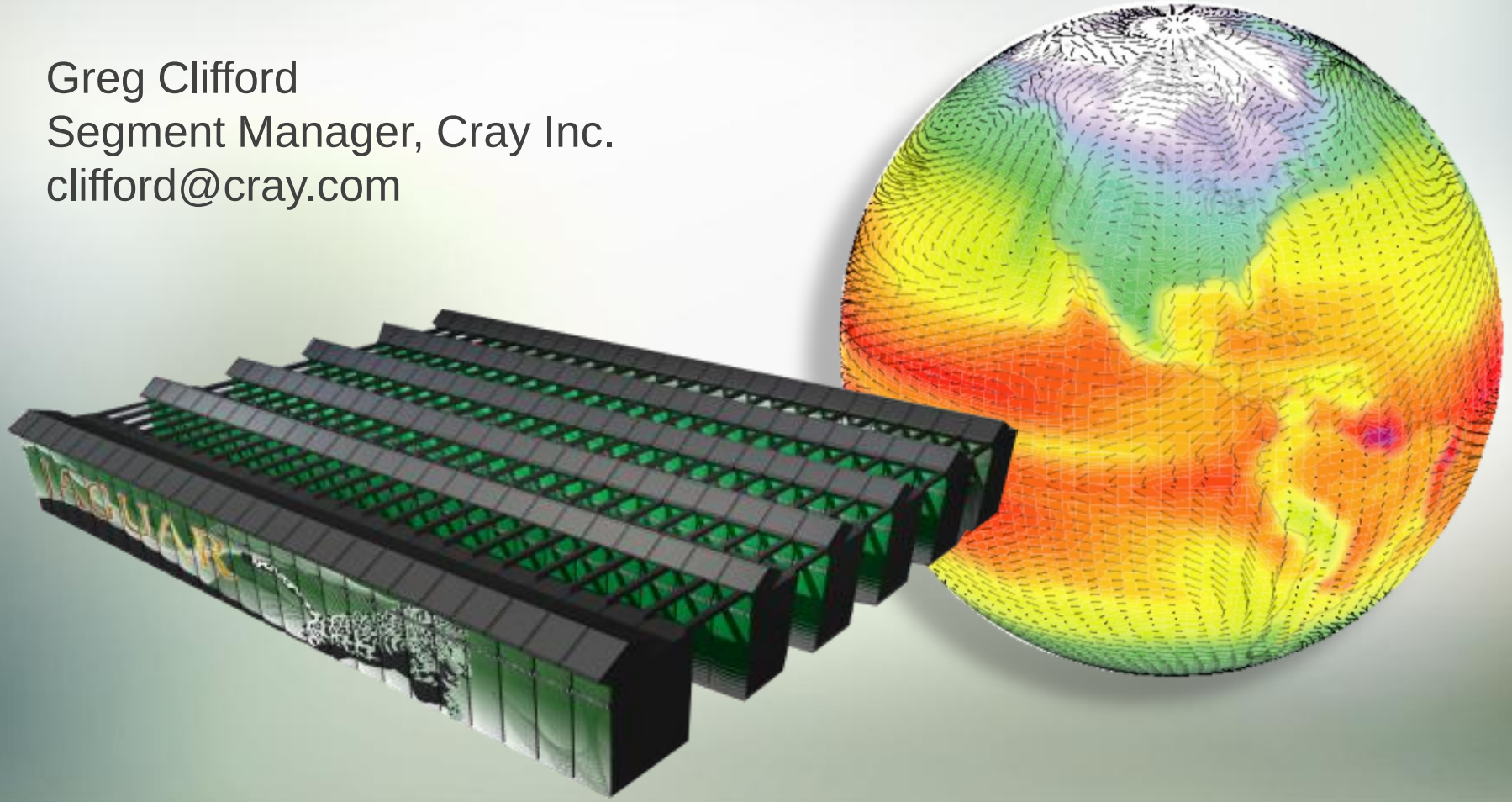
Breakthrough Science via Extreme Scalability

Greg Clifford
Segment Manager, Cray Inc.
clifford@cray.com



Presentation Agenda

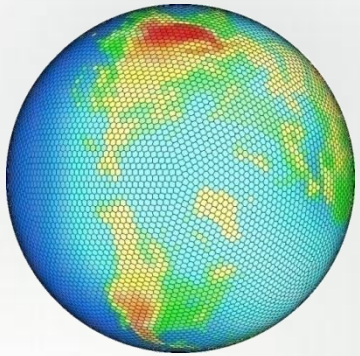
- Cray's focus
- The requirement for highly scalable systems
- Cray XE6 technology
- The path to Exascale computing

Cray Inc. Overview

- 35 year legacy focused on building the worlds fastest computer.
- 850 employees world wide
  - Growing in a tough economy
- Cray XT6 first computer to deliver a PetaFLOP/s in a production environment (Jaguar system at Oakridge)
- A full range of products
  - From the Cray CX1 to the Cray XE6
  - Options includes: unsurpassed scalability, GPUs, SMP to 128 cores & 2 Tbytes, AMD and Intel, InfiniBand and Gemini, high performance IO, …

Cray Development Focus

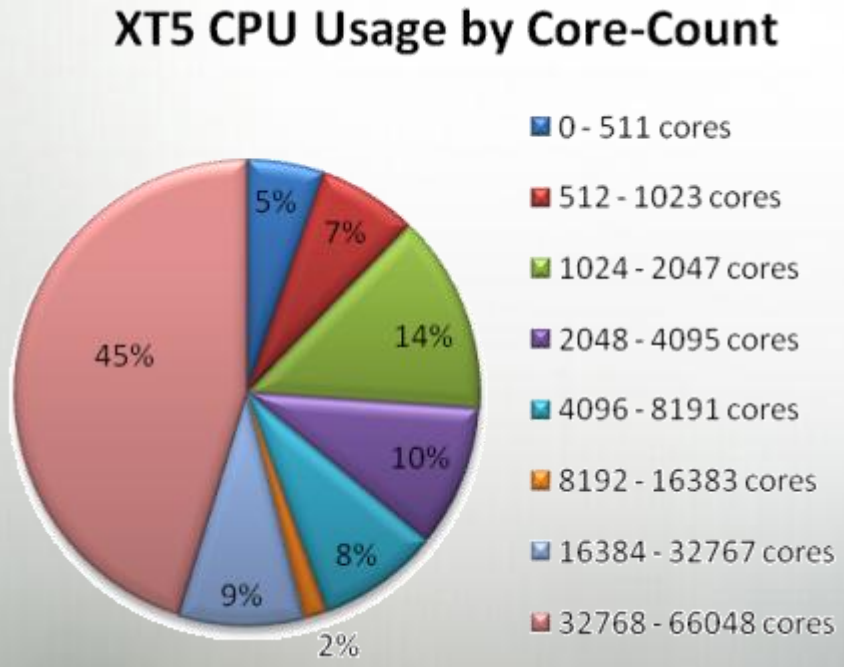
- **Designed for "mission critical" HPC environments:**
  - "when you can not afford to be wrong"
  - Sustainable performance on production applications
  - Reliability
- **Complete HPC environment.**
  - Focus on productivity
  - Cray Linux Environment, Compilers, libraries, etc
  - Partner with industry leaders (e.g. PGI, Platform, etc)
  - Compatible with Open Source World
- **Unsurpassed scalability/performance (compute, I/O and software)**
  - **Proprietary system interconnect (Cray Gemini router)**
  - **Performance on "grand challenge" applications**



Scalability in Production:
Usage Pattern – UT's Kraken Machine

Award(U. Tennessee/ORNL)	Sep, 2007
Cray XT3: 7K cores, 40 TF	Jun, 2008
Cray XT4: 18K cores,166 TF	Aug 18, 2008
Cray XT5: 65K cores, 600 TF	Feb 2, 2009
Cray XT5+: ~100K cores, 1 PF	Oct, 2009

Kraken and Krakettes!

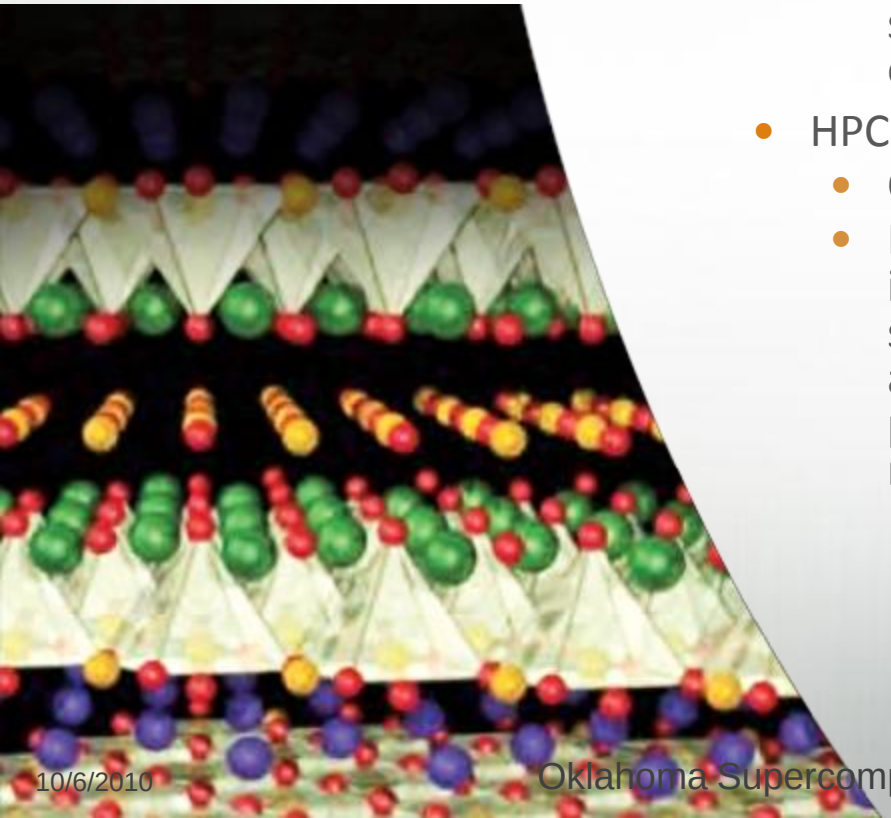


NICS is specializing on true capability applications, plus high performance file and archival systems.

Typical job size on IB cluster at TACC is ~300 cores

Unlocking the Mysteries of Superconducting Materials

OAK RIDGE National Laboratory



- Science Challenge
  - Find a superconductor that will exhibit its desirable characteristics – strong magnetic properties and the ability to conduct electricity without resistance or energy loss – without artificial cooling
- Computational Challenge:
  - Study chemical disorder in high temperature superconductors and the repulsion between electrons on the same atom
- HPC Solution
  - Cray XT5™ supercomputer "Jaguar"
  - Modified the algorithms and software design of its DCA++ code to maximize speed without sacrificing accuracy, achieving 1.352 petaflops and the first simulations with enough computing power to move beyond perfectly ordered materials

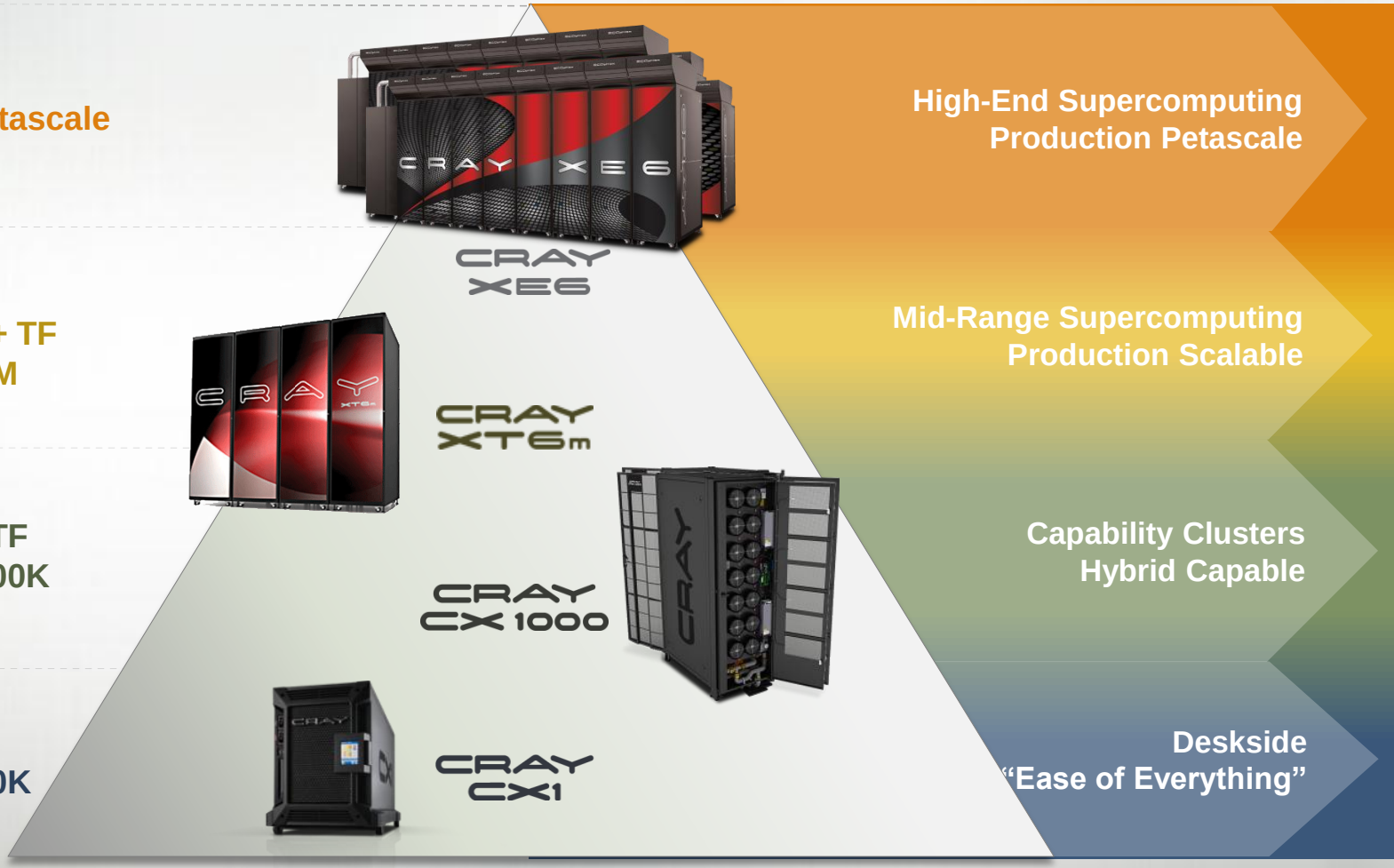
Understanding superconductors may lead to saving significant amounts of energy

Gordon Bell prize awarded to ORNL team

Materials simulation breaks 1.3 petaflops

- A team led by ORNL's Thomas Schulthess received the prestigious 2008 Association for Computing Machinery (ACM) Gordon Bell Prize at SC08
- The award was given to the team for attaining the fastest performance ever in a scientific supercomputing application
- The team achieved **1.352 petaflops on ORNL's Cray XT Jaguar** supercomputer with a simulation of superconductors
- By modifying the algorithms and software design of the DCA++ code, the team was able to boost its performance tenfold

Cray Product Portfolio



The Cray XE6

Scalable Performance

Gemini Interconnect for Multicore era

CLE3.x with ESM

Sustained Petaflops

1M+ cores

Improved Msg. Latency

Production Efficiency

ECOphlex Cooling

Network Resiliency

Warm Swap Blades

NodeKARE

Can Upgrade XT5/6

Adaptive Supercomputing

CLE3.x with CCM

X86/Linux Env.

Mature Software Ecosystem

Multiple File Systems

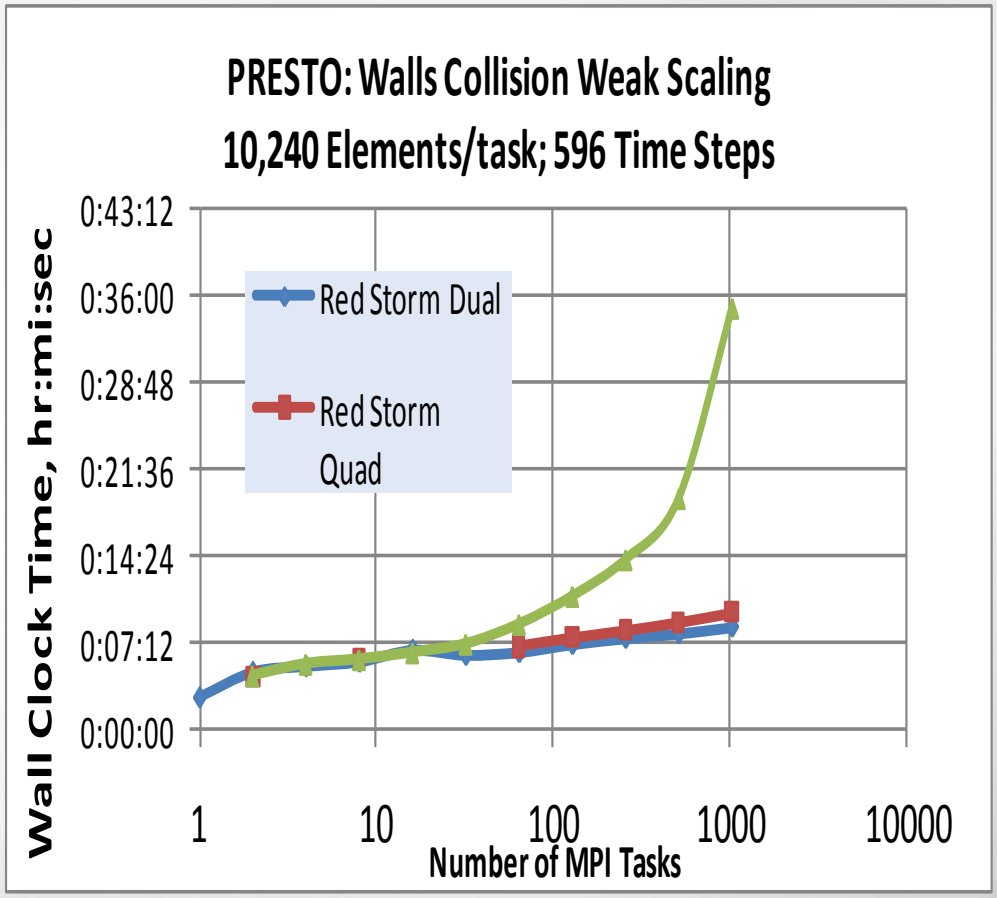
Cray XE 6 success: Over 5 PF's & $200M sold

Scalability Example: Cray interconnect vs. InfiniBand

- Explicit Lagrangian mechanics with contact
- Model: Two sets of brick-walls colliding
- Weak scaling analysis with 80 bricks/PE, each discretized with 4x4x8 elements
- Contact algorithm communications dominates the run time

- ***The rapid increase in run time after 64 processors on TLCC (Intel/InfiiniBand) can be directly related to the poor performance on Intel/IB system for random small-to-medium size messages***

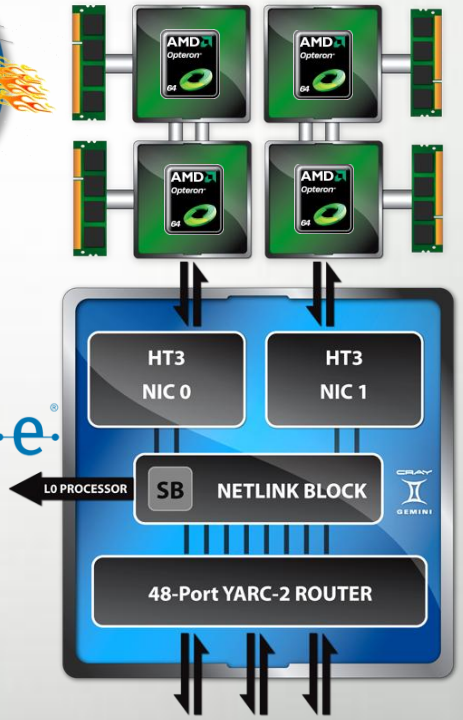
- TLCC/Quad run time ratio at 1024 is 4X.



Ref. Sandia National Labs, "Investigating the balance between capacity and capability workloads across large scale computing platforms", M.Rajan et. al.

Cray XE6: Built for Scalable Performance

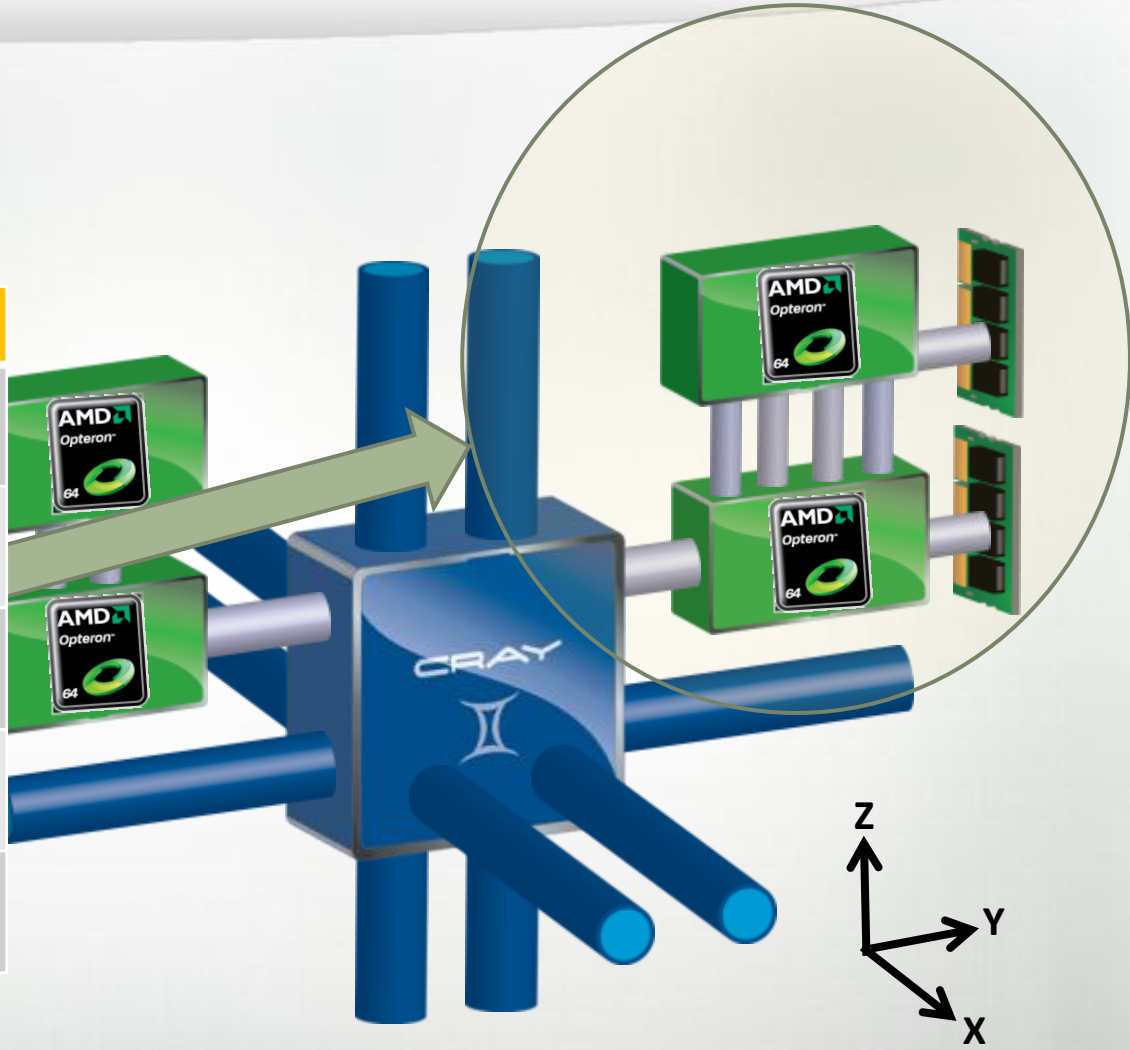
- **Gemini network improves performance**
  - 3-D Torus Scalability to Petaflops
  - Global Address Space
  - High Messaging Rates
  - Low Latency
  - 1M+ core scalability
- **AMD Opteron 6100 Series Processors**
  - 12 and 8-core performance
- **Extreme Scalability Mode(ESM)**
- **Cray Performance Environment**
  - Optimized Libraries and Communication
- **Improved Parallel I/O with Lustre 1.8**



lustre®

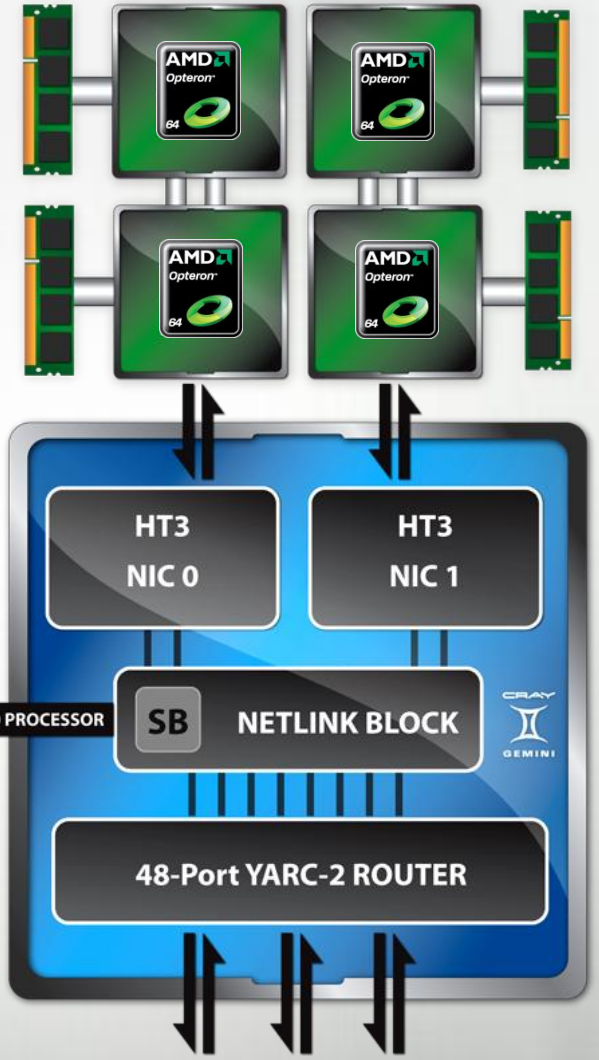
Cray XE6 Node and Gemini Interconnect

Node Characteristics	
Number of Cores	24 (Magny Cours)
Peak Performance MC-12 (2.2)	211 Gflops/sec
Peak Performance MC-8 (2.4)	153 Gflops/sec
Memory Size	32 GB per node 64 GB per node
Memory Bandwidth (Peak)	83.5 GB/sec



Cray Gemini Interconnect ASIC

- MPI Support
  - ~1.2 μs latency
  - ~15M independent messages/sec/NIC
  - BTE for large messages
  - FMA stores for small messages
  - One-sided MPI
- Advanced Synchronization and Communication Features
  - Globally addressable memory
  - Atomic memory operations
  - Pipelined global loads and stores
  - ~25M (65M) independent (indexed) Puts/sec/NIC
  - Efficient support for UPC, CAF, and Global Arrays
- Embedded high-performance router
  - Adaptive routing
  - Scales to over 100,000 endpoints
  - *Advanced resiliency features*



Cray XT6m Supercomputer

Cray MPP product for the mid-range HPC market using proven Cray XT6 technologies

- **Leading Price/Performance**
  - Under $300K
- Divisional/Supercomputing HPC configurations
  - 1-6 Cabinets
- “Right-sized” Interconnect
  - 2D Torus Topology
- Proven “petascale” hardware and software technologies
- New “Customer Assist” Service Plan

Recent Cray Systems in the Academic Community

Cray's position and focus on HPC combined with the introduction of the Cray XT6m has produced many new Academic customers for the Cray architecture.

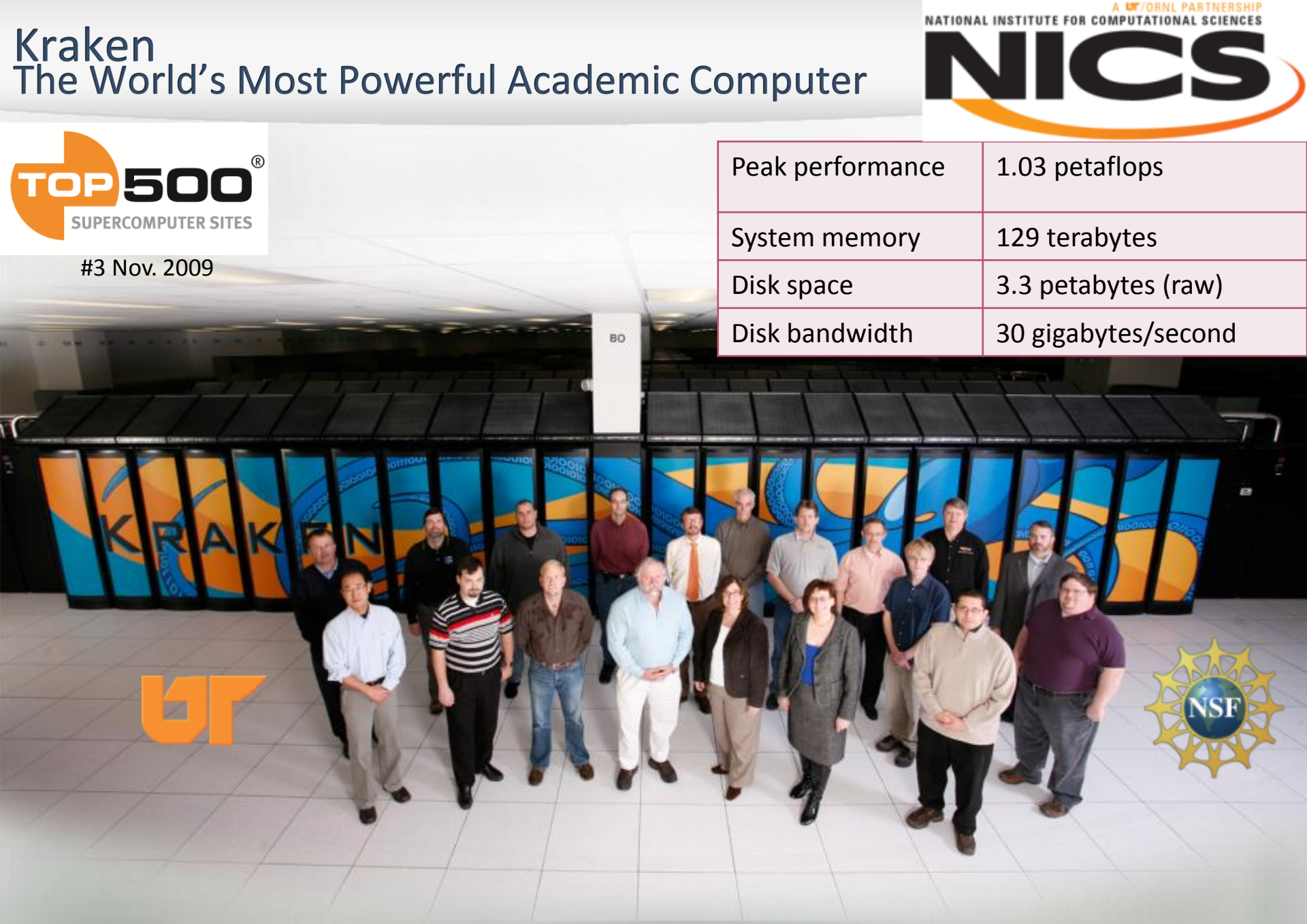
Kraken
The World's Most Powerful Academic Computer

A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

TOP500® SUPERCOMPUTER SITES

#3 Nov. 2009

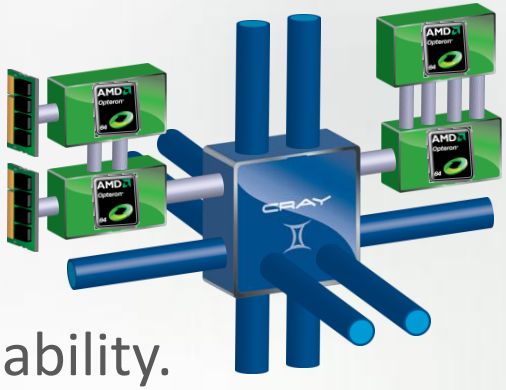
Peak performance	1.03 petaflops
System memory	129 terabytes
Disk space	3.3 petabytes (raw)
Disk bandwidth	30 gigabytes/second



Why the Cray XE6 in Higher Education?

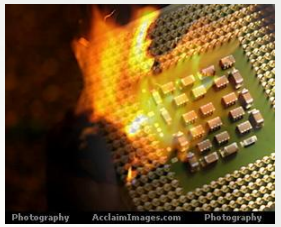
- **Low introductory price with the "m" series of Cray products.**
  - Leverage the Cray technology at a much lower price point.
- **Gemini interconnect provides better PGAS support (UPC, Co-Array, Chapel) – New class of applications.**
  - This has been a key issue with several of our academic customers.
- **Access to "Cray community" of HPC developers.**
  - Compatibility with large, national resources like Kraken, NERSC, ORNL, HECToR, etc.
  - Local resource for initial development and research with the ability to scale up to the largest system in the world.
- **Compatible with Open Source World**
  - Enhanced support for MPI,
- **GPU roadmap and scalability.**
  - Scalable performance combined with GPU accelerators.
  - *"Imagine what you could do with a Cray"*

HPC opportunities and challenges



- **Multi-core to "many core"**
  - Currently at 24 cores/node
  - Next generation will be 16 cores.
  - Increased requirement for application scalability.
- **Applications development/programming models**
  - PGAS programming model increasingly important
  - Ease of programming is critical for wide spread acceptance
  - PGAS hardware support in Gemini interconnect.
- **Accelerators (e.g. GPUs)**
  - Currently in the Cray CX product family has GPUs in the CX1 and CX1000
  - Sept 21, 2010: "Cray to add GPU's to XE6…"

Cray's Exascale Focus



- Major challenges to reach an Exaflop
  - Power
  - Programming difficulty
  - Concurrency (exposing parallelism)
  - Resiliency
- Cray is actively working on the Exascale challenges
  - Reducing PUE to create energy-efficient datacenters
  - Designing innovative, energy-efficient interconnects
  - Pursuing a high-performance, power-efficient, accelerated computing roadmap
  - Reducing OS jitter and creating extreme-scale operating systems
  - Enhancing system reliability through hardware techniques and OS resiliency
  - Focusing on programming tools, libraries and compilers for productivity
  - Pioneering new programming models: CAF, UPC,Chapel, GPU directives
  - Researching how to make applications resilient at extreme scale
- Cray is working closely with the mainstream processor vendors
  - Overall technical directions are similar (heterogeneous computing)
  - Cray is building our systems and software to be compatible
  - Consulting with them on needs for HPC applications and systems

Breaking Sustained Performance Barriers

1 GF – 1988: Cray Y-MP; 8 Processors

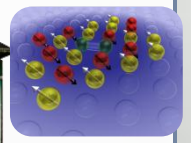
- Static finite element analysis

1 TF – 1998: Cray T3E; 1,024 Processors

- Modeling of metallic magnet atoms

1 PF – 2008: Cray XT5; 150,000 Processors

- Superconductive materials



1 EF – ~2018: ~10,000,000 Processors

Image Courtesy of Jamison Daniel,
National Center for Computational Sciences,
Oak Ridge National Laboratory.

Simulation Carbon-Land Model Intercomparison Project (C-LAMP) on Jaguar

The instantaneous net ecosystem exchange (NEE) of CO_2 is shown as colors projected onto the land surface from a C-LAMP simulation during July 2004. Green represents an uptake by the biosphere (negative NEE) while red represented a net flux into the atmosphere (positive NEE).

Thank you for your attention.

CRAY
THE SUPERCOMPUTER COMPANY

Backup slides

Recent XT & XE Customers

HLRS

fmi.fi

The Institute of Statistical Mathematics

Railway Technical Research Institute

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

UNIVERSITÄT DUISBURG ESSEN

SAHA INSTITUTE OF NUCLEAR PHYSICS
CALCUTTA
ACHIEVEMENT OF KNOWLEDGE

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

INPE

KMA
KOREA METEOROLOGICAL ADMINISTRATION

ARSC

ERDC
MSRC

NERSC

Sandia National Laboratories

Los Alamos
NATIONAL LABORATORY
EST. 1943

AIR FORCE RESEARCH LABORATORY

CRAY
THE SUPERCOMPUTER COMPANY

Software

CLE3, An Adaptive Linux OS designed specifically for HPC

CRAY LINUX ENVIRONMENT CLE3

ESM – *Extreme Scalability Mode*

- No compromise *scalability*
- Low-Noise Kernel for scalability
- Native Comm. & Optimized MPI
- Application-specific performance tuning and scaling

CCM –*Cluster Compatibility Mode*

- No compromise *compatibility*
- Fully standard x86/Linux
- Standardized Communication Layer
- Out-of-the-box ISV Installation
- ISV applications simply install and run

CLE3 run mode is set by the user on a job-by-job basis to provide full flexibility

Cray Software Ecosystem