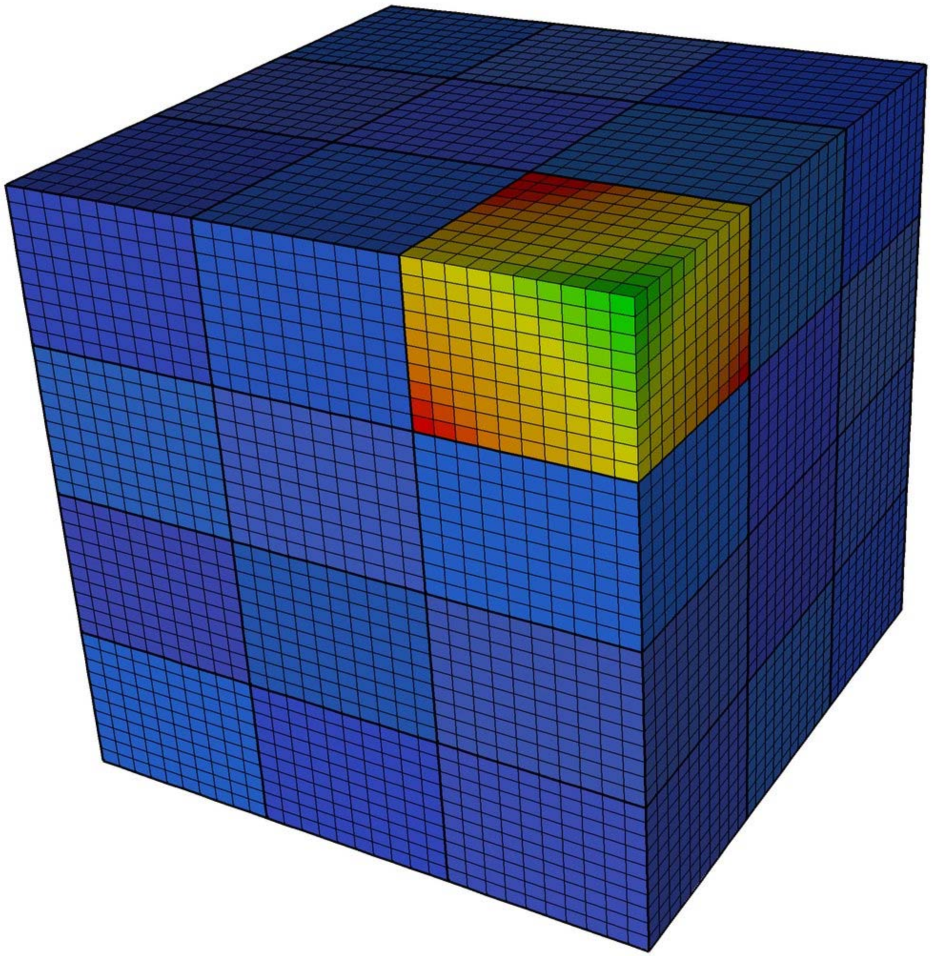
- S3D is a massively parallel direct numerical solver (DNS) for the full compressible Navier-Stokes, total energy, species and mass continuity equations coupled with detailed chemistry.
- Dept. of Energy marquee application for combustion
- Highly scalable code, written entirely in Fortran with MPI



K. Spafford, J. Meredith *et al.*, *“Accelerating S3D: A GPGPU Case Study,” in Seventh International Workshop on Algorithms, Models, and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar 2009). Delft, The Netherlands, 2009*

Challenges: Tailoring S3D to the GPU

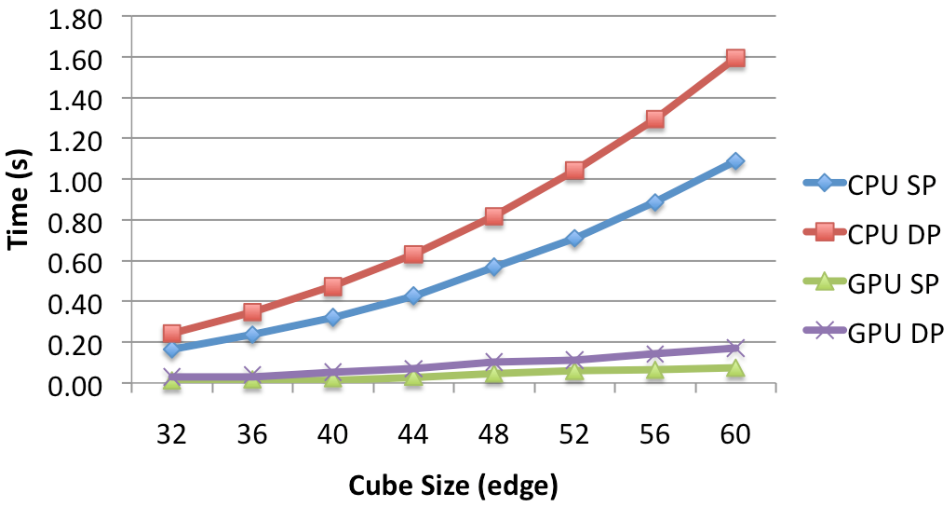
- Main Tasks
  - S3D has a huge code base written in Fortran, porting the entire application without proof-of-concept is infeasible—must choose a kernel for acceleration
  - Expose fine-grained data parallelism
  - Modify algorithms to exploit the GPUs memory hierarchy
- Chemistry calculations – GetRates kernel
  - Based on temperature, pressure, and mass fraction, GetRates computes the speed at which chemical reactions are occurring
  - Comprises roughly 45-60% of runtime, varying with grid size
    - Expected to increase with more complex chemistry
  - Operates on a regular, structured 3-D grid



Exploiting the GPU Memory Hierarchy

- Coalesce reads/writes to global memory
  - Single most important factor for performance
- Interleave accesses to global memory with intense calculation
- Use shared memory when possible
- Constant memory: almost as fast as register access on cache hit
- Transfer data to/from GPU only when absolutely necessary

- SP: 13-14.5x DP: 8.1x-9.3x



Keeneland – In a nutshell

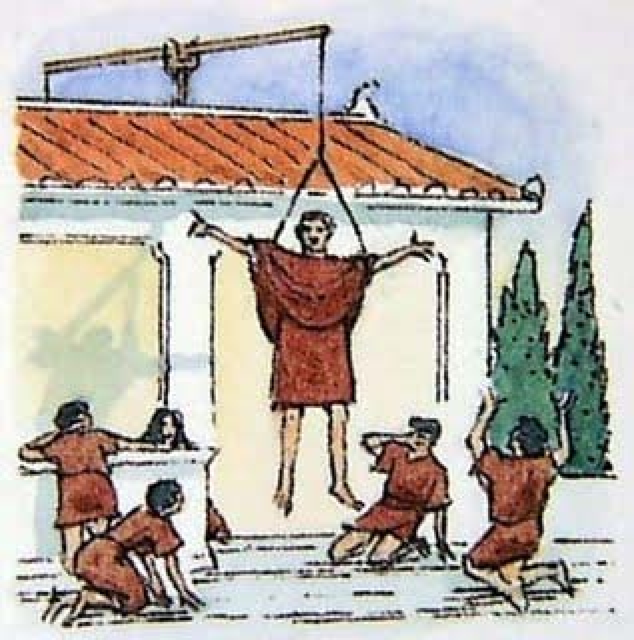
- NSF Track 2D/XHPC
- NSF award
  - $12M
  - 5 years from 2010
- Team
  - Georgia Tech, Lead
  - ORNL, UTK, NICS
- Field two GPGPU accelerated HPC systems
  - Initial Delivery, 2010
  - Final Delivery, 2012
  - FD decision point
  - Operations and support
  - Apps and software development
- Strong emphasis on outreach and education
  - Tutorials, Classes, Workshops

DOE/ORNL is considering this option

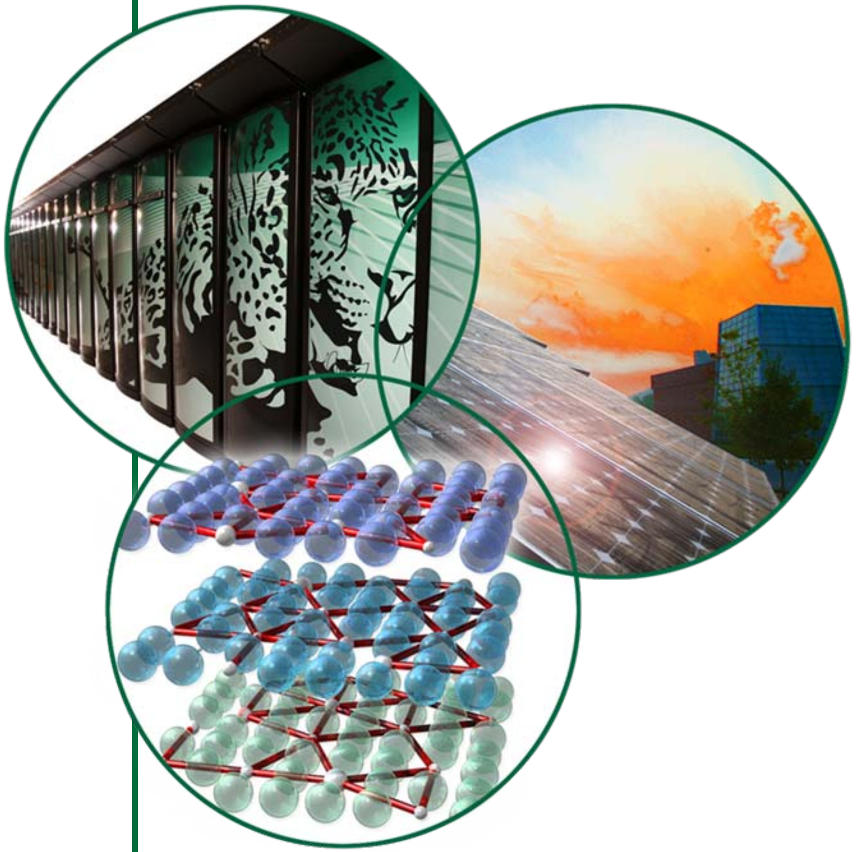
- 2011
- 10-20 PF
- Multicore Commodity processors paired with accelerators
- Applications
  - Tier 1
    - S3D
    - DCA++
    - CAM-HOMME
    - LAMMPS
    - MADNESS
  - Tier 2
    - Chimera
    - Denovo
    - AORSA
    - gWL-LSMS
    - GTC

Deus ex machina?

- Drama
  - Increasing appetite for scientific simulation
  - Increasing size of systems leading to huge facilities
- Enter accelerators/GPUs
- Will GPUs untangle this plot?

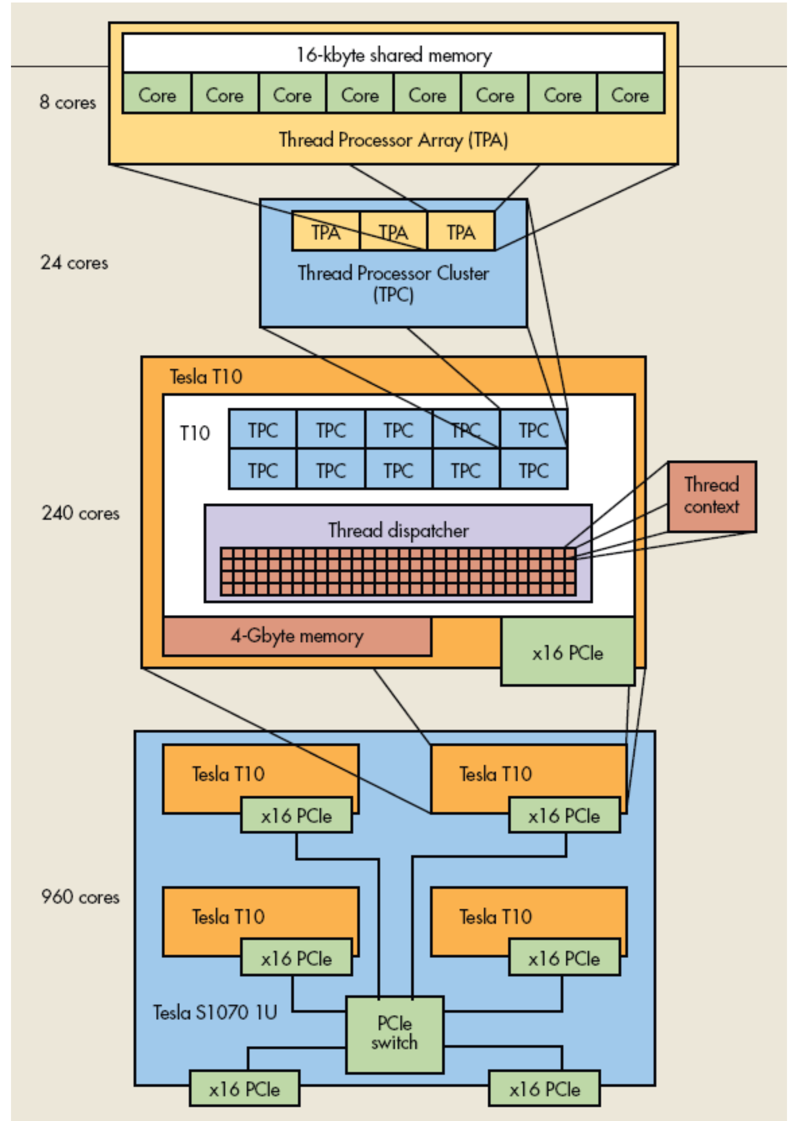


Bonus Slides



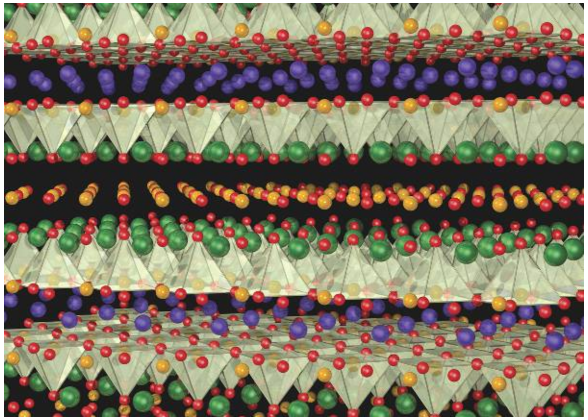
Concurrency - Observations

- S3D has many levels of concurrency
  - MPI
  - Host – GPU including data orchestration
  - SIMT
  - One GPU can have 1000s of thread contexts
  - Cores have internal concurrency

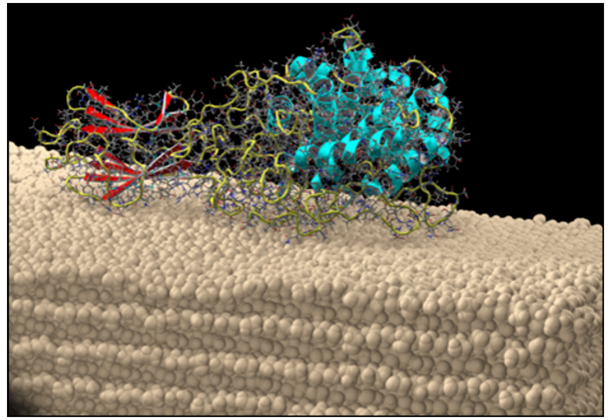


http://electronicdesign.com

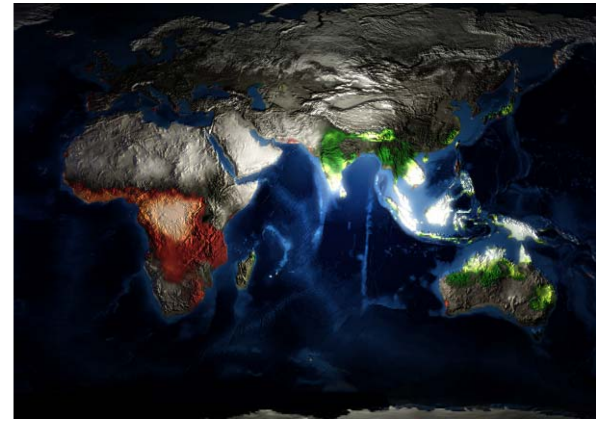
Leadership computing is advancing scientific discovery



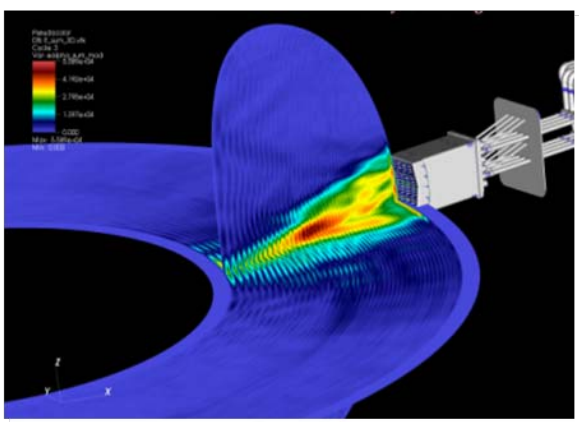
Resolved decades-long controversy about modeling physics of high temperature superconducting cuprates



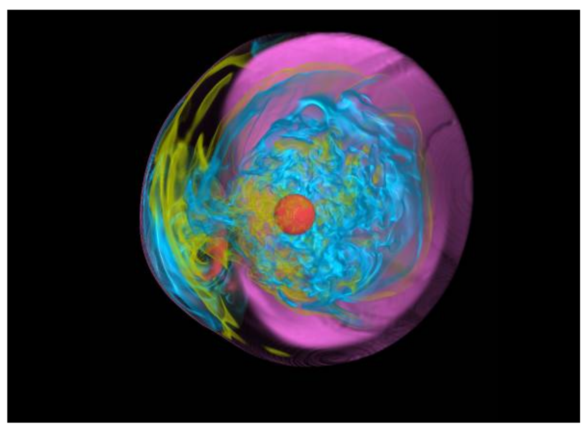
New insights into protein structure and function leading to better understanding of cellulose-to-ethanol conversion



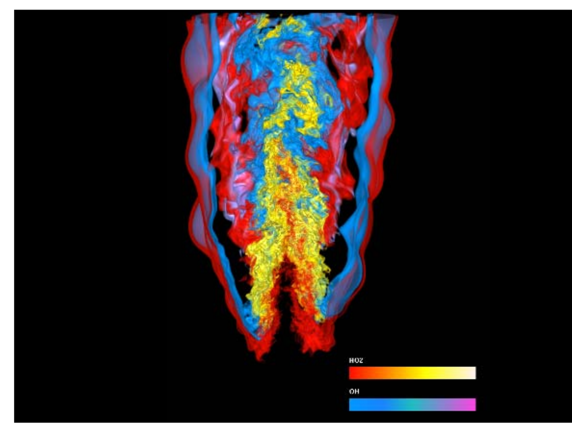
Addition of vegetation models in climate code for global, dynamic CO_2 exploration



First fully 3D plasma simulations shed new light on engineering superheated ionic gas in ITER



Fundamental instability of supernova shocks discovered directly through simulation



First 3-D simulation of flame that resolves chemical composition, temperature, and flow

ORNL FTG – Capabilities and Project Sampling

- Emerging architectures
  - Graphics Processors (e.g., NVIDIA)
  - FPGAs (e.g., Xilinx, Cray, SRC)
  - IBM CELL
  - Massive Multithreaded architectures (e.g., Cray XMT, Niagara 1 & 2)
  - Large Memory – Global addressing
- Performance
  - Scalable tools and infrastructures
    - mpiP MPI performance tool (over 10k downloads)
    - MR-Net scalable infrastructure
  - Prediction
    - Performance models up to 1M cores
    - Simulation of memory behavior
  - Ongoing Benchmarking
    - Nehalem, Shanghai, BlueGene/P
- Programming Systems
  - MPI3 One-sided and RMW operations
  - Global Distributed Transaction Memory support (e.g., Chapel global `atomic` ops)
- System Software
  - Parallel IO
    - Improved scalability for MPI-IO Lustre on Cray
    - PNFS over Lustre
  - TLB support for superpages
  - Virtualization for improved system management and scalability testing

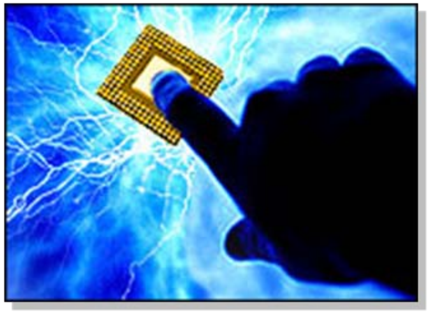
ORNL Future Technologies Group

http://ft.ornl.gov/

Mission - performs basic research in enabling technologies for future generations of high-end computing, including experimental computing architectures and software, with the goal of improving the performance, reliability, and productivity of these systems for our sponsors.

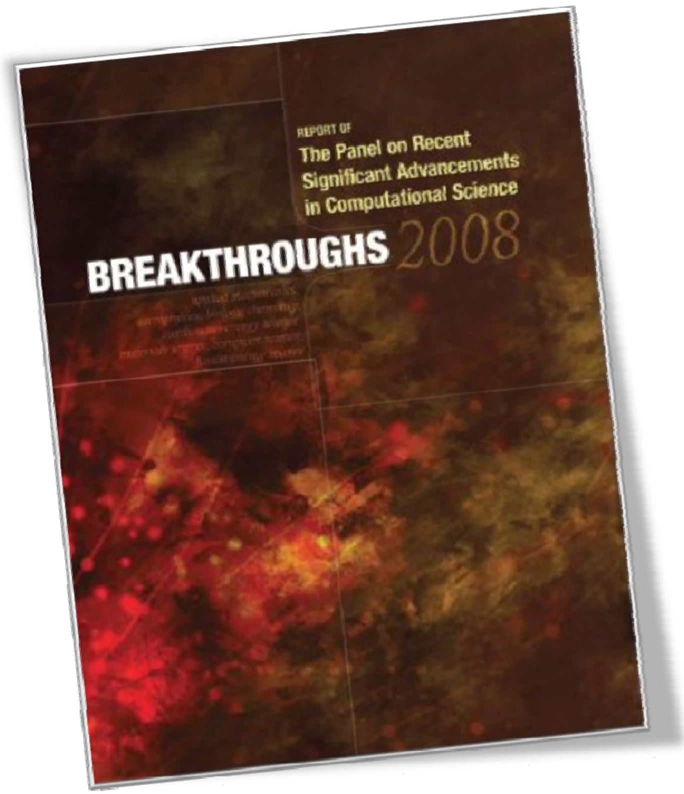
- Staff
  - Jeffrey Vetter (Group Leader)
  - Sadaf Alam
  - David Bader (joint w/ Georgia Institute of Technology)
  - Micah Beck (joint w/ University of Tennessee-Knoxville)
  - Anthony Danalis (joint w/ University of Tennessee-Knoxville)
  - Deb Holder (Group Secretary)
  - Gabriel Marin (in April)
  - Collin McCurdy
  - Jeremy Meredith
  - Kenneth Roche
  - Philip Roth
  - Thomas Sterling (joint w/ Louisiana State University)
  - Olaf Storaasli
  - Vinod Tipparaju
  - Weikuan Yu (joint w/ Auburn University)
  - Visitors

- Sponsors
  - DOE Office of Science
  - DARPA
  - Intelligence Community
  - LDRD
  - NCCS
  - NIH
  - SBIR
  - DOE Institute for Advanced Architectures and Algorithms



ORNL Leadership Computing Facility: Delivering important breakthroughs

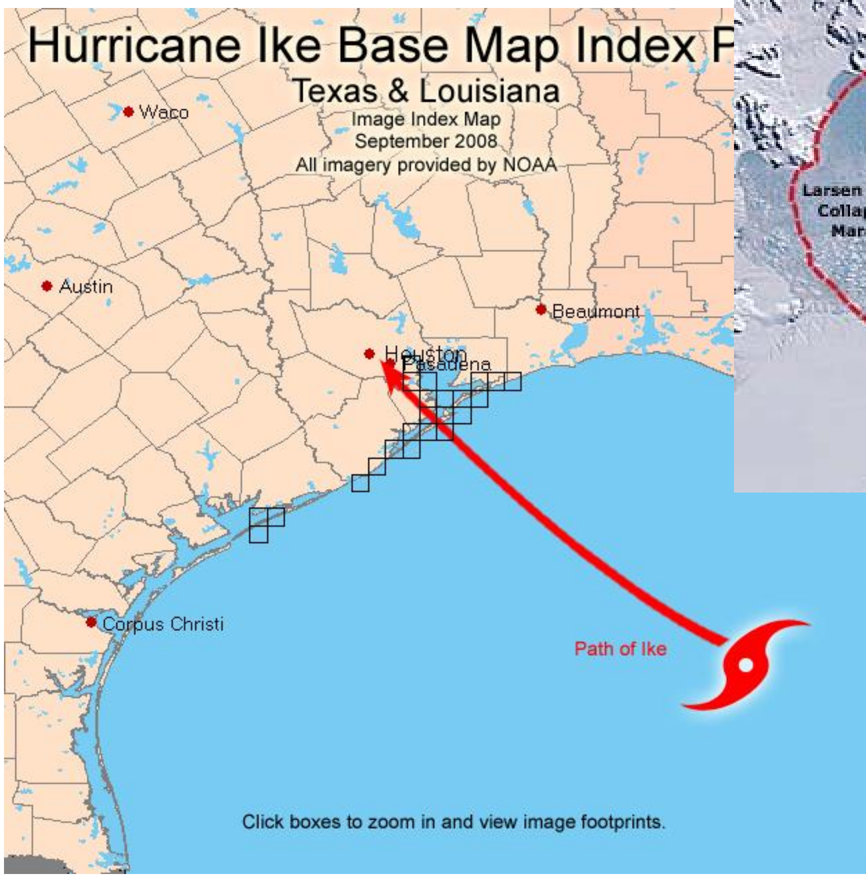
- February 2008: DOE's Office of Advanced Scientific Computing Research chartered a panel of computational scientists, applied mathematicians, and computer scientists to identify recent breakthroughs in computation science and enabling technologies
  - 10 major recent advances in simulation were identified
  - 5 of 10 were achieved using the Cray XT4 Jaguar at ORNL

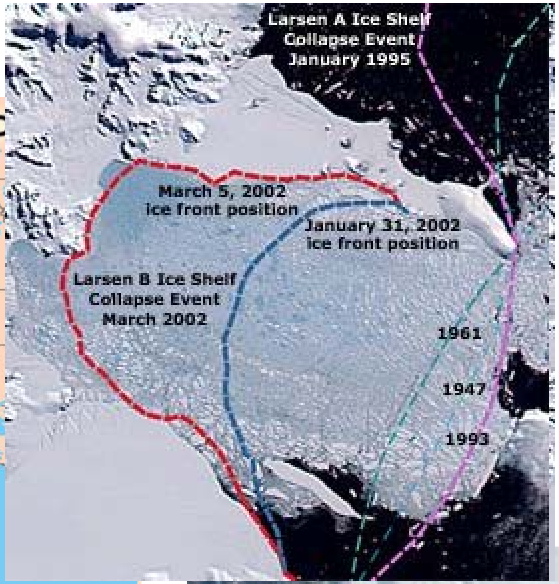


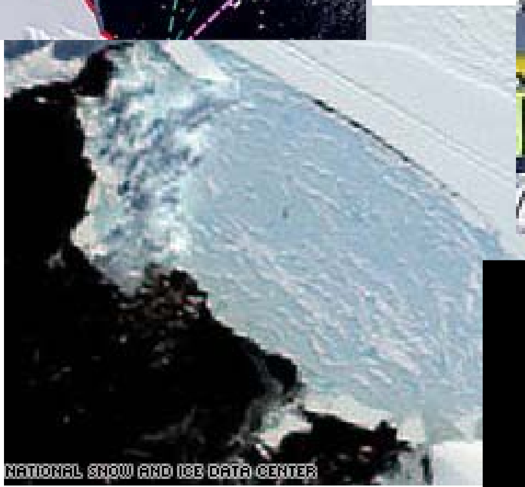
Jaguar XT5: Delivering leadership computing

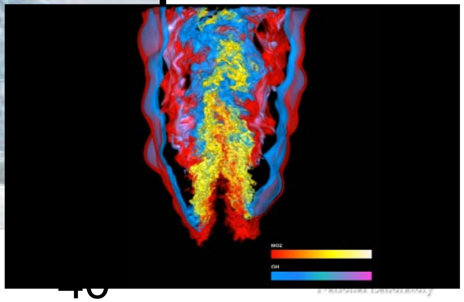
- Job size distribution indicates leadership usage
  - At least 30K cores: 50% of utilization
  - At least 45K cores: 32% of utilization
  - At least 90K cores: 18% of utilization
- 2009 Gordon Bell submissions: 5 from applications running on Jaguar XT5
- Application readiness and scaling demonstrate effectiveness of ORNL Leadership Computing Facility model

Changing Priorities of Workloads: Systems Must Perform Well for All









Application Complexity

- Today's applications are quickly becoming more complex than earlier applications
  - Exascale Town Hall report [1]
- Increased fidelity
- Answering more complex questions
  - Scientists [2,3] working on model that combines
    - Climate model
    - Energy economics
    - Population movements, demographics
- New application areas
  - Search
  - Massive Data
- Multiple phases of an multi-physics application
  - E.g., Climate
- Multiple languages, libraries, runtime systems

Applications design and implementation are already complex;
Writing and optimizing code for each new architecture and programming model is impractical (and only going to happen w/ heroic efforts/funding.)

[1] H. Simon, T. Zacharia, and R. Stevens, Eds., Modeling and Simulation at the Exascale for Energy and the Environment, 2007.
[2] S.W. Hadley, D.J. Erickson et al., "Responses of energy use to climate change: A climate modeling study," *Geophysical Research Letters*, 33(17), 2006.
[3] D. Vergano, "Air conditioning fuels warming," in USA Today, 2006.