# Titan: ORNL's New System for Scientific Computing


Buddy Bland
Jim Rogers
Galen Shipman

*presented to:*

Cray Users Group 2011
Fairbanks, Alaska
May 24, 2011

# Abstract:

**Titan: ORNL's New System for Scientific Computing**

ORNL is planning to install a 10-20 petaflops computer system over the next 18 months that will be the next generation system for scientific computing for the U.S. Department of Energy. While there will be many similarities to the existing Jaguar system, there will also be architectural differences. In this paper, we discuss the accelerator based architecture of Titan and the reasons for our decision to go in this direction. We also discuss our choice of the file systems to support the system.

# Outline

- Success of Red Storm, XT3/4/5/6, XE6 architecture
  - Awards
  - Science Accomplishments
- What is beyond Jaguar
- Titan:  ORNL's new system for scientific computing
- Why Accelerators?
- Programming tools for Titan

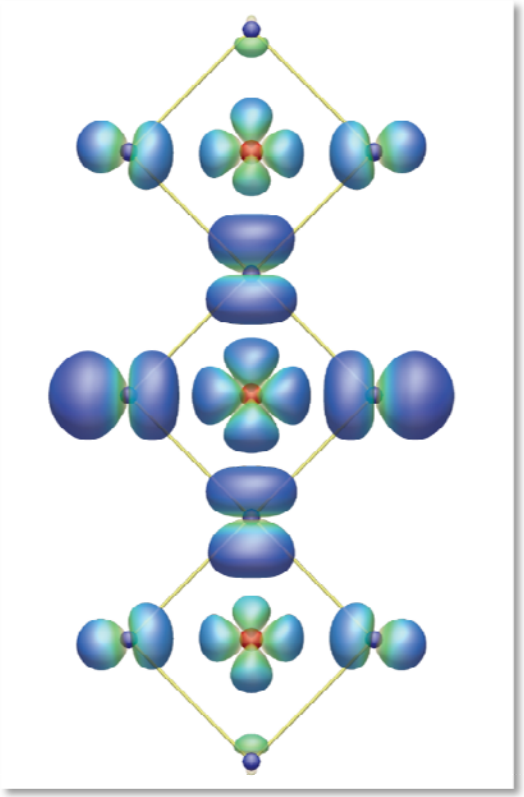
# 2010 Gordon Bell Winner

A high fidelity numerical simulation of blood flow by directly resolving the interactions of 200 million deformable red blood cells flowing in plasma. This simulation amounts to 90 billion unknowns in space, with numerical experiments typically requiring O(1000) time steps. This breakthrough is based on novel algorithms that we designed to enable distributed memory, shared memory, and vectorized - streaming parallelism. We present results on CPU and hybrid CPU-GPU platforms, including the new NVIDIA Fermi architecture and 200,000 cores of ORNL's Jaguar system. For the latter, we achieve over **0.7 Petaflop/s** sustained performance. **Our work demonstrates the successful simulation of complex phenomena using heterogeneous architectures and programming models at the petascale.**

Abtin Rahimian, GT
Ilya Lashuk, GT
Aparna Chandramowlishwaran, GT
Dhairya Malhotra, GT
Logan Moon, GT
Aashay Shringarpure, GT
Richard Vuduc, GT
**George Biros, GT**
Jeffrey Vetter, **ORNL**/GT
Rahul Sampath, **ORNL**
Denis Zorin, NYU
Shravan Veerapaneni, NYU

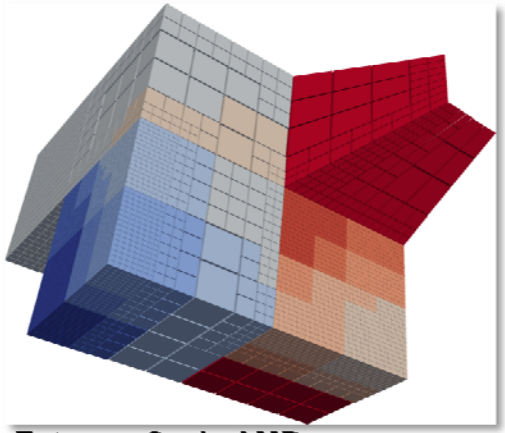
# 2010 Gordon Bell Honorable Mention

Methods based on the many-body Green's function are generally accepted as the path forward beyond Kohn-Sham based density functional theory, in order to compute from first principles electronic structure of materials with strong correlations and excited state properties in nano- and materials science. Here we present an efficient method to compute the screened Coulomb interaction W, the crucial and computationally most demanding ingredient in the GW method, within the framework of the all-electron Linearized Augmented Plane Wave method. We use the method to compute from first principles the frequency dependent screened Hubbard U parameter for $La_2CuO_4$, the canonical parent compound of several cuprate high-temperature superconductors. **These results were computed at scale on the Cray XT5 at ORNL, sustaining 1.30 petaflop.**



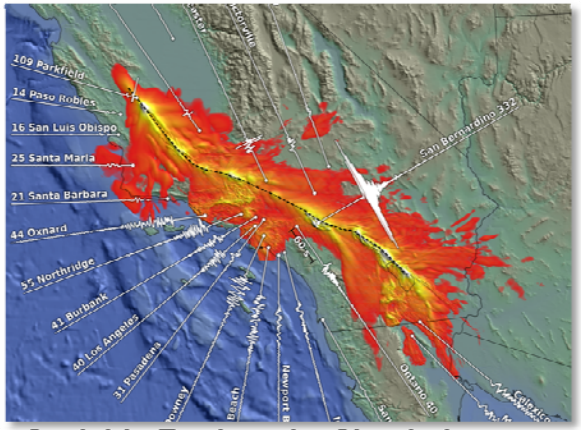
Anton Kozhevnikov - ETH Zürich
Adolfo G. Eguiluz - University of Tennessee, Knoxville
Thomas C. Schulthess - ETH Zürich

# 4 of 6 ACM 2010 Gordon Bell Prize Finalists use Jaguar – Winner of 4 prizes in 2008, 2009, & 2010



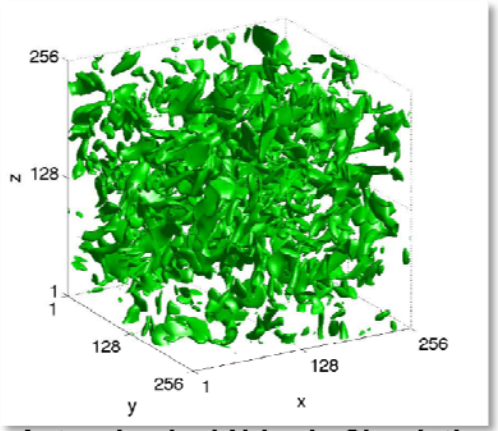
## Extreme-Scale AMR

Scaling difficult Adaptive Mesh Refinement techniques to over 224,000 cores on **Jaguar** demonstrating excellent scaling.



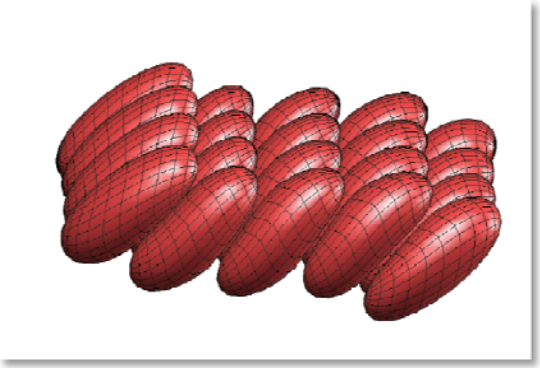

## Scalable Earthquake Simulation

Largest simulation of an earthquake ever performed shows a magnitude-8 quake and its impact on the region. Run on 223,074 cores on **Jaguar** opening new territory for earthquake science.



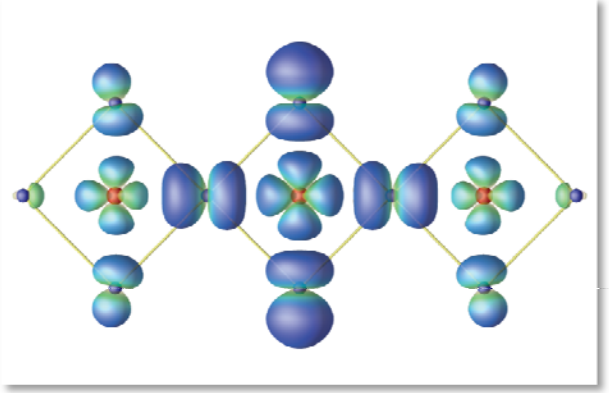

## Astrophysical N-body Simulation

An astrophysical N-body simulation with 3,278,982,596 particles using a treecode algorithm shows a sustained performance of 190.5 Teraflops on DEGIMA, a 144 node GPU cluster at Nagasaki U.



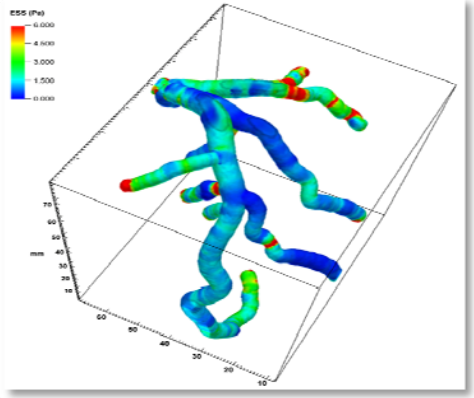
## DNS of Blood Flow on 200K Cores

The first high-fidelity petascale direct numerical simulation of blood flow that directly resolves the interactions of 200 million deformable red blood cells in plasma. Runs on GPUs, and achieves 700 Teraflops on 200k cores of **Jaguar**.



## Simulation of Excited Electron States

First principle simulations of excited state properties and strong electron correlations that are based on many-body Green's function approach run on **Jaguar** at over 1.3 Petaflops. These methods, that are generally accepted as a path forward beyond standard density functional theory, have now become tractable, and with continued increase in supercomputer performance they will become mainstream.



## Full Heart-Circulation System

The first large-scale simulation of blood flow in the coronary artieries, with a realistic description of human arterial geometry at spatial resolutions from centimeters down to 10 microns run on Jugene at Jülich.

# HPC Challenge Awards

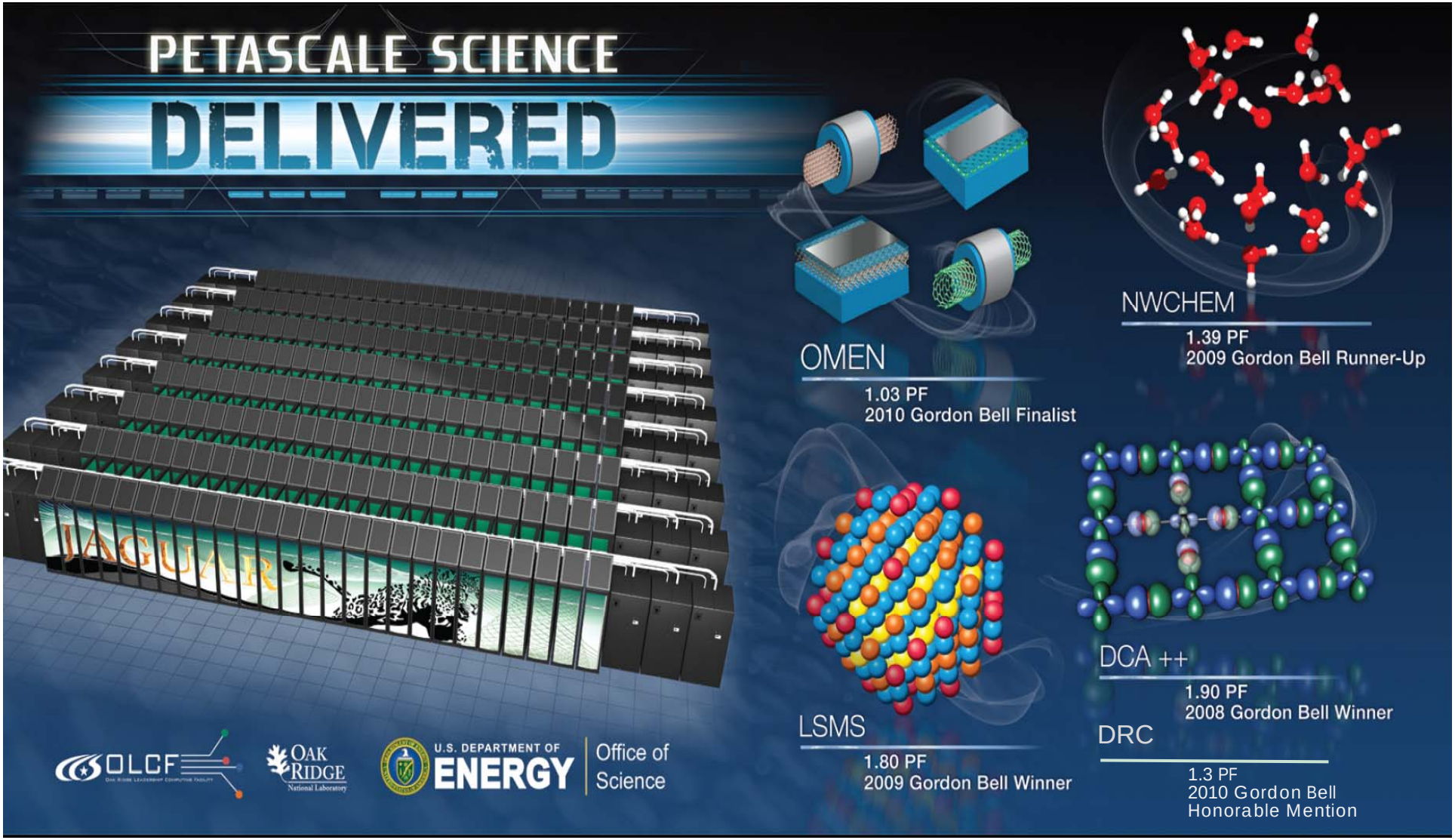
- HPC Challenge awards are given out annually at the Supercomputing conference
- Awards in four categories, result published for two others; tests many aspects of the computer's performance and balance
- Must submit results for all benchmarks to be considered
- Jaguar won 2 of 4 awards and placed 2nd and 3rd in the other two
- Jaguar had the highest performance on the other benchmarks
- Jaguar won 3 of 4 awards in 2009

| G-HPL (TF) | | EP-Stream (GB/s) | | G-FFT (TF) | | G-Random Access (GUPS) | | EP-DGEMM (TF) | | PTRANS (GB/s) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ORNL** | **1533** | **ORNL** | **398** | JAMSTEC | 12 | LLNL | 117 | **ORNL** | **2147** | **ORNL** | **13,723** |
| NICS | 736 | LLNL | 267 | **ORNL** | **11** | ANL | 103 | JAMSTEC | 1305 | SNL | 4,994 |
| LLNL | 368 | JAMSTEC | 233 | NICS | 8 | **ORNL** | **38** | NICS | 951 | LLNL | 4,666 |




Cray Users Group 2011, Buddy Bland

# We are delivering Petascale Science Today! Five applications running over 1 Petaflops



# BMI Uses Jaguar to Overhaul Long-Haul Trucks

## *Simulating energy-efficient trucks has the potential to significantly cut fuel costs*

- BMI Corporation, an engineering services firm has teamed up with ORNL, NASA, and several BMI corporate partners with large trucking fleets to increase the efficiency of tractor trailers.
- These rigs carry 75 percent of all US freight and supply 80 percent of its communities with 100 percent of their consumables. However, they also average 6 miles per gallon or less and annually emit some 423 million pounds of $CO_2$.
- To beef up its computing power, BMI applied for and received a grant through the ORNL Industrial HPC Partnerships Program for time on Jaguar.
- BMI engineers are now creating the most complex truck and trailer model ever simulated using NASA's FUN3D application for computational fluid dynamics analysis.
- BMI's ultimate goal is to design a sleek, aerodynamic truck with a lower drag coefficient than that of a low-drag car and anticipated fuel efficiencies running as high as 50 percent.

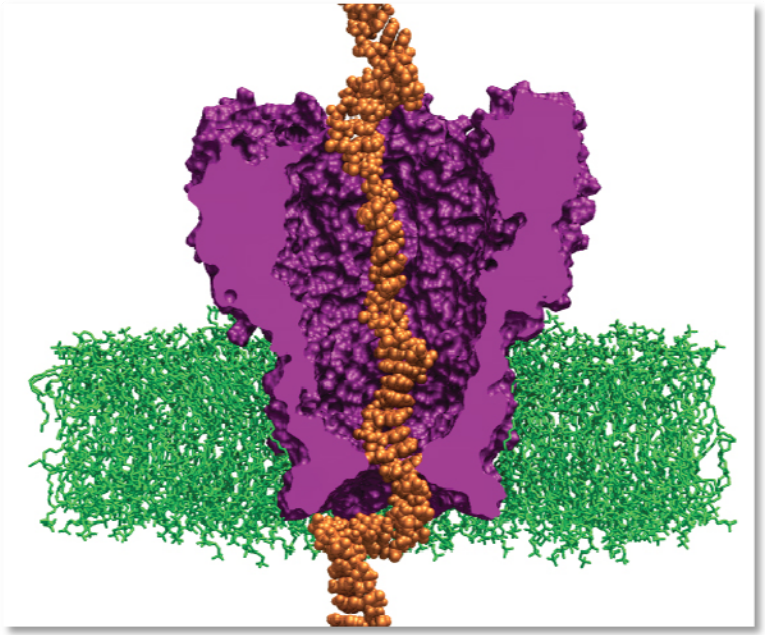
Trailers equipped with this front tray fairing and other BMI Corp. SmartTruck UnderTray components can achieve between 7 and 12 percent improvements in fuel mileage. (Photo courtesy of BMI SmartTruck)

"What Jaguar and FUN3D allow us to do is to break the truck into hundreds of pieces. We examine the drag on each piece and determine how it interacts across the entire system. Working at that level of detail and resolving the flow on each component is something you can't do with a small cluster." - BMI founder and CEO Mike Henderson

# Whole-Genome Sequencing Simulated on Supercomputers

*Scientists work to make personalized genomics affordable and quick for patients*

- A team led by Aleksei Aksimentiev of the University of Illinois–Urbana-Champaign is working to help create machines for personal genome sequencing that will be more accessible to hospitals, i.e. sequencing that can be performed in less than a day and for less than $1000.
- The research team uses the code NAMD, which calculates minimum energy states of atoms in a large biomolecular systems to determine their structure.
- The scientists were able to model the movement of DNA molecules through a MspA nanopore, confirming the possibility of a 100-fold slowdown of the strand through the nanopore. This is essential to enable single-nucleotide resolution, making genome sequencing viable and affordable.
- The ability to make genome sequencing affordable will enable such programs as the Cancer Genome Project, which characterizes DNA mutations in cancer cells in various tissues throughout all stages of cancer development.



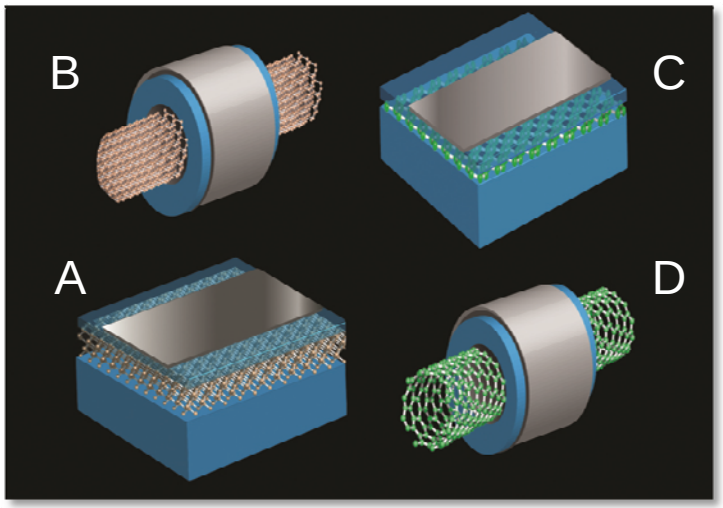
Scientists simulate DNA interacting with an engineered protein. The system may slow DNA strands travelling through pores enough to read a patient's individual genome. Image courtesy of Aleksei Aksimentiev.

# The Flight of the Electrons

*Team uses Jaguar to explore technology at the nanoscale*

- Gerhard Klimeck and Mathieu Luisier are exploring ways to advance computer chip design once transistors can get no smaller
- Applications—NEMO 3D and OMEN-- model electrons traveling through electronic devices
- Reached 1.03 petaflops on Jaguar using 220,000 processors
- Applications realistically model electron flow in 3D devices in reasonable time, making simulations practical for use in engineering design cycles

"Our understanding of electron flow in these structures is different. As you make things very small, you expose the quantum mechanical nature of the electrons. They can go around corners. They can tunnel. They can do all kinds of crazy stuff."—Gerhard Klimeck, leader of Purdue University's Nanoelectronic Modeling Group

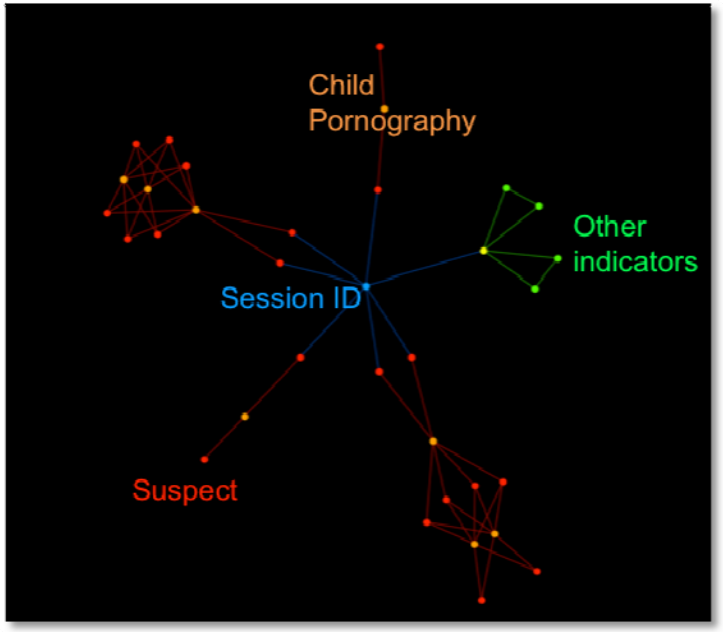



(a) Single-gate and doublegate ultra-thin-body FET made of Si, Ge, or III-V semiconductors; (b) gateall-around nanowire FET; (c) graphene nanoribbon FET; and (d) coaxially gated carbon nanotube FET. Image courtesy Gerhard Klimeck, Purdue University.

# Jaguar Pounces on Child Predators

## *Oak Ridge supercomputer will help identify producers of child pornography*

- To accelerate the acquisition of information needed to arrest child predators, law enforcement officers have teamed with experts at ORNL to speedily analyze the activities on file-sharing networks that pinpoint porn producers.
- Principle Investigator Thomas Patton has received funding from an industry sponsor and an allocation of 1 million processor hours on Jaguar.
- Patton has used part of the allocation for initial runs to test some clustering algorithms and later will use the remainder for development and testing using data provided by law enforcement.
- In a complex analysis, such as online child porn distribution, prioritization can save time. Data clusters can be distributed on the hundreds of thousands of processors of a leadership-class supercomputer for analysis.




File-sharing logs from law enforcement. The blue dot represents the file-sharing session. Red dots are computers offering files, and orange dots are child pornography files offered. Green dots indicate other potential crimes. Image credit: Analysis done by Tom Potok.

**"Across the globe criminals are using technology to facilitate the sexual exploitation of children. Police are overwhelmed and outnumbered. These Oak Ridge scientists are the good guys we've been waiting for. Their computers will become child rescue engines."**
**- Grier Weeks, executive director of the National Association to Protect Children, or PROTECT**

# What do we need to go beyond Jaguar

- Weak scaling of apps has run its course
- Need more powerful nodes for strong scaling
  - Faster processors but using much less power per GFLOPS
  - More memory
  - Better interconnect
- Hierarchical programming model to expose more parallelism
  - Distributed memory among nodes — 100K – 1M way parallelism
  - Threads within nodes — 10s – 100s of threads per node
  - Vectors within the threads — 10s – 100s of vector elements/ thread

1 Billion way parallelism needed for exascale systems



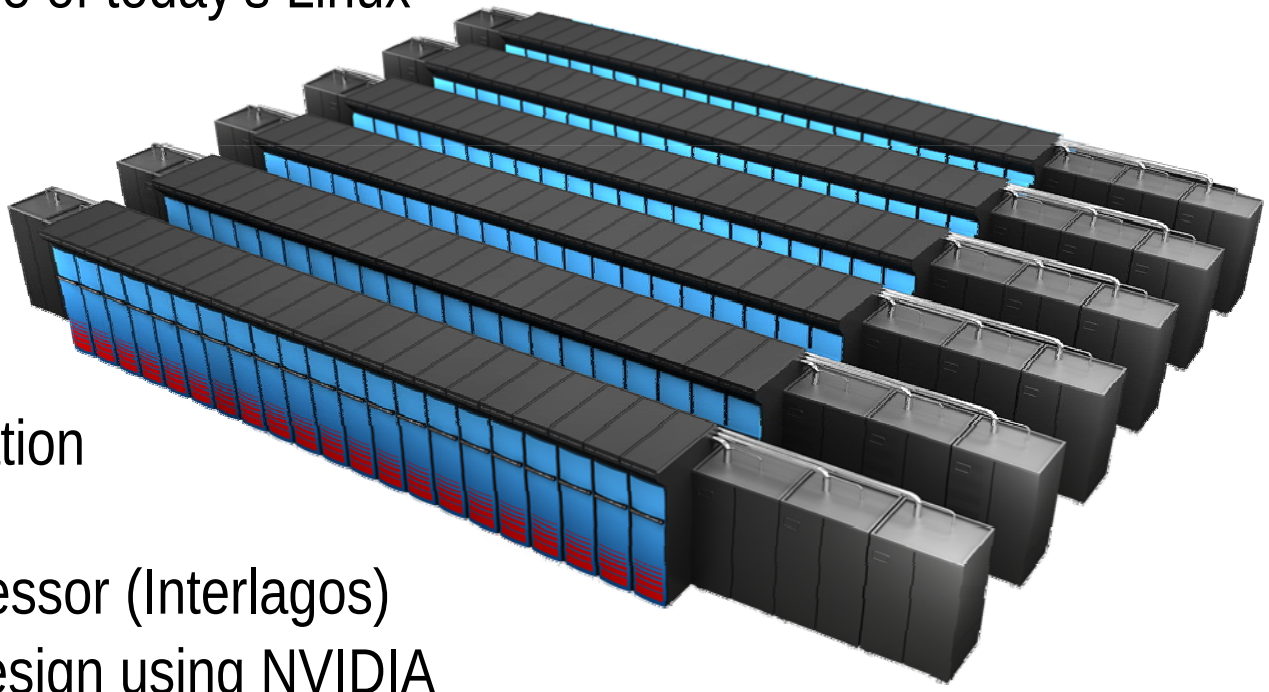
Cray Users Group 2011, Buddy Bland

# Disclaimer

ORNL has been given permission by the US Department of Energy to begin negotiations with Cray for a 10-20 petaflops system for delivery in 2012. The descriptions that follow are based on current thinking and plans, not a firm contract. We do not currently have a contract in place for any future system.

# ORNL's "Titan" System Goals

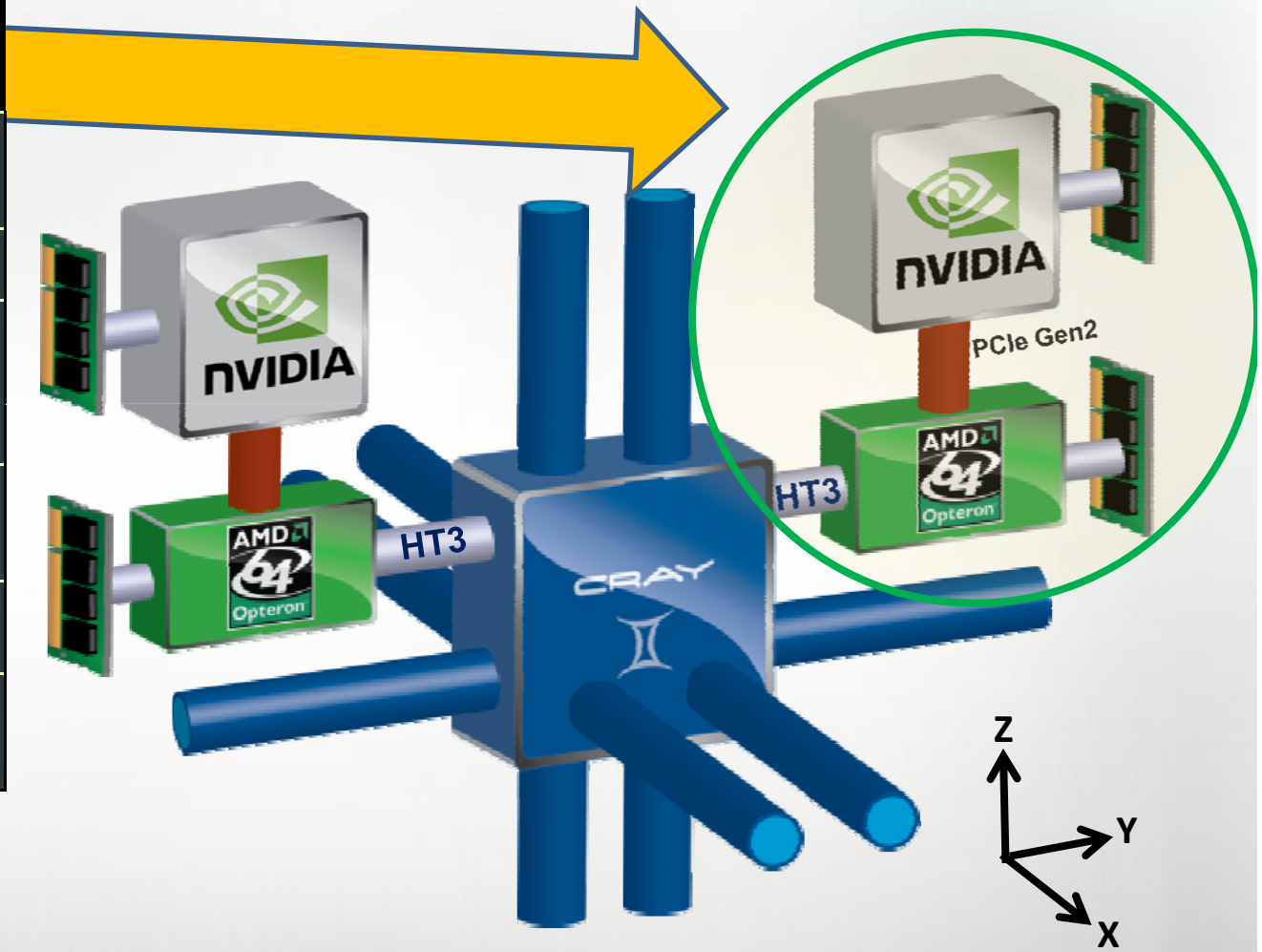
- Similar number of cabinets, cabinet design, and cooling as Jaguar
- Operating system upgrade of today's Linux operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- AMD Opteron 6200 processor (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
- 10-20 PF peak performance
  - Performance based on available funds
- Larger memory - more than 2x more memory per node than Jaguar



- **~1,000 accelerators late 2011 for application development**
- **~7,000 – 13,000 accelerators in late 2012**

# Cray XK6 Compute Node

| XK6 Compute Node Characteristics |
| --- |
| AMD Opteron 6200 Interlagos 16 core processor |
| Tesla X2090 @ 665 GF |
| Host Memory 16 or 32GB 1600 MHz DDR3 |
| Tesla X090 Memory 6GB GDDR5 capacity |
| Gemini High Speed Interconnect |
| Upgradeable to NVIDIA's KEPLER many-core processor |




Slide courtesy of Cray, Inc.

Cray Users Group 2011, Buddy Bland

# AMD 6200 “Interlagos” processor has 8 dual-core Bulldozer modules

## Bulldozer

**What it is:**

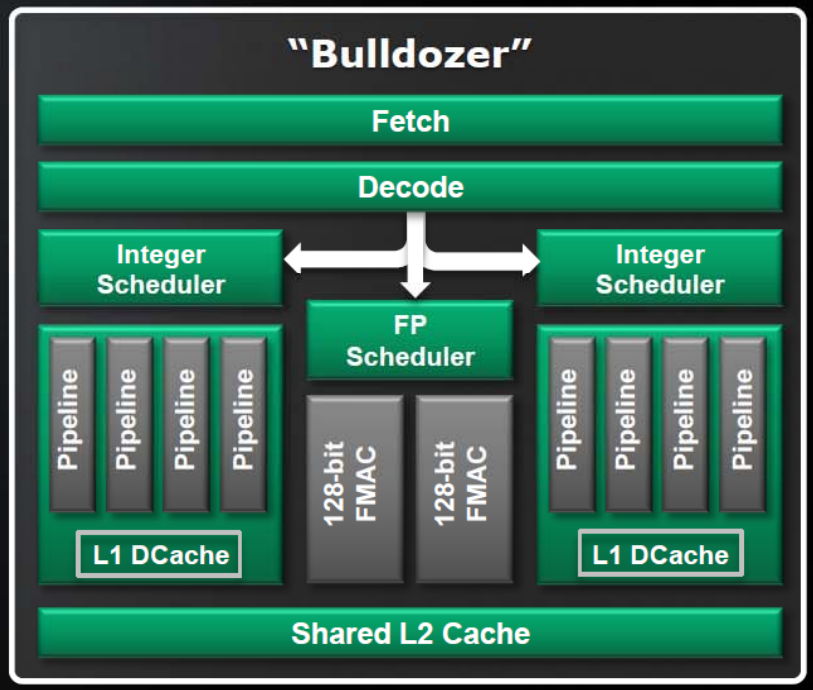
- A monolithic dual core building block that supports two threads of execution

**How it works:**

- Shares latency-tolerant functionality
- Smoothes bursty/inefficient usage
- Dynamic resource allocation between threads

**Customer Benefits:**

- Greater scalability and predictability than two threads sharing a single core
- Throughput advantages for multi-threaded workloads without significant loss on serial single-threaded workload components
- When only one thread is active, it has full access to all shared resources
- Estimated average of 80% of the CMP performance with much less area and power *

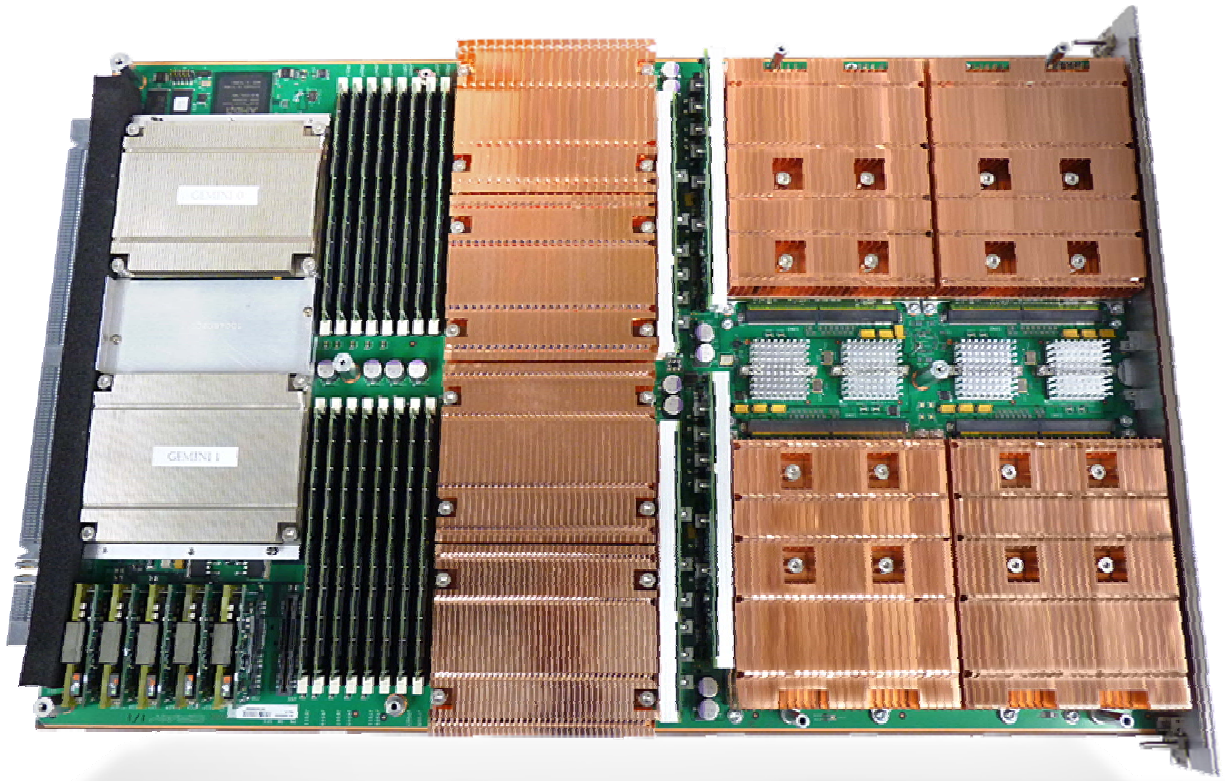





*Based on internal AMD modeling using benchmark simulations

AMD The future is fusion

# XK6 Compute Blade



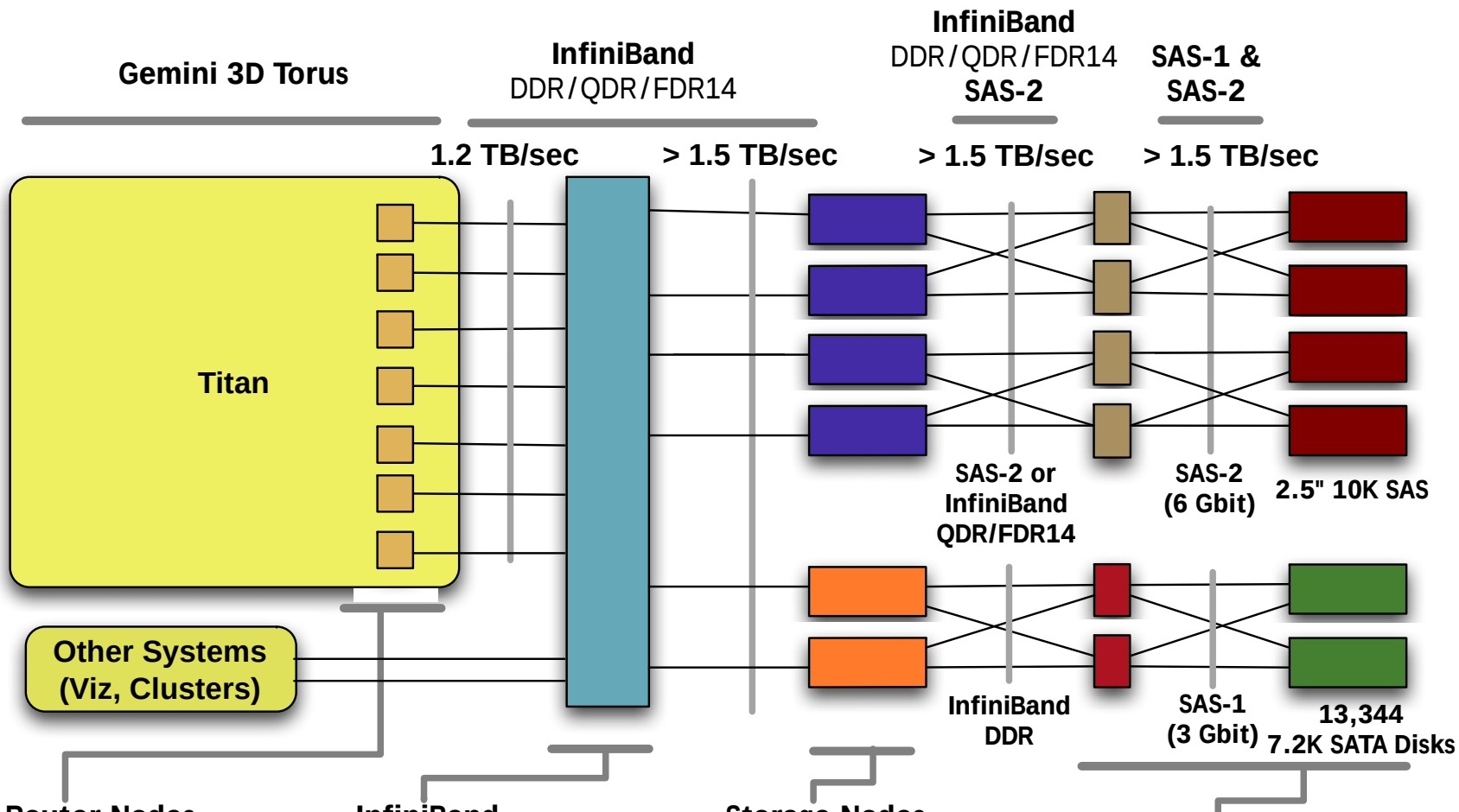
Slide courtesy of Cray, Inc.

Cray Users Group 2011, Buddy Bland

# What about the file system?


- We will continue to use Lustre™ as our file system
  - High-performance, scalable, center-wide accessibility
  - The only open-source solution
  - Strong community of users, developers, and integrators
- Expand our Spider file system infrastructure
  - Increase bandwidth by up to 1 TB/sec
  - Increase capacity by 10 – 30 Petabytes (disk technology dependent)
- Targeting Lustre 2.x
  - Enhanced metadata performance and resiliency under development
    - (NRE contract with Whamcloud)
  - Leveraging OpenSFS activities
    - Community engagement (requirements gathering, architecture, testing, etc.)
    - Next-generation feature development (accelerating the community roadmap)

# Titan Parallel I/O Architecture




**Lustre Router Nodes**
Forward I/O operations from Compute Clients to Storage Nodes
384 "XIO nodes"

**InfiniBand**
switch complex
1200+ port 32 or 54 Gbit/sec
*3000+ port 20 Gbit/sec InfiniBand switch complex*

**Storage Nodes**
run parallel file system software and manage incoming FS traffic.

**Enterprise Storage**
controllers and large racks of disks are connected via SAS

# Why did we choose to use an accelerator based system?

- I will show two slides that Steve Scott presented at the SciDAC 2010 conference that summarize the reasons
- See the full presentation at:

  http://computing.ornl.gov/workshops/scidac2010/presentations/s_scott.pdf

# Processor Architecture: Power vs. Single Thread Performance

- Multi-core architectures are a good first response to power issues
  - Performance through parallelism, not frequency
  - Exploit on-chip locality
- However, conventional processor architectures are optimized for single thread performance rather than energy efficiency
  - Fast clock rate with latency(performance)-optimized memory structures
  - Wide superscalar instruction issue with dynamic conflict detection
  - Heavy use of speculative execution and replay traps
  - Large structures supporting various types of predictions
  - Relatively little energy spent on actual ALU operations
- Could be much more energy efficient with multiple simple processors, exploiting vector/SIMD parallelism and a slower clock rate
- But serial thread performance is really important (Amdahl's Law):
  - If you get great parallel speedup, but hurt serial performance, then you end up with a niche processor (less generally applicable, harder to program)



Courtesy of Steve Scott
presented at SciDAC 2010

# Exascale Conclusion: Heterogeneous Computing

- To achieve scale and sustained performance per {$,watt}, must adopt:
  - …a *heterogeneous* node architecture
    - fast serial threads coupled to many efficient parallel threads
  - …a deep, explicitly managed memory hierarchy
    - to better exploit locality, improve predictability, and reduce overhead
  - …a microarchitecture to exploit parallelism at all levels of a code
    - distributed memory, shared memory, vector/SIMD, multithreaded
    - (related to the "concurrency" challenge—leave no parallelism untapped)

- This sounds a lot like GPU accelerators…
- NVIDIA Fermi™ has made GPUs feasible for HPC
  - Robust error protection and strong DP FP, plus programming enhancements
- Expect GPUs to make continued and significant inroads into HPC
  - Compelling technical reasons + high volume market
- Programmability remains primary barrier to adoption
  - Cray is focusing on compilers, tools and libraries to make GPUs easier to use
  - There are also some structural issues that limit applicability of current designs…
- Technical direction for Exascale:
  - Unified node with "CPU" and "accelerator" on chip sharing common memory
  - Very interesting processor roadmaps coming from Intel, AMD and NVIDIA….



Courtesy of Steve Scott
presented at SciDAC 2010

# How do you program these nodes?

GPU Tools BOF Wed @ 2:00

Rich Graham Thur @ 11:30

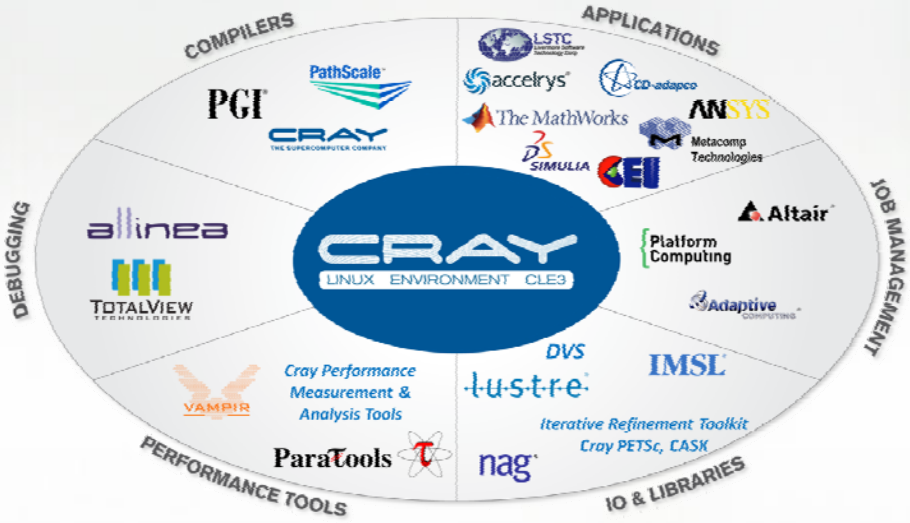
- We are working with several vendors on programming environment tools for Titan
  - Allinea DDT debugger scales to full system size and with ORNL support will be able to debug heterogeneous (x86/GPU) apps
  - CAPS HMPP compiler supports both C and Fortran compilation for heterogeneous nodes. ORNL is supporting C++ development
  - ORNL has worked with the Vampir team at TUD to add support for profiling codes on heterogeneous nodes
  - PGI has a compiler that generates code for the accelerators

DDT Talk Thur @ 9:30

Brent LeBack PGI Wed @ 2:00

# Software for the Cray XK6

## XK6 will benefit from the Cray Software Ecosystem for XE6




**BOF on Cray GPU Prog Tools Wed @ 4:15**

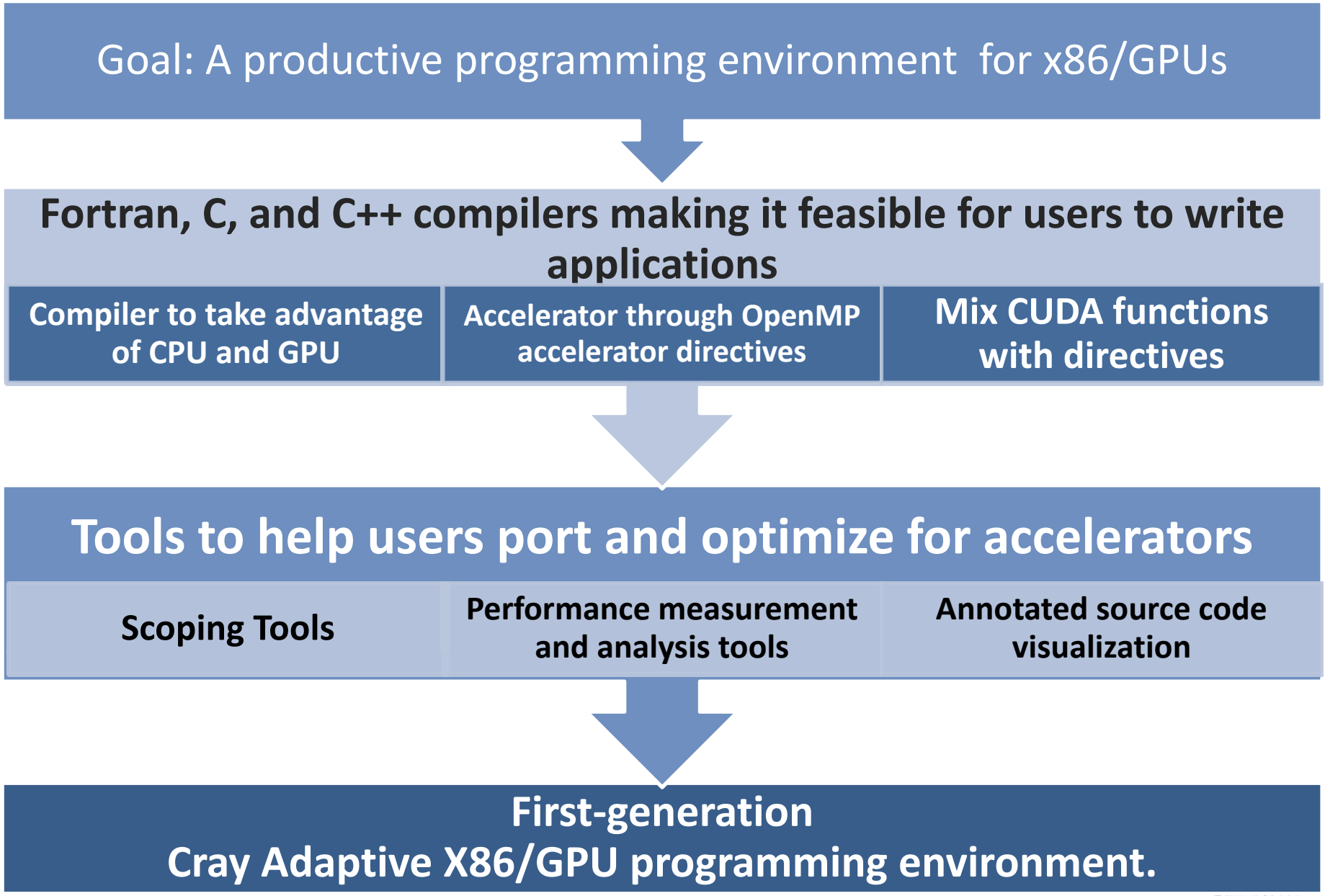
## And will be extended

- Cray Adaptive GPU/x86 Software Environment
- This will roll out throughout 2011-2012

**PGI for GPUs** | **CUDA NVIDIA SDK** | **Enhanced Libraries** | **Cray Compiler for GPUs** | **Cray Optimization Tools for GPUs** | **Debuggers**

Slide courtesy of Cray Inc.

# Cray Adaptive GPU/x86 Software Environment





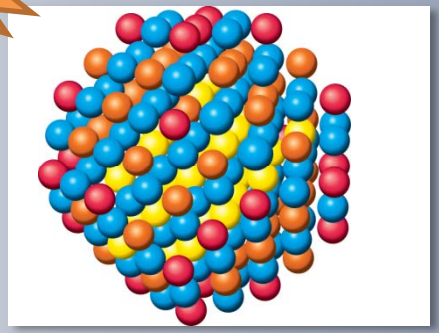

Slide courtesy of Cray Inc.

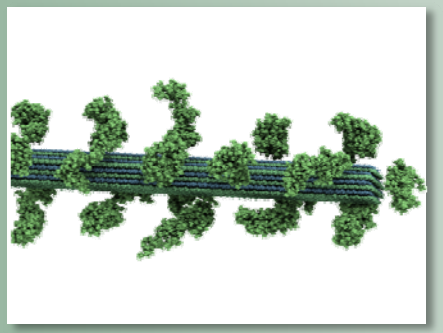
# Titan: Early Science Applications

Thur @ 1:00

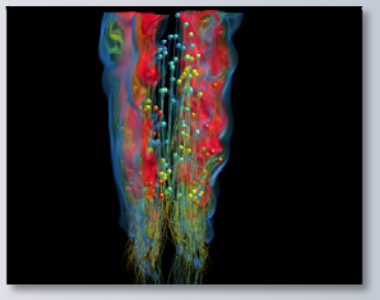
Bronson Messer will talk about this effort Wed. @ 11:15

WL-LSMS
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.



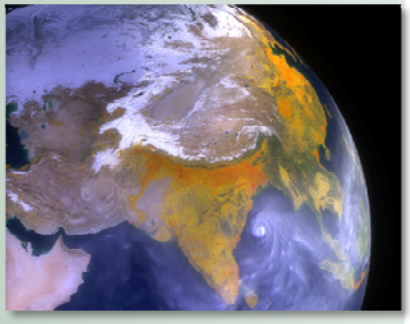


LAMMPS
Biofuels: An atomistic model of cellulose (blue) surrounded by lignin molecules comprising a total of 3.3 million atoms. Water not shown.

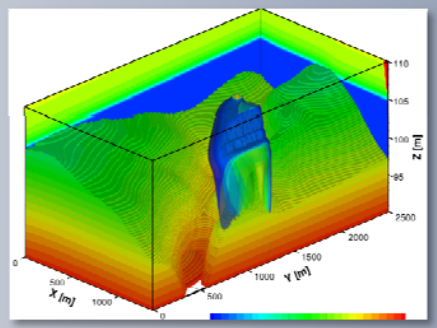


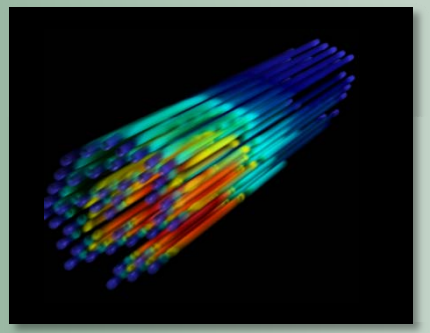
S3D
How are going to efficiently burn next generation diesel/bio fuels?
.

CAM / HOMME
Answer questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns/statistics and tropical storms



PFLOTRAN
Stability and viability of large scale $CO_2$ sequestration; predictive containment groundwater transport

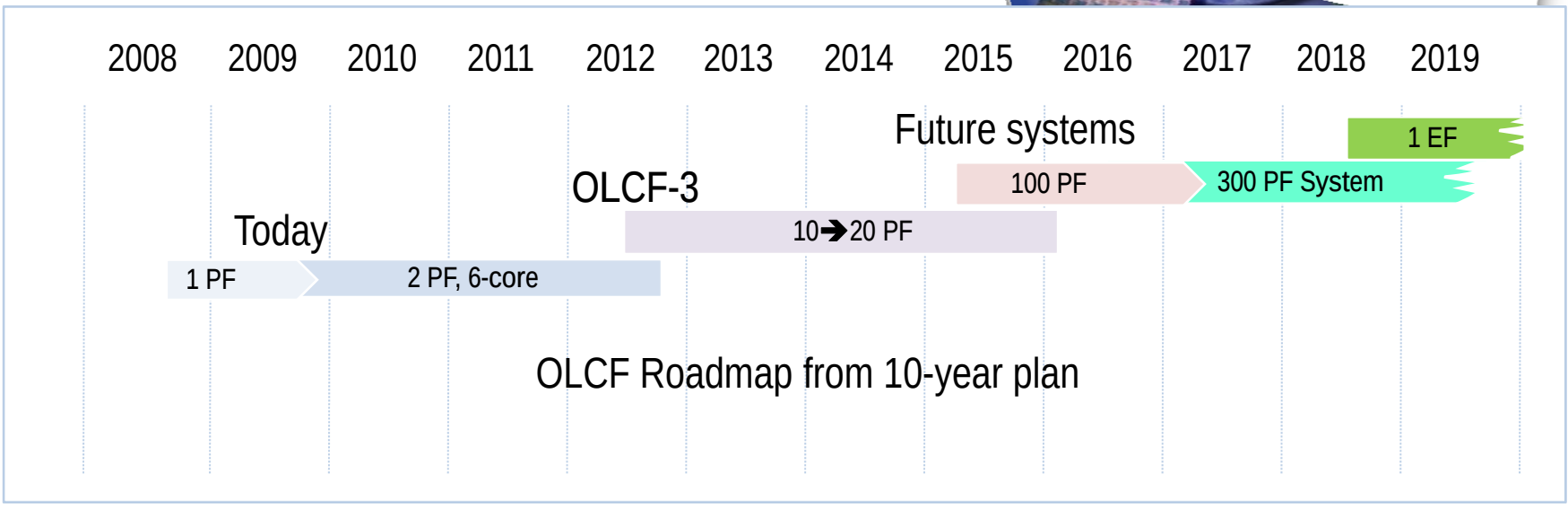


Denovo
Unprecedented high-fidelity radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

# 10 Year Strategy: 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
  - OLCF-3 Project: New 10-20 petaflops computer based on early hybrid multi-core technology

# Questions?

Buddy Bland
Email: BlandAS@ORNL.GOV

The research and activities described in this presentation were performed using the resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.