Jaguar and Kraken -The World's Most Powerful Computer Systems

OAK RIDGE LEADERSHIP COMPUTING FACILITY

CRAY USER GROUP INCORPORATED

Arthur Bland

Cray Users' Group 2010 Meeting

Edinburgh, UK

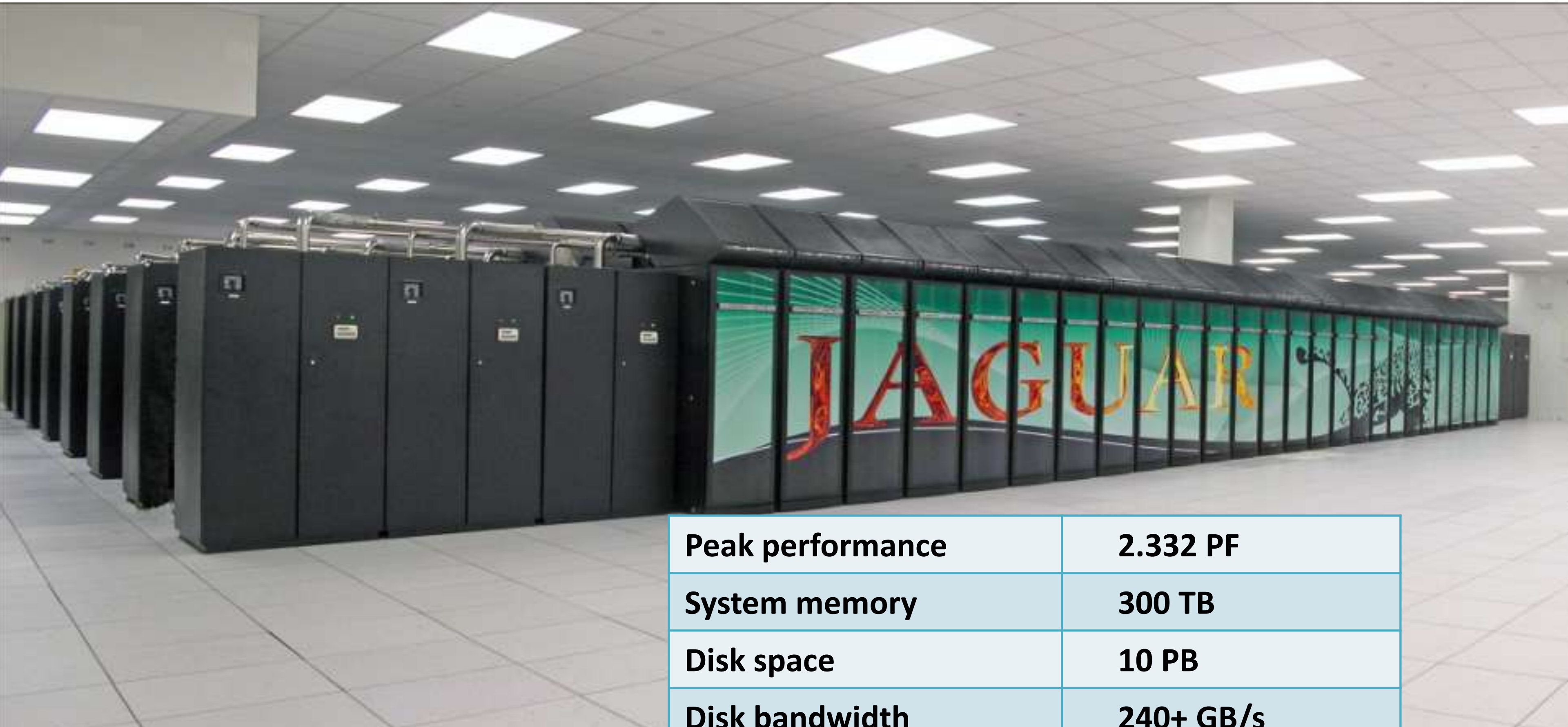
May 25, 2010

Abstract & Outline

At the SC'09 conference in November 2009, Jaguar and Kraken, both located at ORNL, were crowned as the world's fastest computers (#1 & #3) by the web site www.Top500.org. In this paper, we will describe the systems, present results from a number of benchmarks and applications, and talk about future computing in the Oak Ridge Leadership Computing Facility.

- Cray computer systems at ORNL
- System Architecture
- Awards and Results
- Science Results
- Exascale Roadmap

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



Peak performance	2.332 PF
System memory	300 TB
Disk space	10 PB
Disk bandwidth	240+ GB/s
Compute Nodes	18,688
AMD "Istanbul" Sockets	37,376
Size	4,600 $feet^2$
Cabinets	200 (8 rows of 25 cabinets)

Based on the Sandia & Cray designed Red Storm System

Kraken
World’s most powerful academic computer

Peak performance	1.03 petaflops
System memory	129 TB
Disk space	3.3 PB
Disk bandwidth	30 GB/s
Compute Nodes	8,256
AMD “Istanbul” Sockets	16,512
Size	2,100 $feet^2$
Cabinets	88 (4 rows of 22)

Climate Modeling Research System

Part of a research collaboration in climate science between ORNL and NOAA (National Oceanographic and Atmospheric Administration)

- Phased System Delivery
  - CMRS.1 (June 2010) 260 TF
  - CMRS.2 (June 2011) 720 TF
  - CMRS.1UPG (Feb 2012) 386 TF
  - Aggregate in June 2011: 980 TF
  - Aggregate in Feb 2012: 1106 TF
- Total System Memory
  - 248 TB DDR3-1333
- File Systems
  - 4.6 PB of disk (formatted)
  - External Lustre

Athena and Jaguar – Cray XT4

Athena

Peak Performance	166 TF
System Memory	18 TB
Disk Space	100 TB
Disk Bandwidth	10 GB/s
Compute Nodes	4,512
AMD 4-core Sockets	4,512
Size	800 $feet^2$
Cabinets	48

Jaguar

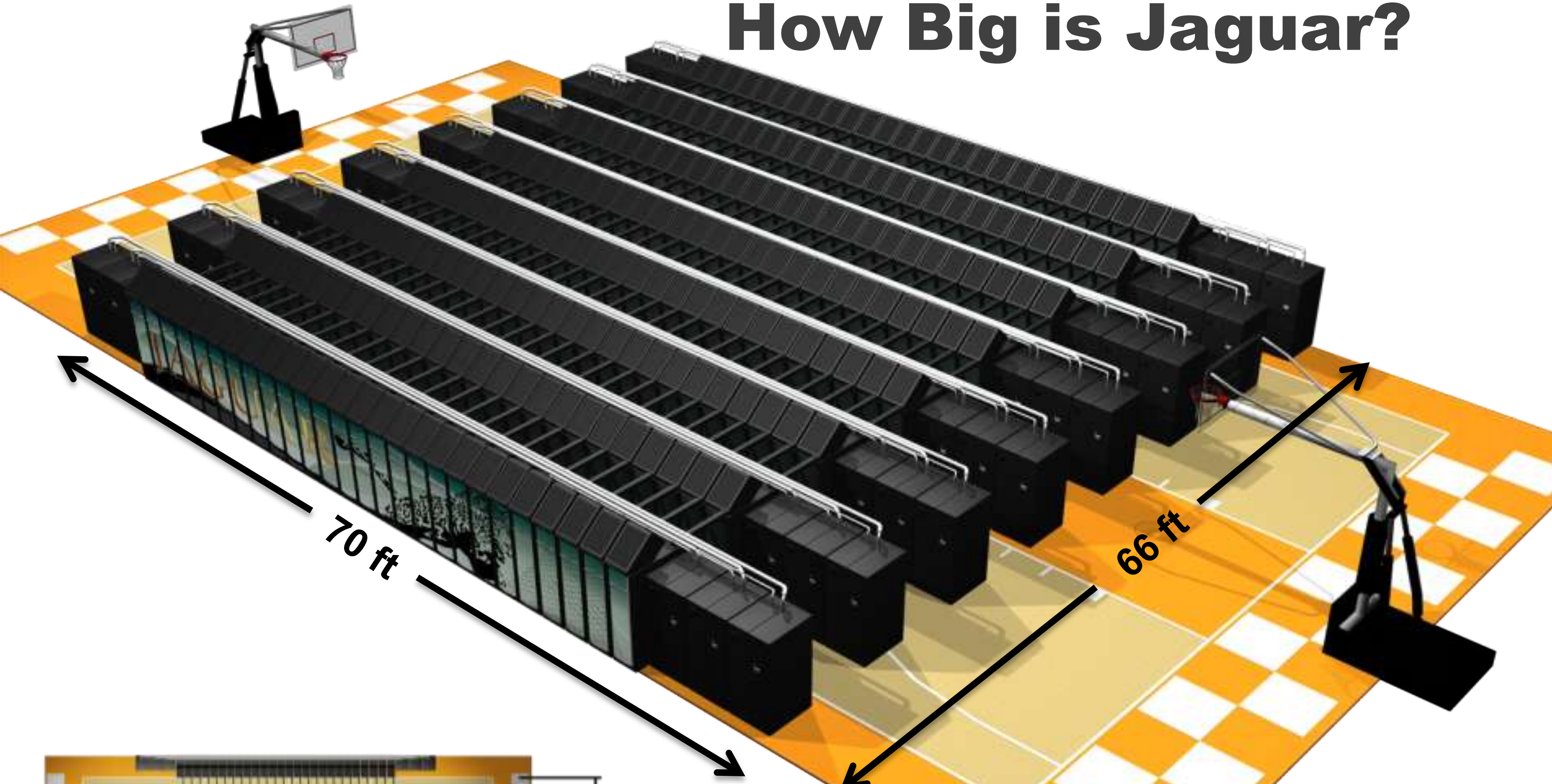
Peak Performance	263 TF
System Memory	62 TB
Disk Space	900 TB + 10 PB
Disk Bandwidth	44 GB/s
Compute Nodes	7,832
AMD 4-core Sockets	7,832
Size	1,400 $feet^2$
Cabinets	84

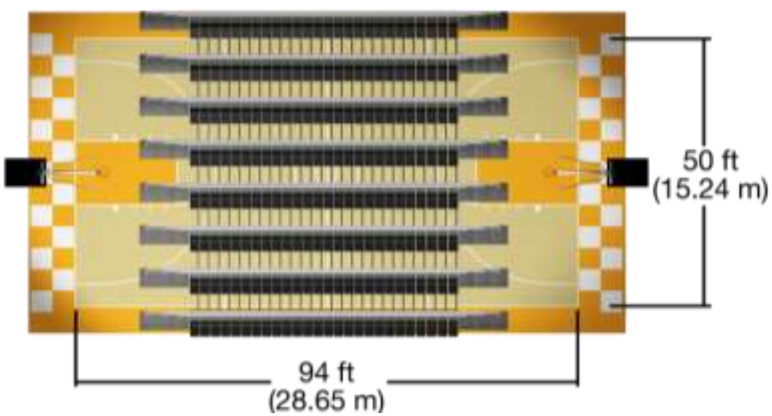
Cray XT Systems at ORNL

Characteristic	Jaguar XT5	Kraken XT5	Jaguar XT4	Athena XT4	NOAA "Baker"*	Total @ ORNL
Peak performance (TF)	2,332	1,030	263	166	1,106	**4,897 TF**
System memory (TB)	300	129	62	18	248	**757 TB**
Disk space (PB)	10	3.3	0.9	0.1	4.6	**18.9 PB**
Disk bandwidth (GB/s)	240	30	44	10	104	**428**
Compute Nodes	18,688	8,256	7,832	4,512	3,760	**43,048**
AMD Opteron Sockets	37,376	16,512	7,832	4,512	7,520	**73,752**
Size ($feet^2$)	4,600	2,100	1,400	800	1,000	**25,000**
Cabinets	200	88	84	48	40	**460**

*coming soon

How Big is Jaguar?





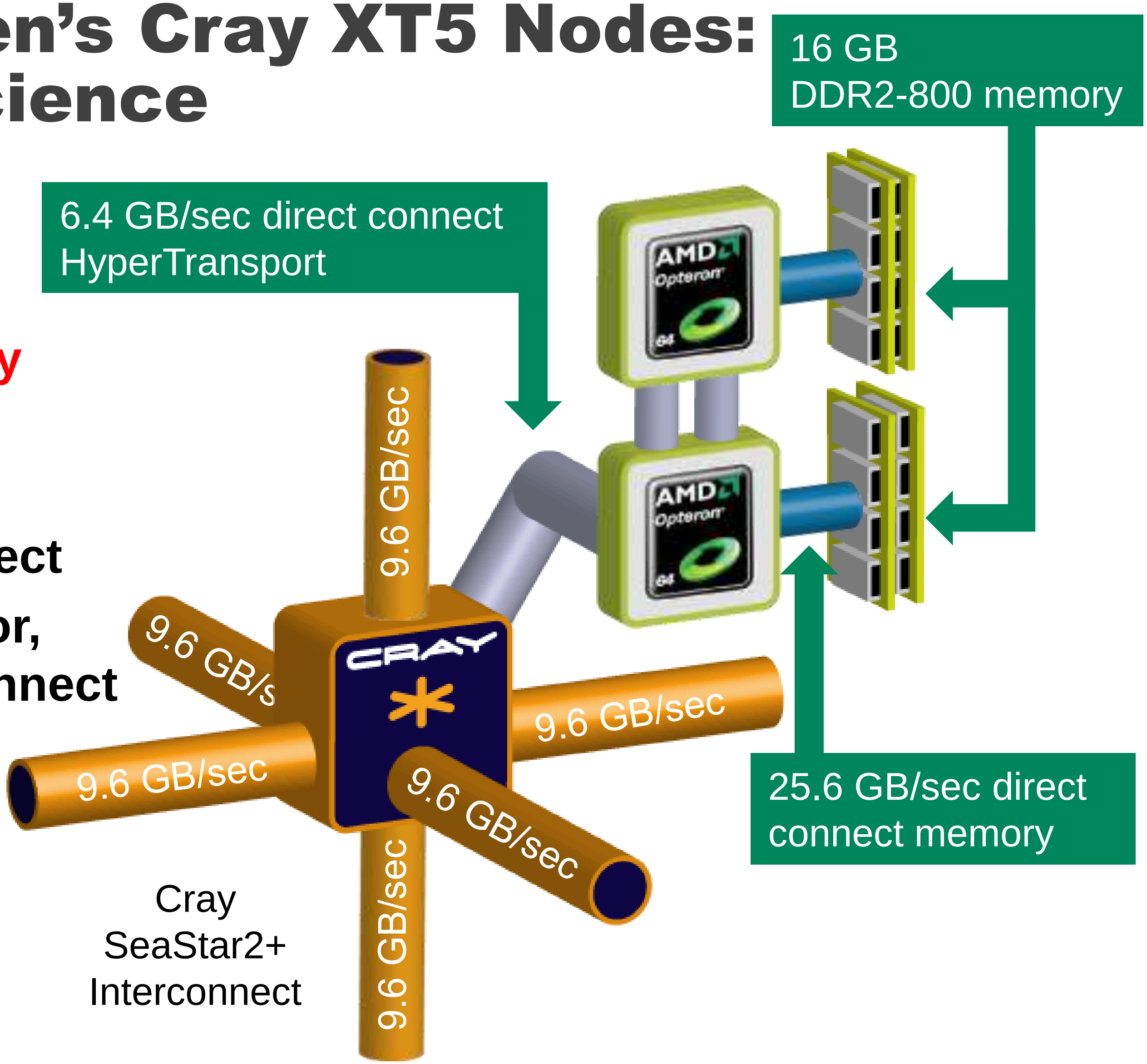
- 4,600 feet^2
- 7.6 megawatts (peak) 5.2 MW (avg.)
- 2,300 tons of Air Conditioning

Jaguar and Kraken were upgraded to AMD’s Istanbul 6-Core Processors

- Both Cray XT5 systems were upgraded from 2.3 GHz quad-core processors to 2.6 GHz 6-core processors.
- Increased Jaguar’s peak performance to 2.3 Petaflops and Kraken to 1.03 PF
- Upgrades were done in steps, keeping part of the systems available
- Benefits:
  - Increased allocatable hours by 50%
  - Increased memory bandwidth by 20%
  - Decreased memory errors by 33%
  - Increased performance by 69%

Jaguar & Kraken's Cray XT5 Nodes: Designed for science

- **Powerful node improves scalability**
- **Large shared memory**
- **OpenMP Support**
- **Low latency, High bandwidth interconnect**
- **Upgradable processor, memory, and interconnect**



GFLOPS	125
Memory (GB)	16
Cores	12
SeaStar2+	1

Center-wide File System

See Spider talk on Wednesday

This technology forms the basis of Cray's External I/O offering, "esFS".

- "Spider" provides a shared, parallel file system for all systems
  - Based on Lustre file system
- Demonstrated bandwidth of over 240 GB/s
- Over 10 PB of RAID-6 Capacity
  - 13,440 1-TB SATA Drives
- 192 Storage servers
- Available from all systems via our high-performance scalable I/O network (Infiniband)
- Currently mounted on over 26,000 client nodes
- ORNL and partners developed, hardened, and scaled key router technology

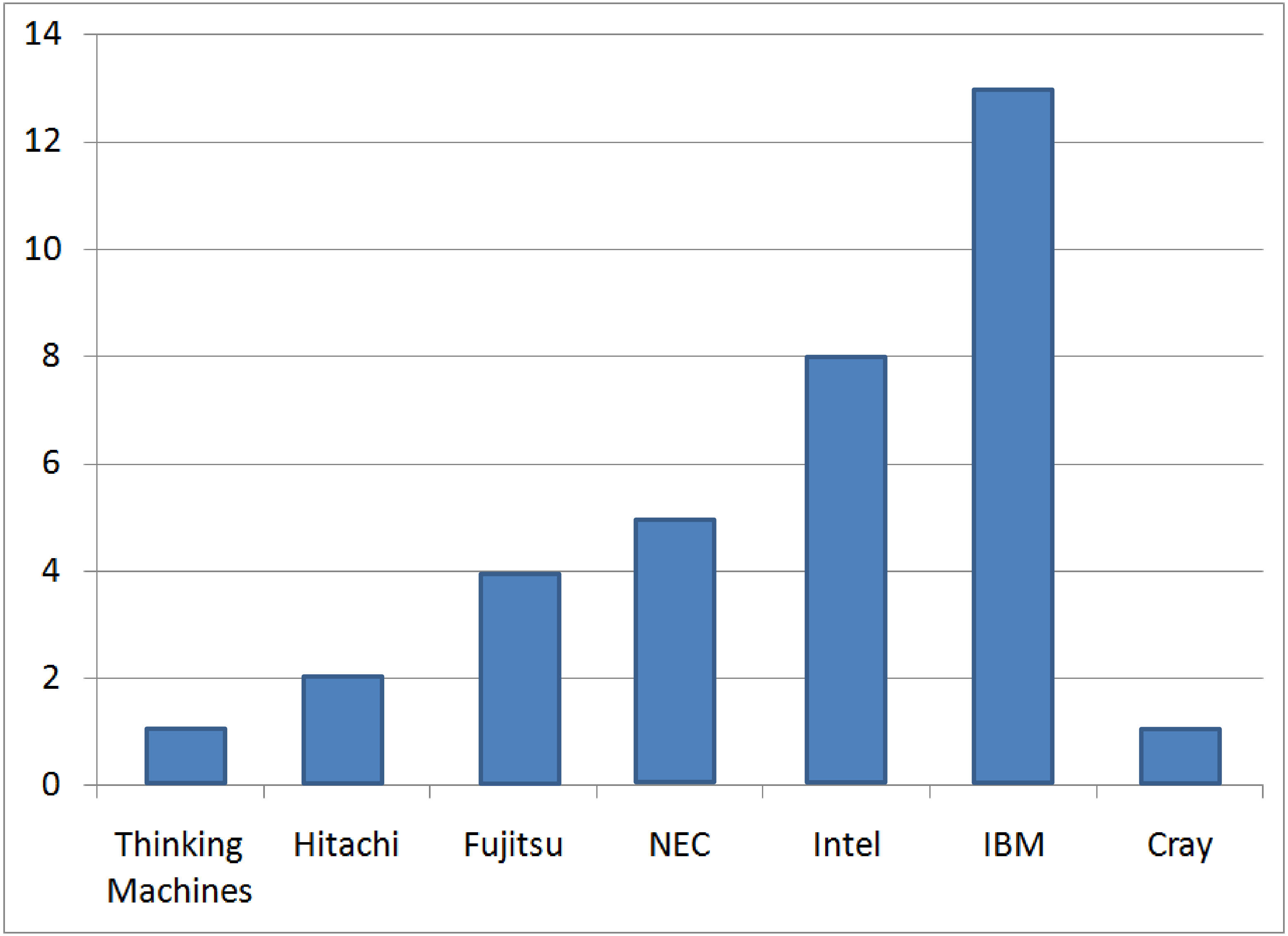
HPC Challenge Benchmarks

- **Tests many aspects of the computer's performance and balance**
- HPC Challenge awards are given out annually at the Supercomputing conference
- Awards in four categories, result published for two others
- Must submit results for all benchmarks to be considered
- **Jaguar** won 3 of 4 awards and placed 3rd in fourth
- **Jaguar** had the highest performance on the other benchmarks
- **Kraken** placed 2nd on three applications

G-HPL (TF)		EP-Stream (GB/s)		G-FFT (TF)		G-Random Access (GUPS)		EP-DGEMM (TF)		PTRANS (GB/s)	
ORNL	**1533**	ORNL	**398**	ORNL	**11**	LLNL	117	ORNL	**2147**	ORNL	**13,723**
NICS	**736**	LLNL	267	NICS	**8**	ANL	103	NICS	**951**	SNL	4,994
LLNL	368	JAMSTEC	173	JAMSTEC	7	ORNL	**38**	LLNL	363	LLNL	4,666

http://icl.cs.utk.edu/hpcc/

How many times has Cray been #1 on the Top500 List?



There have been 34 Top500 lists, starting in June 1993

HPLinpack Results

November 2009: http://www.top500.org/lists/2009/11

Jaguar PF

- 1.759 PetaFLOPS
- Over 17 hours to run
- 224,162 cores
- Rank: #1

Kraken

- 831.7 TeraFLOPS
- Over 11 hours to run
- 98,920 cores
- Rank: #3

But… Isn't it interesting that HPL is our 3rd fastest application!

Science Area	Code	Contact	Cores	Total Performance	Notes
Materials	DCA++	Schulthess	213,120	1.9 PF*	2008 Gordon Bell Winner
Materials	WL-LSMS	Eisenbach	223,232	1.8 PF	**2009 Gordon Bell Winner**
Chemistry	NWChem	Apra	224,196	1.4 PF	2009 Gordon Bell Finalist
Nano Materials	OMEN	Klimeck	222,720	860 TF	
Seismology	SPECFEM3D	Carrington	149,784	165 TF	2008 Gordon Bell Finalist
Weather	WRF	Michalakes	150,000	50 TF	
Combustion	S3D	Chen	144,000	83 TF	
Fusion	GTC	PPPL	102,000	20 billion Particles / sec	
Materials	LS3DF	Lin-Wang Wang	147,456	442 TF	2008 Gordon Bell Winner
Chemistry	MADNESS	Harrison	140,000	550+ TF	

2009 Gordon Bell Prize Winner and Finalist

Winner: Peak Performance Award

See talk on Thursday

A Scalable Method for Ab Initio Computation of Free Energies in Nanoscale Systems

- Markus Eisenbach (ORNL)
- Thomas C. Schulthess (ETH Zürich)
- Donald M. Nicholson (ORNL)
- Chenggang Zhou (J.P. Morgan Chase & Co)
- Gregory Brown (Florida State University)
- Jeff Larkin (Cray Inc.)

Finalist: Peak Performance Award

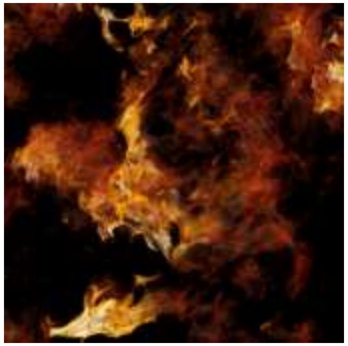
See talk on Thursday

Liquid Water: Obtaining the Right Answer for the Right Reasons

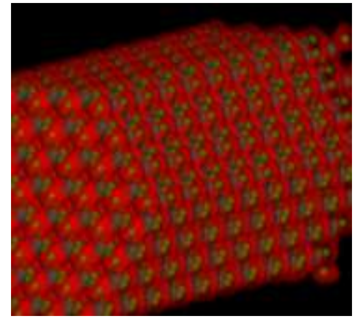
- Edoardo Apra (ORNL)
- Robert J. Harrison (ORNL)
- Vinod Tipparaju (ORNL)
- Wibe A. de Jong (PNNL)
- Sotiris Xantheas (PNNL)
- Alistair Rendell (Australian National University)

Great scientific progress at the petascale

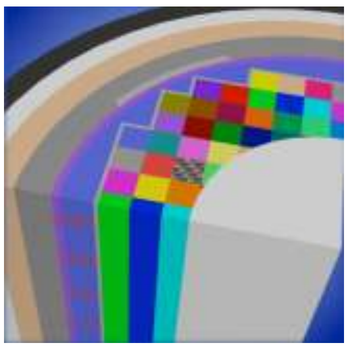
Jaguar is making a difference in energy research



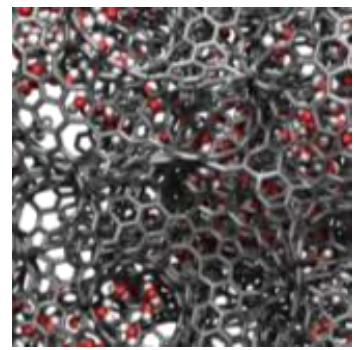
Turbulence
Understanding the statistical geometry of turbulent dispersion of pollutants in the environment



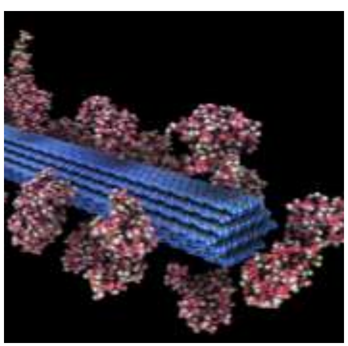
Nano Science
Understanding the atomic and electronic properties of nanostructures in next-generation photovoltaic solar cell materials



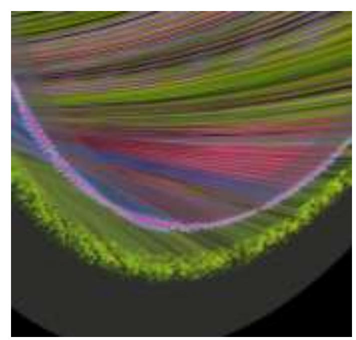
Nuclear Energy
High-fidelity predictive simulation tools for the design of next-generation nuclear reactors to safely increase operating margins



Energy Storage
Understanding the storage and flow of energy in next-generation nanostructured carbon tube supercapacitors



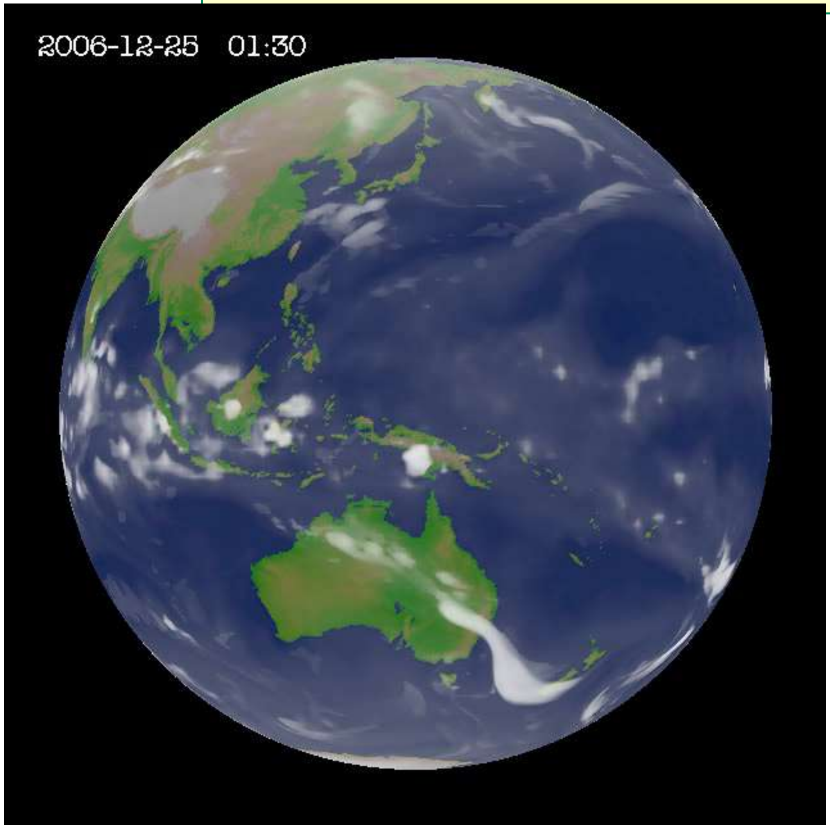
Biofuels
A comprehensive simulation model of lignocellulosic biomass to understand the bottleneck to sustainable and economical ethanol production



Fusion Energy
Understanding anomalous electron energy loss in the National Spherical Torus Experiment

An International, Dedicated High-End Computing Project to Revolutionize Climate Modeling

Runs completed in April 2010. Simulation generated over 1 PB of data



http://ftp.ccsr.u-tokyo.ac.jp/~satoh/nicam/MJO2006/olr_gl11_061225-061231.mpg

Project

Use dedicated HPC resources – Cray XT4 (Athena) at NICS – to simulate global climate change at the highest resolution ever. **Six months of dedicated access.**

Collaborators

COLA	Center for Ocean-Land-Atmosphere Studies, USA
ECMWF	European Center for Medium-Range Weather Forecasts
JAMSTEC	Japan Agency for Marine-Earth Science and Technology
UT	University of Tokyo
NICS	National Institute for Computational Sciences, University of Tennessee

Codes

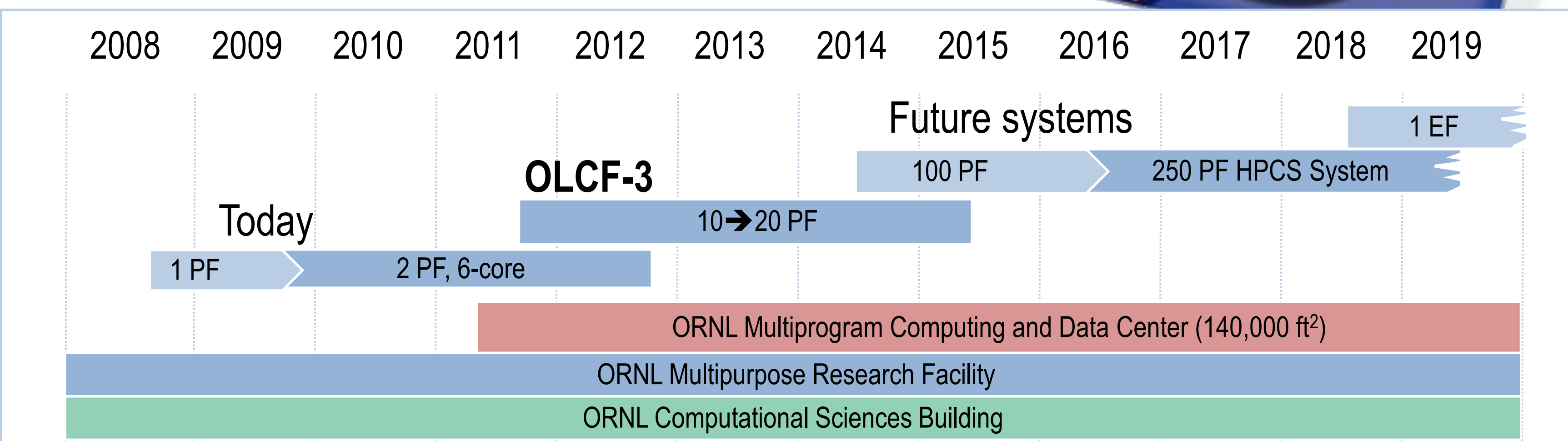
NICAM	Nonhydrostatic, Icosahedral, Atmospheric Model
IFS	ECMWF Integrated Forecast System

Expected Outcomes

- Better understand global mesoscale phenomena in the atmosphere and ocean
- Understand the impact of greenhouse gases on the regional aspects of climate
- Improve the fidelity of models simulating mean climate and extreme events

Moving to the Exascale

- The U.S. Department of Energy requires exaflops computing by 2018 to meet the needs of the science communities that depend on leadership computing
- Our vision: Provide a series of increasingly powerful computer systems and work with user community to scale applications to each of the new computer systems
  - **Today**: Upgrade of Jaguar to 6-core processors in progress
  - **OLCF-3 Project**: New 10-20 petaflops computer based on early DARPA HPCS technology

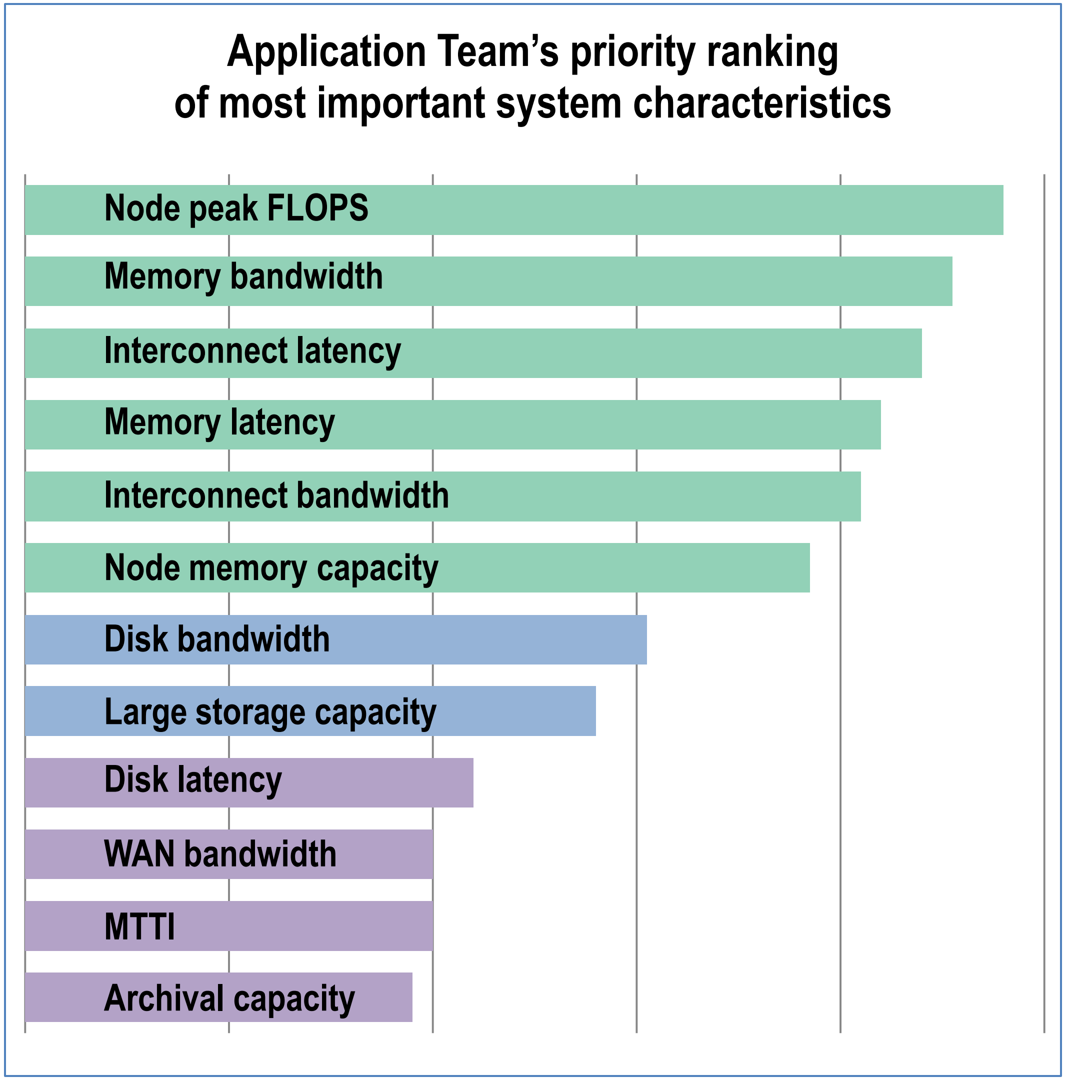


OLCF Roadmap from 10-year plan

What do the Science Codes Need?

What system features do the applications need to deliver the science?

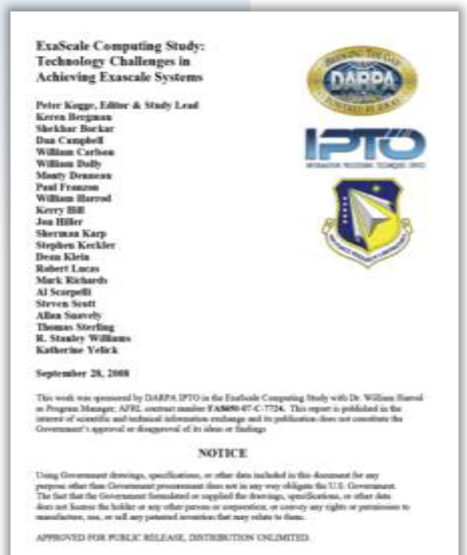
- 10-20 PF in 2011–2012 time frame with 1 EF by end of the decade
- Applications want powerful nodes, not lots of weak nodes
  - Lots of FLOPS and OPS
  - Fast, low-latency memory
  - Memory capacity ≥ 2GB/core
- Strong interconnect



How will we deliver these features, and address the power problem?

DARPA ExaScale Computing Study (Kogge et al.): We can't get to the exascale without radical changes

- Clock rates have reached a plateau and even gone down
- Power and thermal constraints restrict socket performance
- Multi-core sockets are driving up required parallelism and scalability



Future systems will get performance by integrating accelerators on the socket (already happening with GPUs)

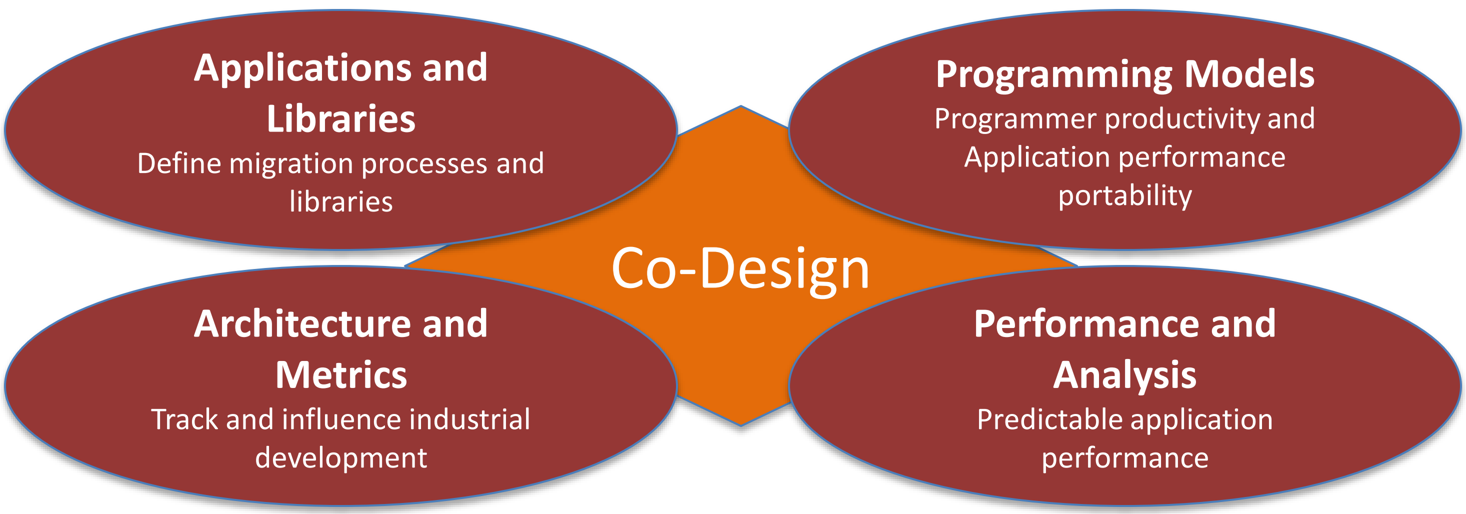
- AMD Fusion™
- Intel Larrabee
- NVIDIA Tesla
- IBM Cell (power + synergistic processing units)
- This has happened before (3090+array processor, 8086+8087, …)

Hybrid Multi-core Consortium (HMC)

http://computing.ornl.gov/HMC/

A multi-organizational partnership to support the effective development (productivity) and execution (performance) of high-end scientific codes on large-scale, accelerator based systems.

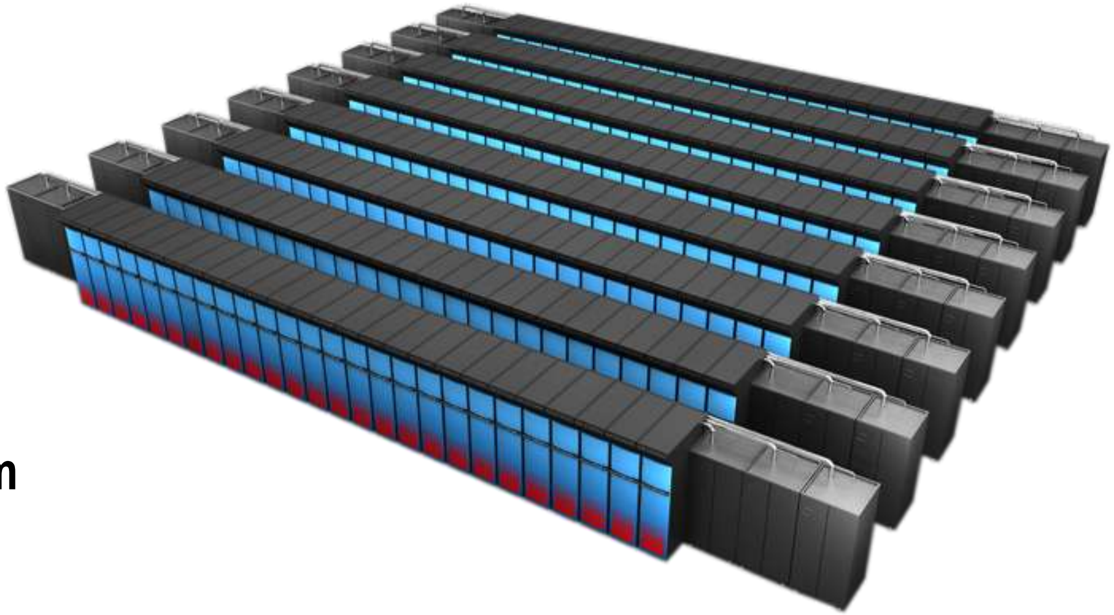
Goal: Facilitate production readiness of hybrid multi-core systems by identifying strategies and processes, based on **co-design** among applications, programming models, and architectures.



Membership is open to all parties with an interest in large-scale systems based on hybrid multi-core technologies

Titan (OLCF-3) system specification

- Similar number of cabinets, cabinet design, and cooling as Jaguar
- Operating system based on Linux
- High-speed, Low latency interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- Heterogeneous node design
- 10-20 PF peak performance
- Much larger memory
- 3x larger and 4x faster file system



Why do we build these large systems?

Movie at http://computing.ornl.gov/SC09/videos/SUPERCOMPOPEN_1Mb.mov

ORNL and NICS Talks at CUG 2010

Monday	Tuesday	Wednesday	Thursday
Guru Kora, (ORNL), RAVEN: RAS Data Analysis Through Visually Enhanced Navigation	**R. Glenn Brook, (NICS),** Interactions Between Application Communication and I/O Traffic on the Cray XT High Speed Network	**David Dillow, (ORNL),** Lessons Learned in Deploying the World's Largest Scale Lustre File System	**Rainer Keller, (ORNL),** MPI Queue Characteristics of Large-scale Applications
Galen Shipman, (ORNL), Correlating Log Messages for System Diagnostics	**Galen Shipman, (ORNL),** Performance Monitoring Tools for Large Scale Systems		**Markus Eisenbach, (ORNL),** Thermodynamics of Magnetic Systems from First Principles: WL-LSMS
Mark Fahey, (NICS), Automatic Library Tracking Database	**Richard Graham, (ORNL),** Effects of Shared Memory and Topology in the Cray XT5 Environment		**Kenneth Matney, Sr., (ORNL),** Parallelism in System Tools
Robert Whitten, Jr., (ORNL), A Pedagogical Approach to User Assistance	**Matthew Ezell, (NICS),** Collecting Application-Level Job Completion Statistics		**Galen Shipman, (ORNL),** Reducing Application Runtime Variability on XT5
James Rosinski, (ORNL), General Purpose Timing Library: A Tool for Characterizing Performance of Parallel & Serial Apps	**Troy Baer, (NICS),** Using Quality of Service for Scheduling on Cray XT Systems		**Patrick Worley, (ORNL),** XGC1: Performance on the 8-core and 12-core Cray XT5 Systems at ORNL
	Arthur Bland, (ORNL), Jaguar and Kraken, The World's Most Powerful Computer Systems		**Bronson Messer, (ORNL),** Evolution of a Petascale Application: Work on CHIMERA
			Edoardo Apra, (ORNL), What's a 200,000 CPU Petaflops Computer Good For?
			Mike McCarty, (NICS), Regression Testing on Petaflops Computational Resources