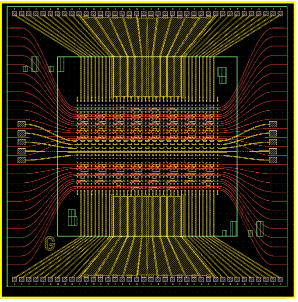
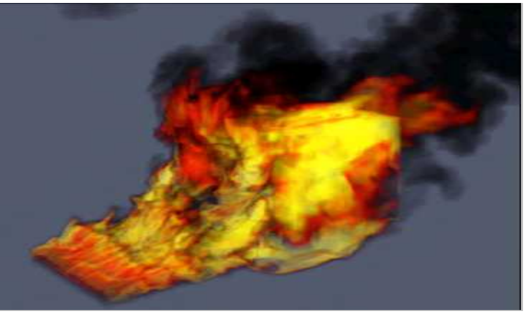
SAND2013-0566P





Sandia National Laboratories

Exceptional service in the national interest

Parallel Scientific Computing at the DOE National Laboratories: Successes and Challenges

Stephen Olivier

Scalable System Software Group

Computer Science Research Institute

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

CSRI Computer Science Research Institute

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Talk Overview

- The DOE labs: Who we are and what we do
- DOE supercomputers and their capabilities
- Present challenges, including multicore/manycore

Talk Overview

- The DOE labs: Who we are and what we do
- DOE supercomputers and their capabilities
- Present challenges, including multicore/manycore

What Are the DOE National Labs?

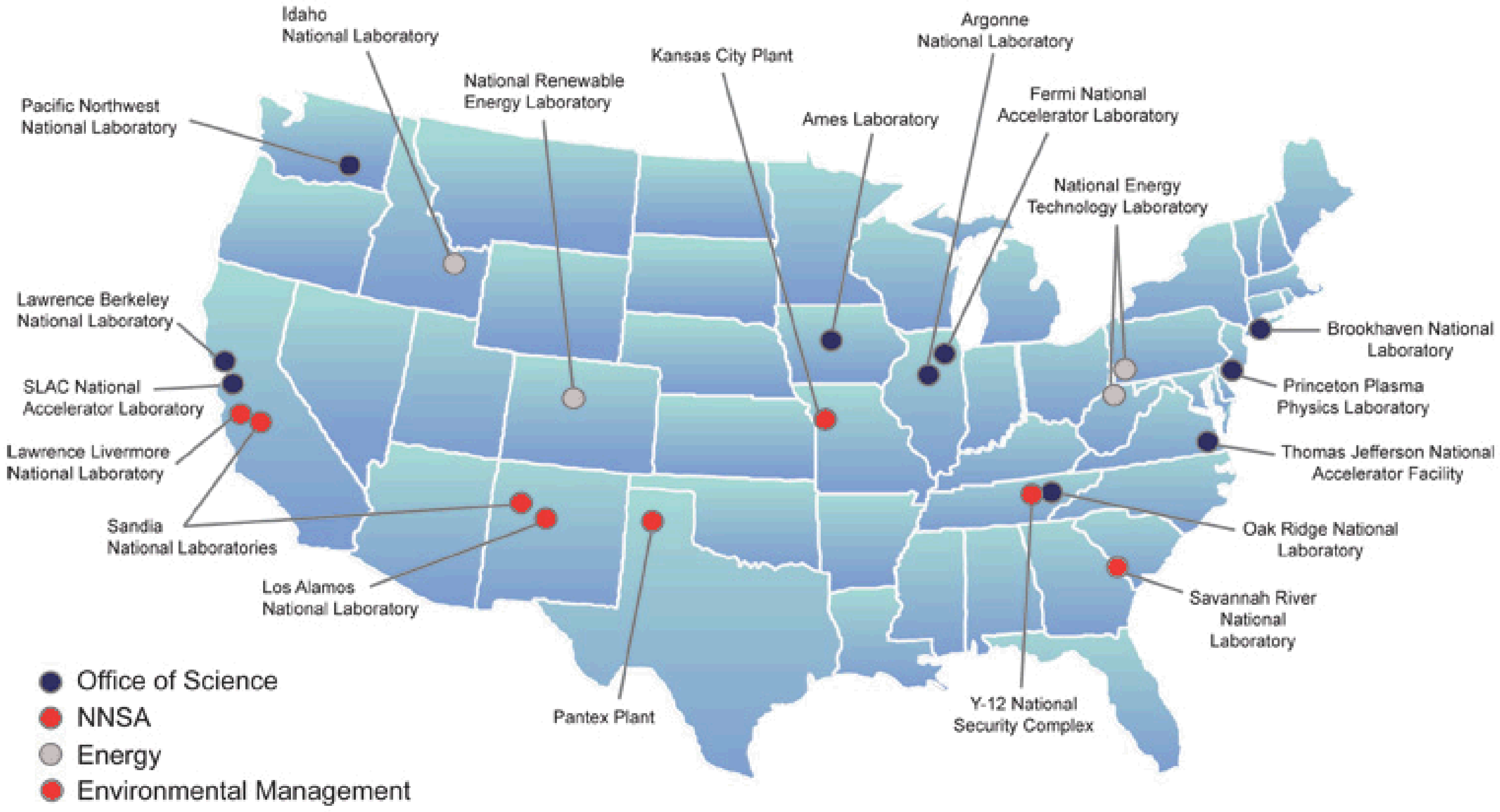
- “Together, the 17 DOE laboratories comprise a preeminent federal research system, providing the Nation with strategic scientific and technological capabilities:
  - Execute long-term government scientific and technological missions, often with complex security, safety, project management, or other operational challenges
  - Develop unique, often multidisciplinary, scientific capabilities beyond the scope of academic and industrial institutions, to benefit the Nation’s researchers and national strategic priorities
  - Develop and sustain critical scientific and technical capabilities to which the government requires assured access”

Source: energy.gov

What Do the National Labs Do?

- Office of Science Labs
  - "Advance the science needed for revolutionary energy breakthroughs
  - Unravel nature's deepest mysteries
  - Provide the Nation's researchers with the most advanced large-scale tools of modern science" - science.energy.gov
- National Nuclear Security Administration (NNSA) Labs
  - "Maintain the safety, security and effectiveness of the nuclear deterrent without nuclear testing" - nnsa.energy.gov
  - Multiprogram activities: leverage science and technology capabilities for other customers in both government and industry

Where Are the DOE National Labs?

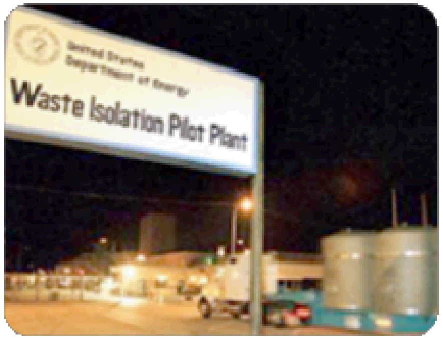

Idaho National Laboratory
Pacific Northwest National Laboratory
National Renewable Energy Laboratory
Kansas City Plant
Argonne National Laboratory
Ames Laboratory
Fermi National Accelerator Laboratory
National Energy Technology Laboratory
Lawrence Berkeley National Laboratory
SLAC National Accelerator Laboratory
Lawrence Livermore National Laboratory
Sandia National Laboratories
Los Alamos National Laboratory
Brookhaven National Laboratory
Princeton Plasma Physics Laboratory
Thomas Jefferson National Accelerator Facility
Oak Ridge National Laboratory
Savannah River National Laboratory
Pantex Plant
Y-12 National Security Complex
Office of Science
NNSA
Energy
Environmental Management

Source: energy.gov

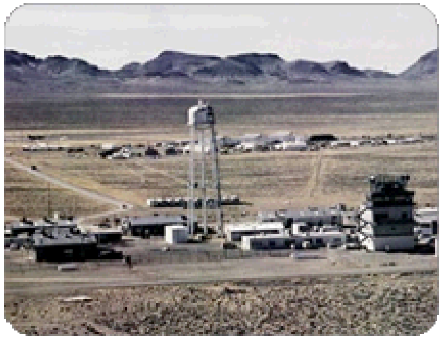
Sandia's Sites

Albuquerque, New Mexico

Carlsbad, New Mexico


United States Department of Energy
Waste Isolation Pilot Plant

Tonopah, Nevada



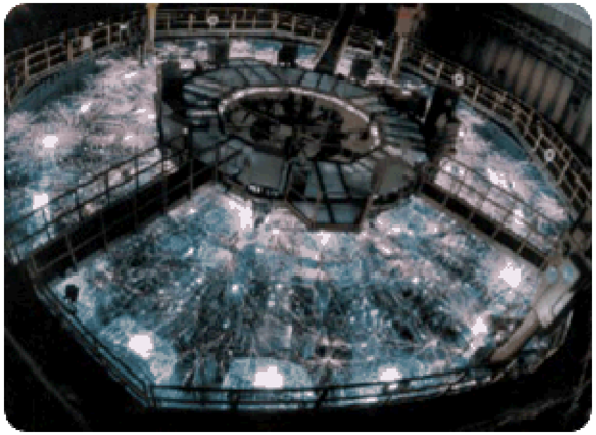
Livermore, California

Amarillo, Texas

Kauai, Hawaii

Nuclear Weapons

Pulsed power and radiation effects sciences

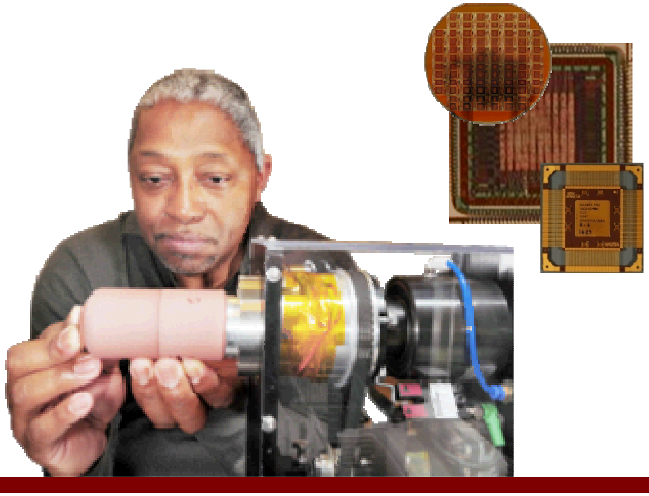


Design agency for nonnuclear components

- Neutron generators
- Arming, fuzing and firing systems
- Safety systems
- Gas transfer systems

Warhead systems engineering and integration

Production agency



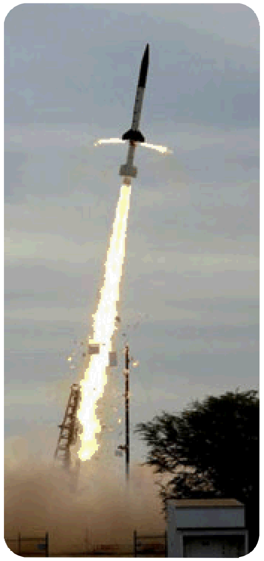
Defense Systems and Assessments

Synthetic aperture radar

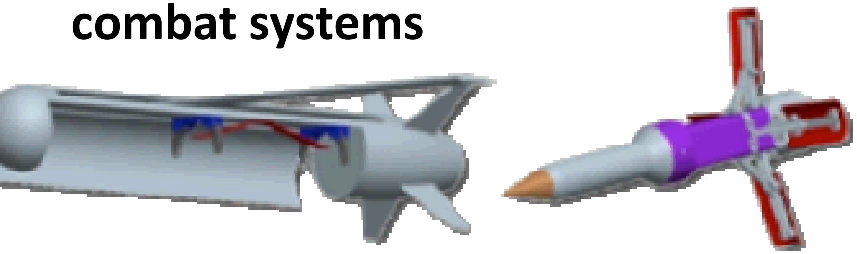
Mower activity
Human footprints

Support for NASA

Support for ballistic missile defense

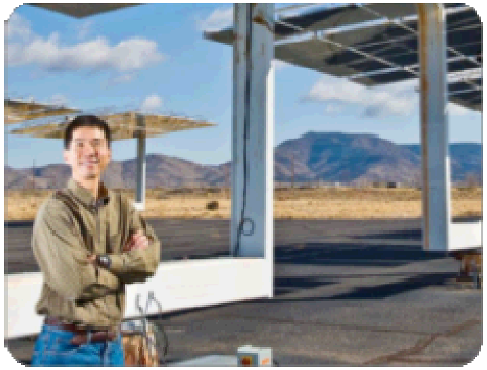


Ground sensors for future combat systems

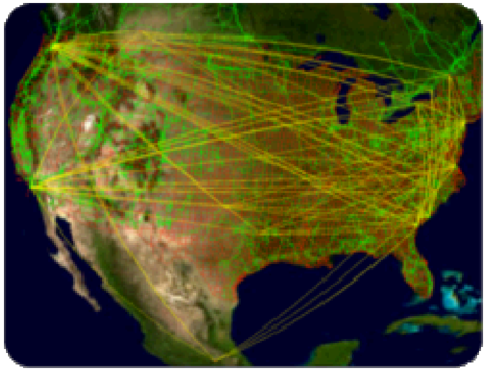


Energy, Climate, and Infrastructure Security

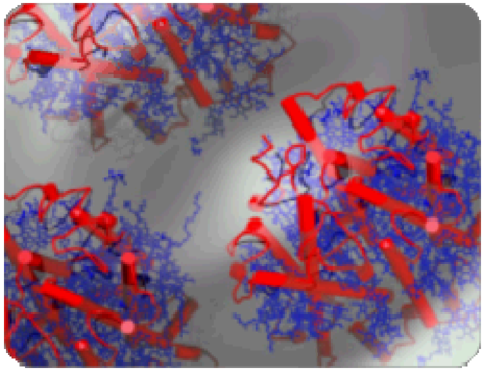
Energy



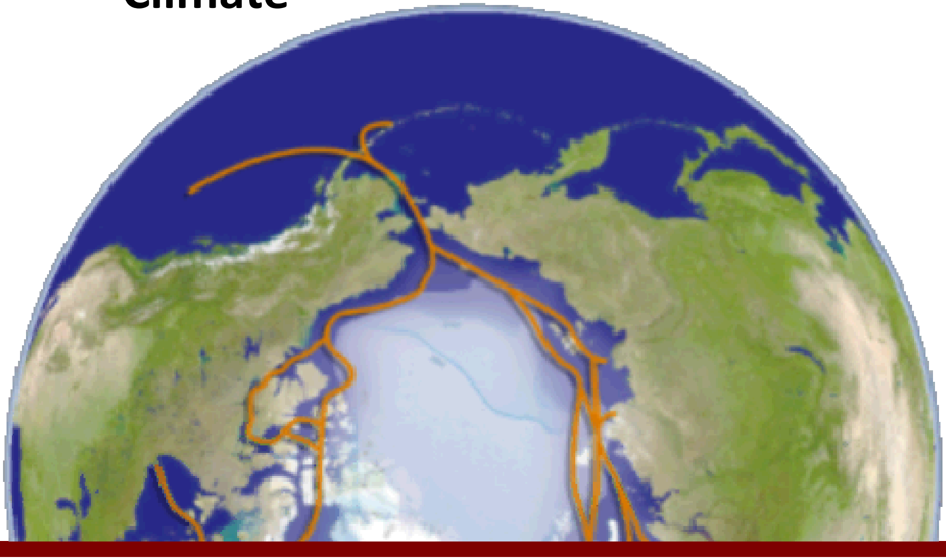
Infrastructure



Crosscuts and enablers



Climate



jbei
Joint BioEnergy Institute

International, Homeland, and Nuclear Security

Critical asset protection

Homeland defense and force protection

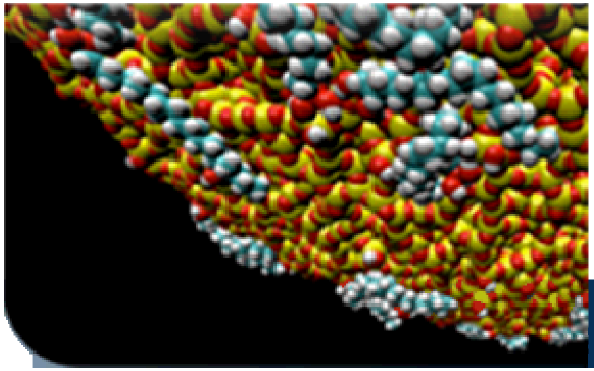
Homeland security programs

Global security

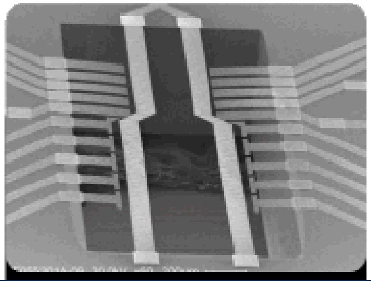
Science and Engineering Foundations

Computing and information science

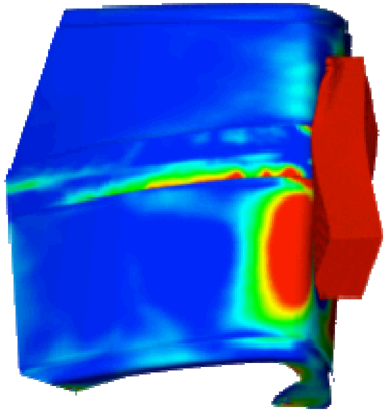
Materials science



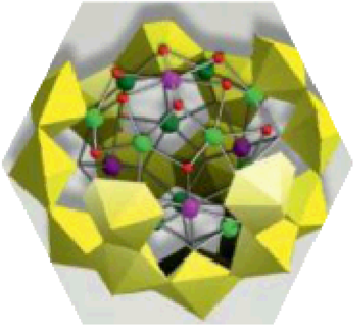
Nanodevices and microsystems



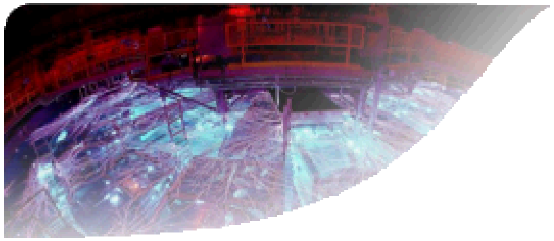
Engineering sciences



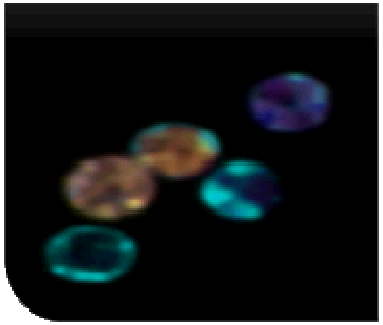
Geoscience



Radiation effects and high-energy density science



Bioscience

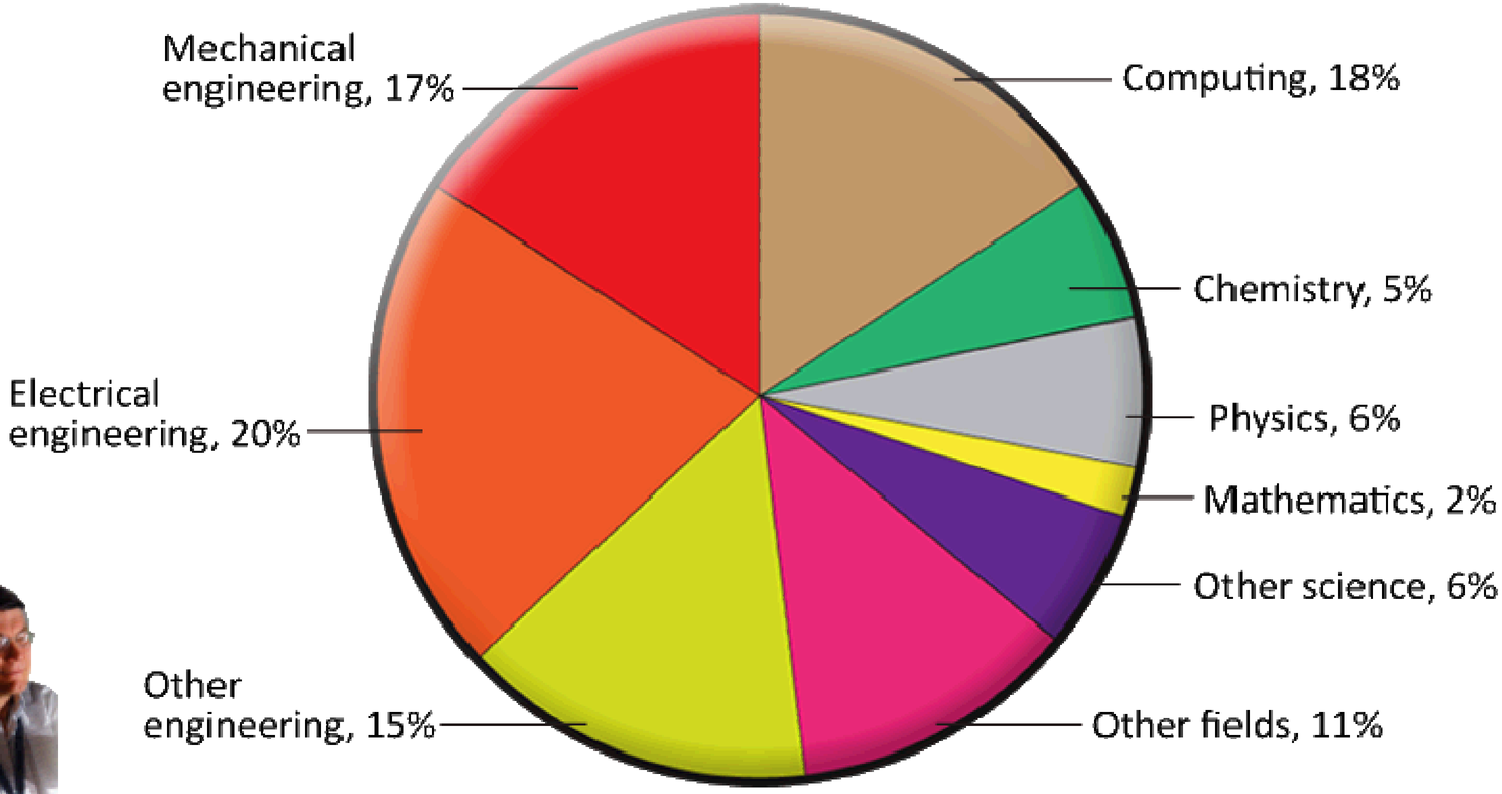


Our Workforce

- Onsite workforce: 11,711
- Regular employees: 9,238
- Gross payroll: ~$981M

Data for FY12 through end of September

Research & Development staff(4,682) by discipline

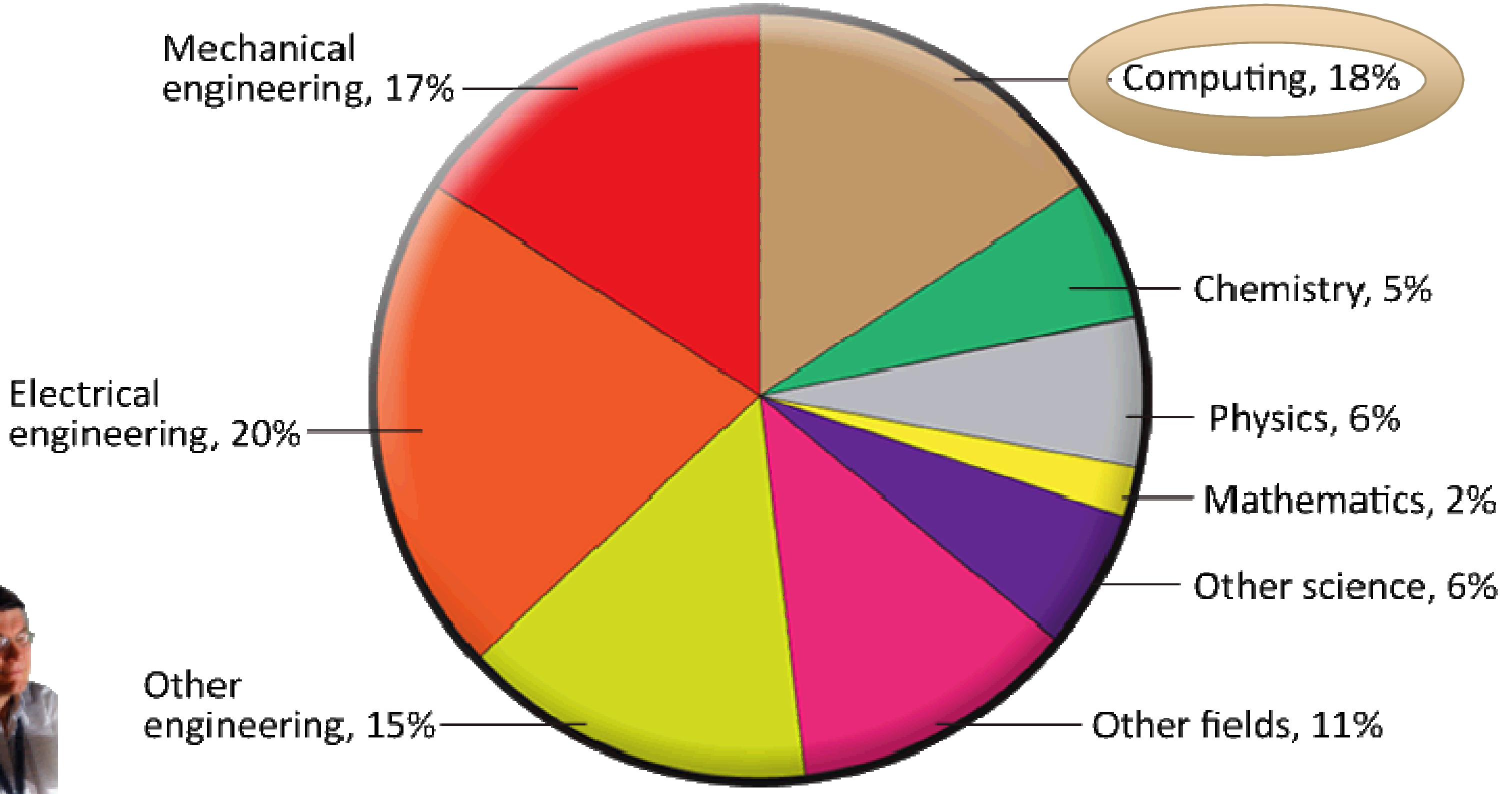


Our Workforce

- Onsite workforce: 11,711
- Regular employees: 9,238
- Gross payroll: ~$981M

Data for FY12 through end of September

Research & Development staff(4,682) by discipline

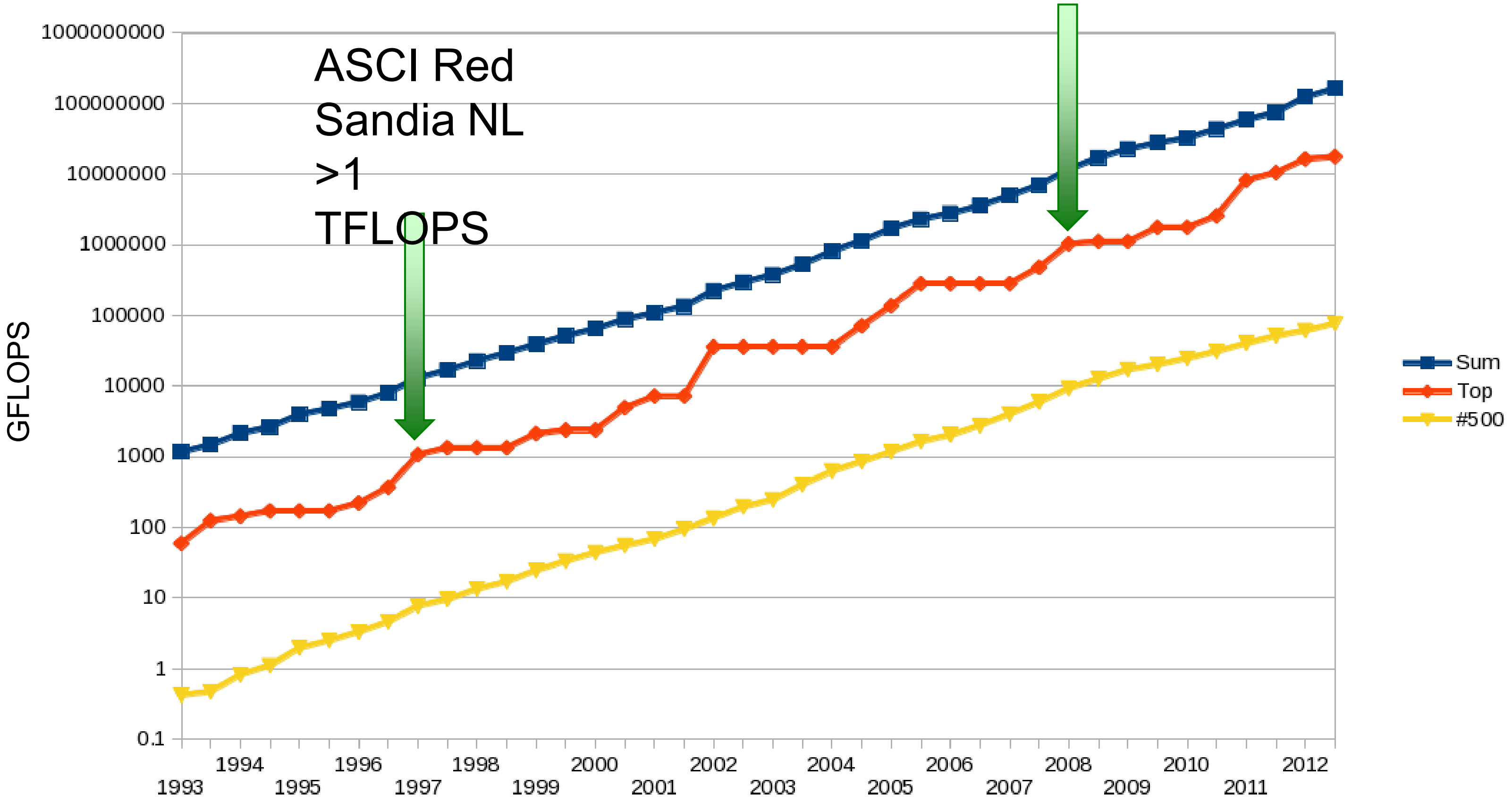


Talk Overview

- The DOE labs: Who we are and what we do
- DOE supercomputers and their capabilities
- Present challenges, including multicore/manycore

Breaking Speed Barriers

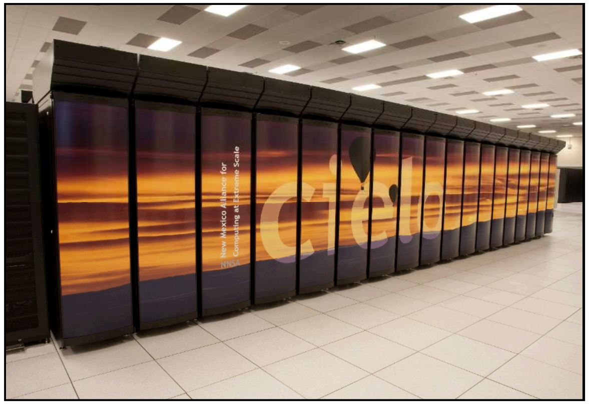
Roadrunner
Los Alamos NL
>1 PFLOPS

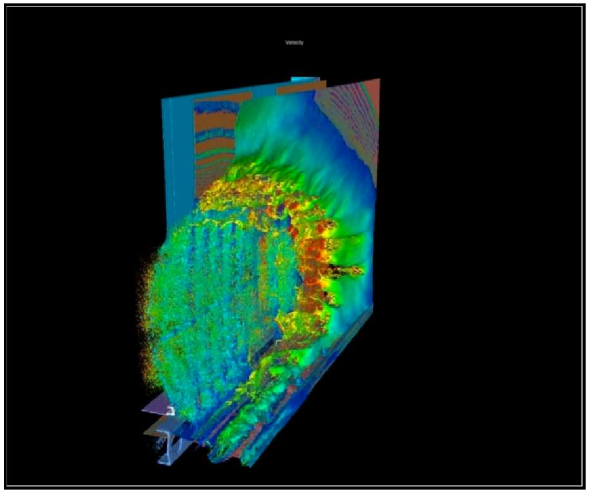


ASCI Red
Sandia NL
>1
TFLOPS

Cielo (LANL/SNL) by Numbers

Operational Time Frame	2011
Theoretical Peak Performance	1,374 TF
HPL (Linpack) Performance	1,110 TF using 142,272 cores
Cabinets	96
# Compute Nodes	8,944
# Compute Cores	143,104
Compute Processor	Dual AMD Opteron™ 6136 eight-core “Magny-Cours” Socket G34 @ 2.4 GHz
Compute Memory	286 TB DDR3 @ 1333 MHz
Compute Memory BW	763 TB/s
Service Nodes	272 AMD Opteron™ 2427 six-core “Istanbul” Socket F @ 2.2 GHz
User Disk Storage	7.6 PB User Available Capacity
Parallel File System	Cray DVS and Panasas PanFS
Parallel File System BW	~160 GB/s
High Speed Interconnect	Cray Gemini 3D Torus in a 16 x 12 x 24 (XYZ) Topology
Bi-section BW	6.57 x 4.38 x 4.38 (XYZ) TB/s
System Foot Print	~3,000 sq ft including Storage
Power Requirement	3,980 KW running HPL
Operating System	Cray Linux Environment





A 32,768-core CTH simulation run on Cielo helps designers understand the response of structures under severe blast loading conditions

Cielo Hardware Architecture

- AMD Magny-Cours Node
  - Dual-socket AMD 6136 Processors
  - 2 x 8 = 16 total cores
  - 2.4 GHz core frequency
  - 32 GB of 1333 DDR3 memory
    - 64 GB for Visualization Nodes
  - 153.6 peak DP GFLOPs
  - 85.3 peak GB/s memory BW
- Gemini High-Speed Interconnect
  - 3D Torus topology
  - 16x12x24

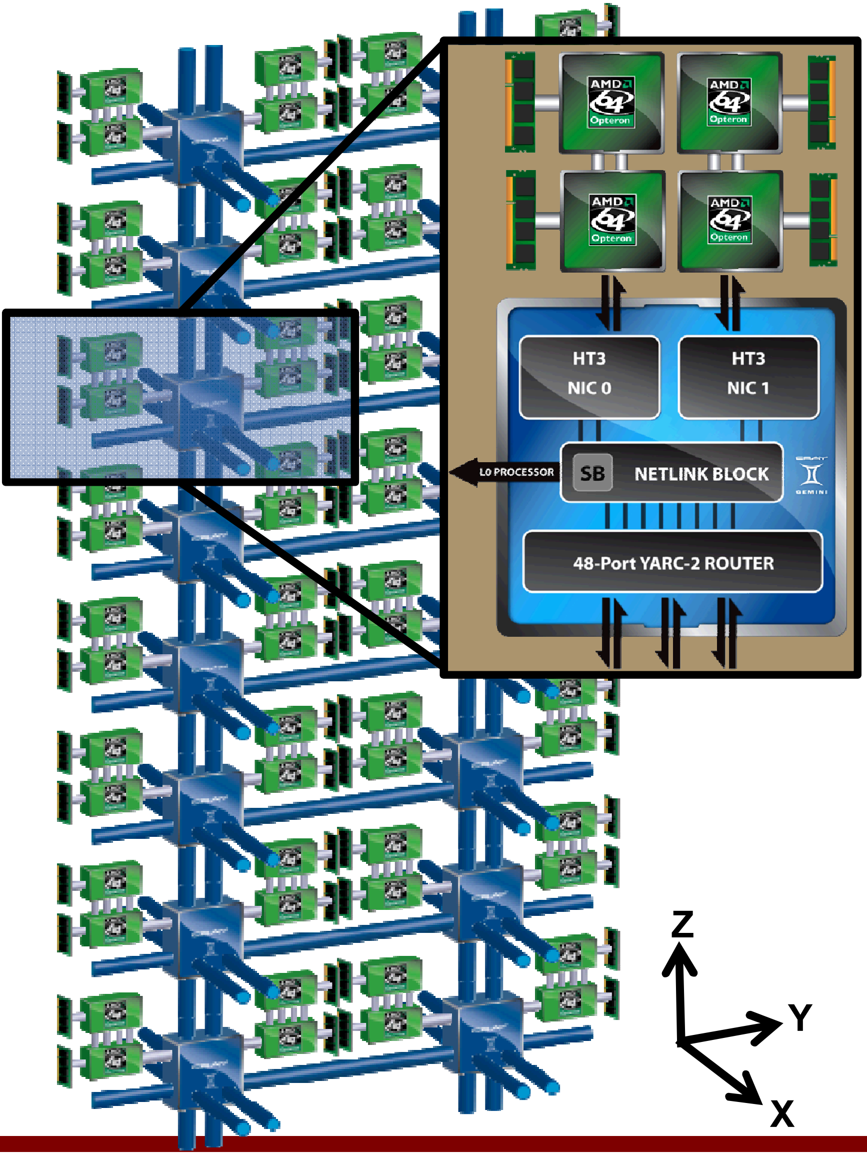
    X bisection: > 6.57 TB/s

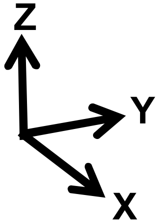
    Y bisection: > 4.38 TB/s

    Z bisection: > 4.38 TB/s
  - Node Injection

    > 6 GB/s/dir sustained BW

    > 8 MMsgs/sec sustained

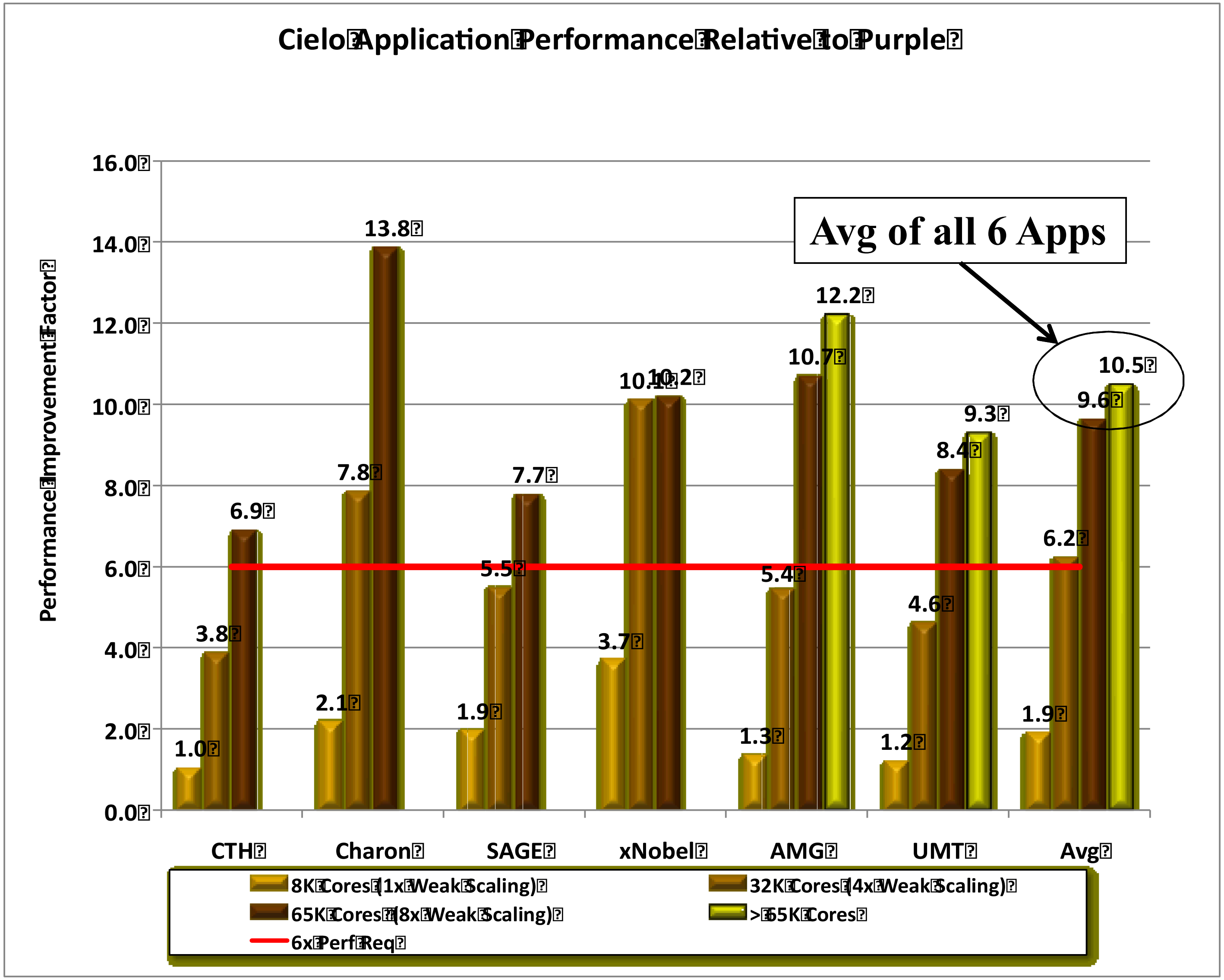




Acceptance Applications

Lab	Code	Fortran	Python	C	C++	MPI	OpenMP	Description
SNL	Charon			X	X	X		A transport reaction code to simulate the performance of semiconductor devices under irradiation
SNL	CTH	X		X		X		Explicit, multi-material shock hydrodynamics code
LANL	xNOBEL	X		X		X		Continuous Adaptive Mesh Refinement (CAMR) code: Hydrodynamics with adaption and high-explosive burn modeling
LANL	SAGE	X		X		X		Multi-dimensional multi-material Eulerian hydrodynamics code with adaptive mesh refinement.
LLNL	AMG2006			X		X	X	Algebraic Multi-Grid linear system solver for unstructured mesh physics packages
LLNL	UMT2006	X	X	X	X	X	X	Single physics package code. Unstructured-Mesh deterministic radiation Transport.

Capability Improvement Summary

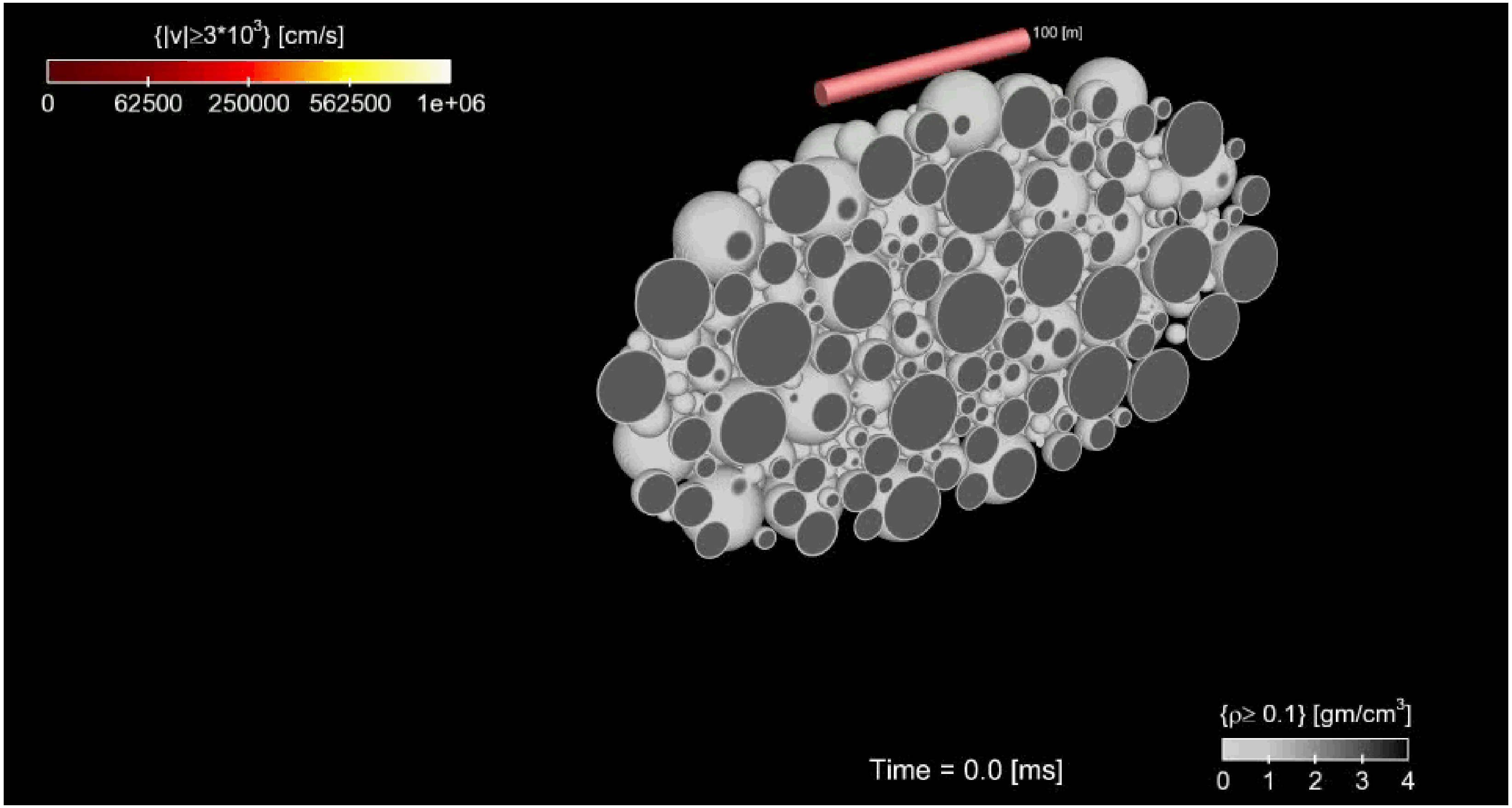

Cielo Application Performance Relative to Purple

Avg of all 6 Apps

3D RAGE simulations on the Cielo supercomputer to simulate a 1Mton surface explosion on Asteroid 25143 Itokawa

- Significant public interest in this topic
  - Several Hollywood movies; interest from government; popular articles
  - We use the shape of the Asteroid Itokawa, which is not a near-Earth hazard, simply to have a nonspherical geometry
- Many methods of Potentially Hazardous Objects (PHOs) mitigation have been proposed:
  - Nuclear options: Explosive disruption; stand-off momentum/velocity transfer
  - Non-nuclear methods: gravity attractors; solar energy absorption (paint) etc.
- For this simulation we use the RAGE hydrocode in 3D with a 1 Mton energy source on the surface of the object
- Here we use realistic (nonspherical) shapes and explore a "rubble piles" composition, i.e., where the asteroid has experienced many disruptive interactions, recombined, and is composed of many smaller "rocks."
- This simulation currently has run only to 25 ms; interesting mitigation occurs after ~5 s for this explosion energy. This is still running on Cielo.
- Never before: Cielo is the first computer big enough to run this problem in 3D.

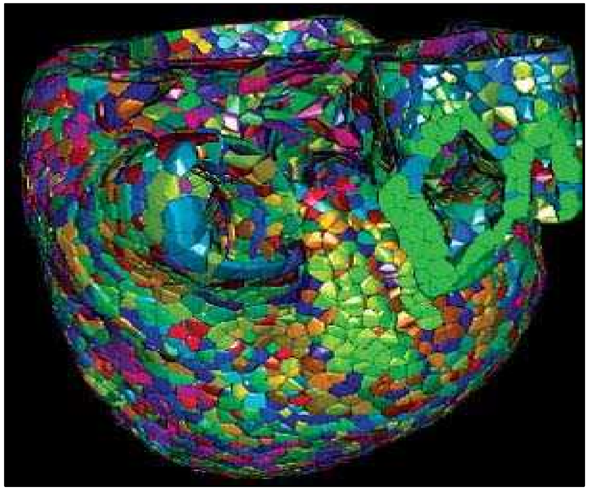
Cool, a Movie!



(Now on YouTube: http://youtu.be/hOcNbAV6SiI)

Sequoia (LLNL) by Numbers

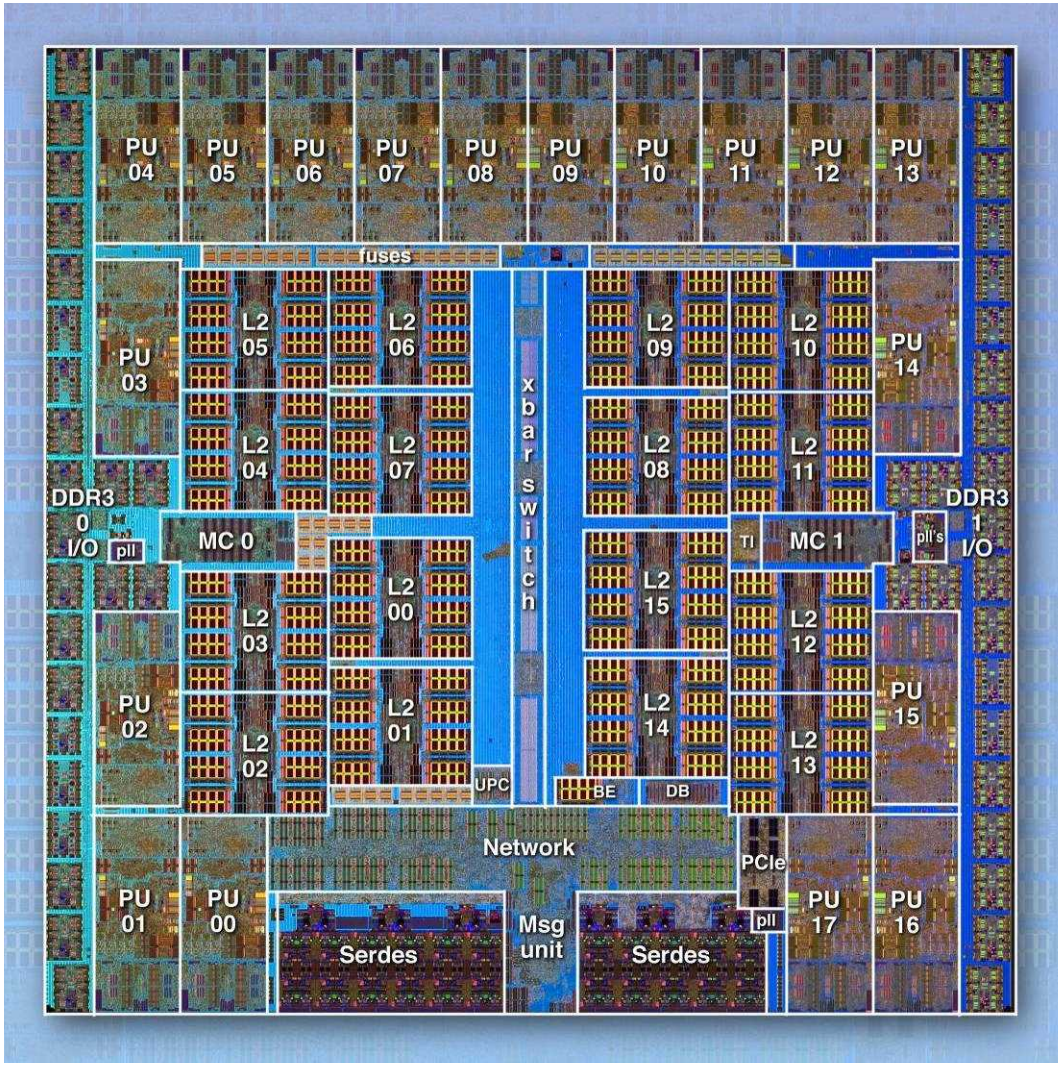
Operational Time Frame	2013
Theoretical Peak Performance	20 PF
HPL (Linpack) Performance	16.3 PF using 96 racks
Cabinets	96
# Compute Nodes	98,304
# Compute Cores	1,572,864 (with 4 threads/core)
Compute Processor	IBM BlueGene/Q Chip: 16 (user-accessible) cores @ 1.6 GHz
Compute Memory	1.5TB DDR3 @ 1333 MHz
Compute Memory BW	4 PB/s
Service Nodes	768 Dual eight-core Intel “Sandybridge” @ 2.6 GHz
User Disk Storage	50 PB User Available Capacity
Parallel File System	Lustre/Netapp
Parallel File System BW	512 GB/s
High Speed Interconnect	5D Torus and Tree-Structured Collective Network
Bi-section BW	60 TB/s
System Foot Print	~4,000 sq ft
Power Requirement	9.6 MW
Operating System	Lightweight Compute Node Kernel



The Cardioid electrophysiology tool run on Sequoia simulates heartbeats at a resolution of 0.05-0.1 mm to study drug-induced arrhythmia.

BlueGene/Q Compute chip

System-on-a-Chip design: integrates processors, memory and networking logic into a single chip

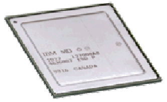


- **16 user + 1 service + 1 redundant cores**
  - Each 4-way multi-threaded, 1.6 GHz 64-bit PPC
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - Each has Quad FPU (4-wide double precision, SIMD)
  - Peak performance 204.8 GFLOPS @ 55 W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - Multiversioned cache, supports atomic ops
- **Dual memory controller**
  - 8–16 GB external DDR3 memory
  - 1.33 Gb/s
  - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
  - 5D Torus topology + external link
    → 5 x 2 + 1 high speed serial links
  - Each 2 GB/s send + 2 GB/s receive
  - DMA, remote put/get, collective operations
- **External (file) IO -- when used as IO chip**
  - PCIe Gen2 x8 interface (4 GB/s Tx + 4 GB/s Rx)
  - Re-uses 2 serial links
  - Interface to Ethernet or Infiniband cards

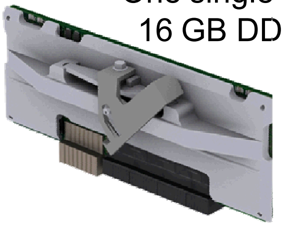
Sequoia: From Chip to System

1. Chip
16 cores

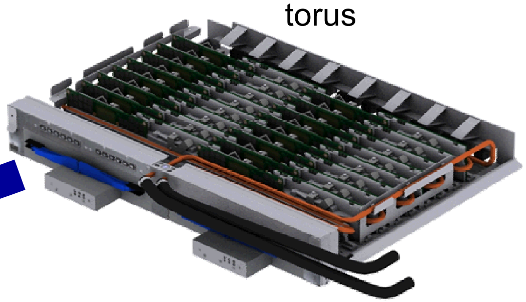
2. Module
Single chip



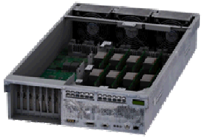
3. Compute card
One single chip module,
16 GB DDR3 memory



4. Node card
32 compute cards,
Optical modules, link chips,
torus



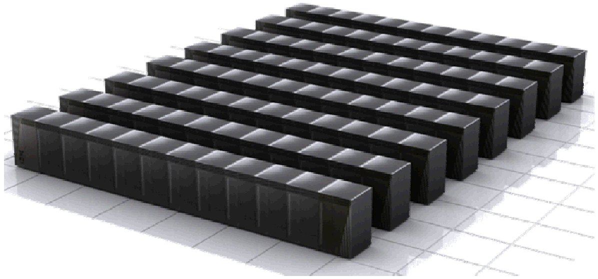
5b. I/O drawer
8 I/O cards
8 PCIe Gen2 slots



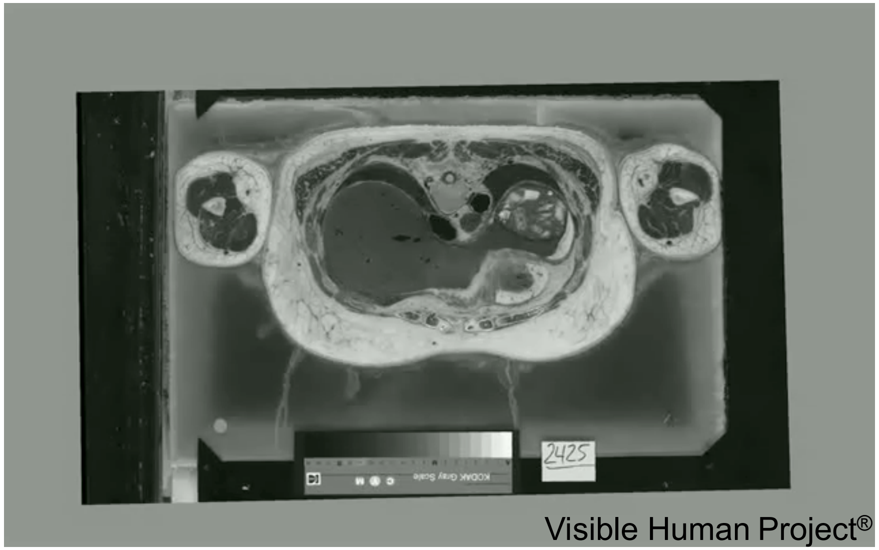
5a. Midplane
16 node cards

6. Rack
2 midplanes
1, 2, or 4 I/O drawers

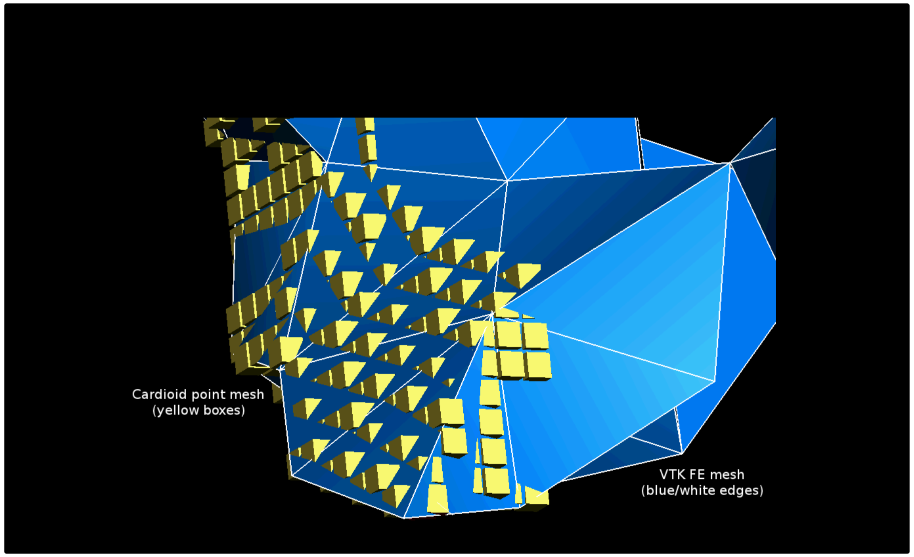
7. System
20 PF/s



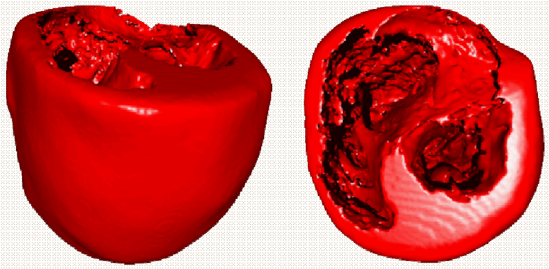
Computational Grid Constructed Using Actual Human Heart Data



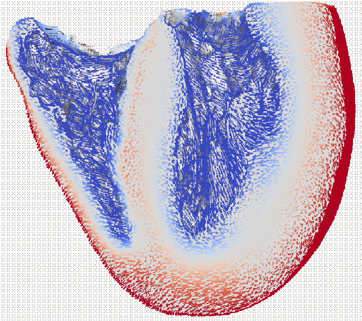
Visible Human Project®



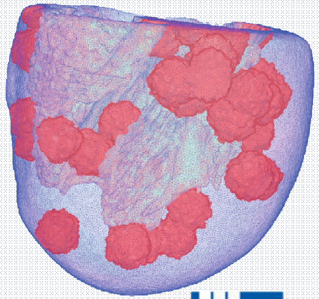
High resolution mesh



Generation of cardiac fibers



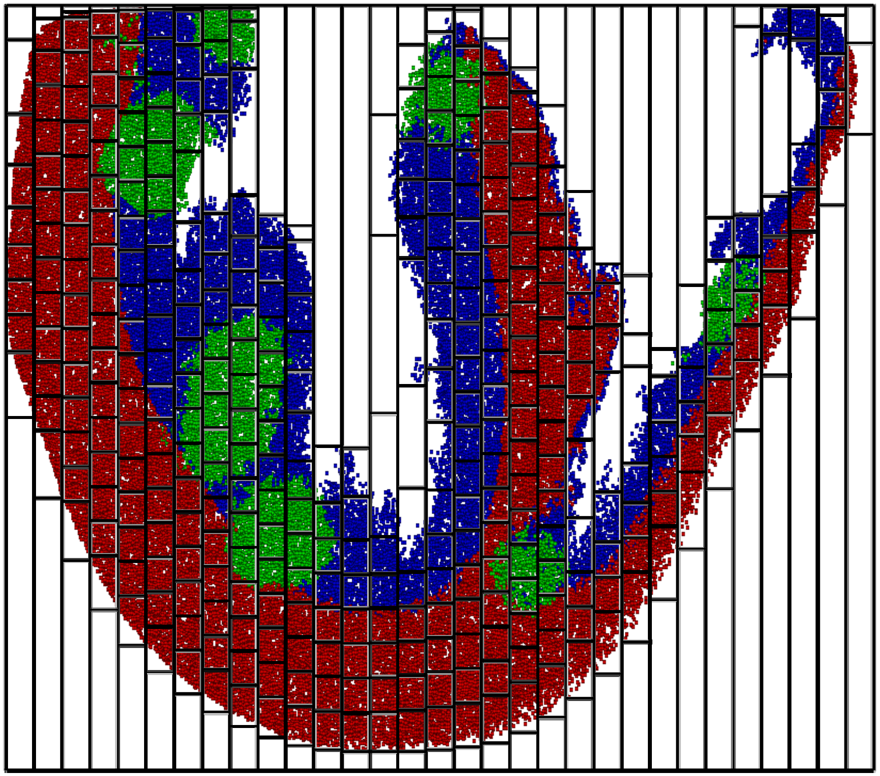
Introducing M-cell islands



Sequoia's Five Levels of Parallelism

- **Nodes**: MPI/SPI to communicate data between disjoint address spaces.
- **Cores**: Leverage shared address space across multiple cores using OpenMP or similar on-node parallelism model.
- **Threads**: Multiple threads per core cover latencies and may allow for better register usage, reduced pipeline stalls, etc.
- **SIMD**: Quad double precision SIMD units execute up to 8 floating point operations per cycle.
- **Functional Units**: Algorithms and threading must balance use of integer and floating point functional units.
- Also consider pipelines, prefetch, DMA, …

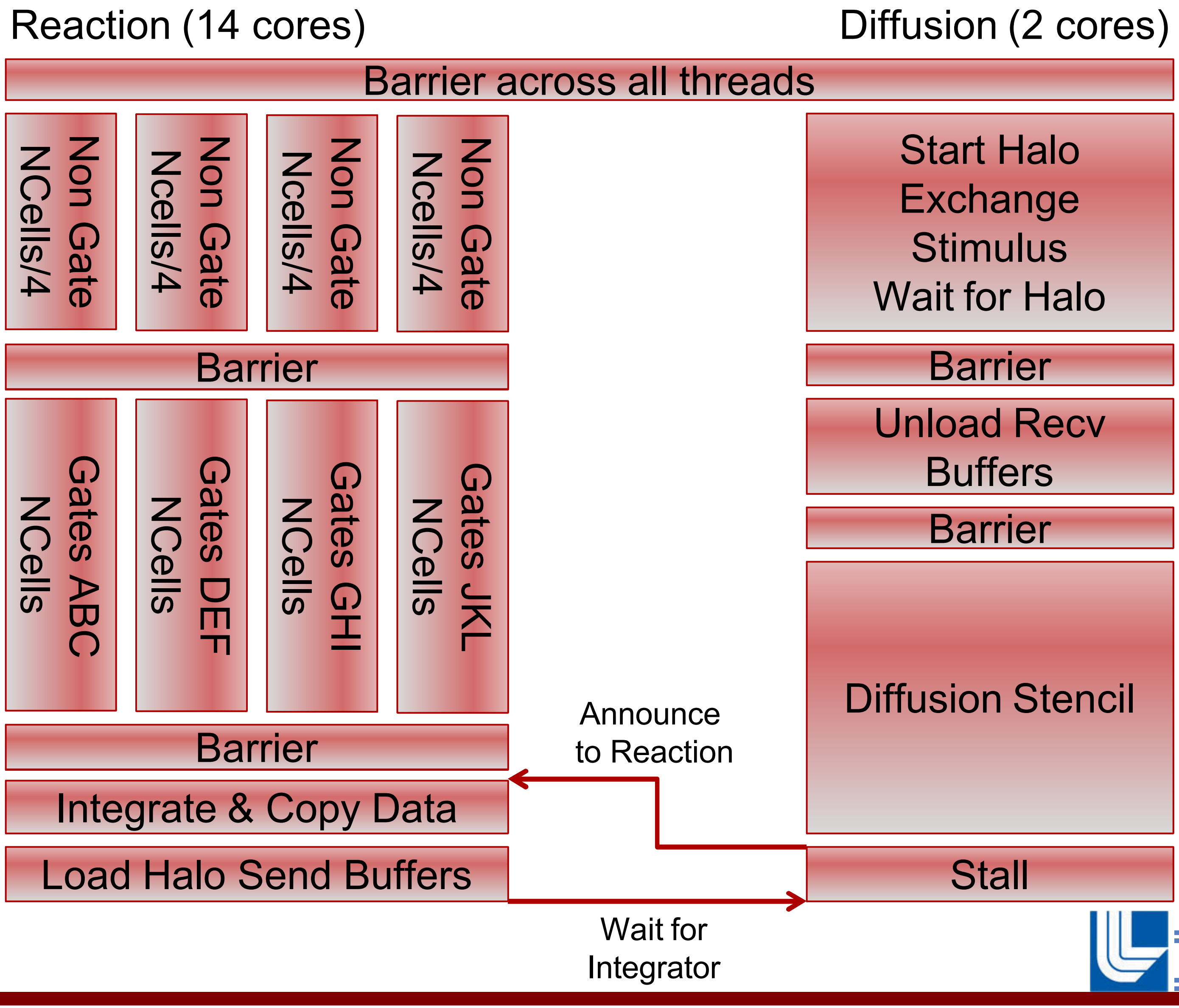
Load Balance for Complex Geometries



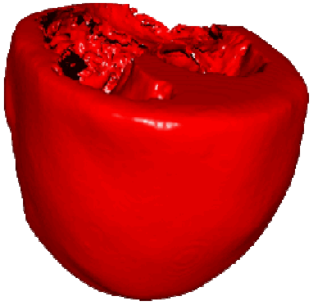
WorkBound Method:

- Choose dimensions of "basic block" that defines work load upper bound.
- Loop over xy-columns, adding planes to a given task until cost function nears or equals upper bound.
- Constraining the x- and y-dimensions to SIMD- and register-friendly sizes yields better time to solution.

Threading Time Line



Overheads of MPI and OpenMP

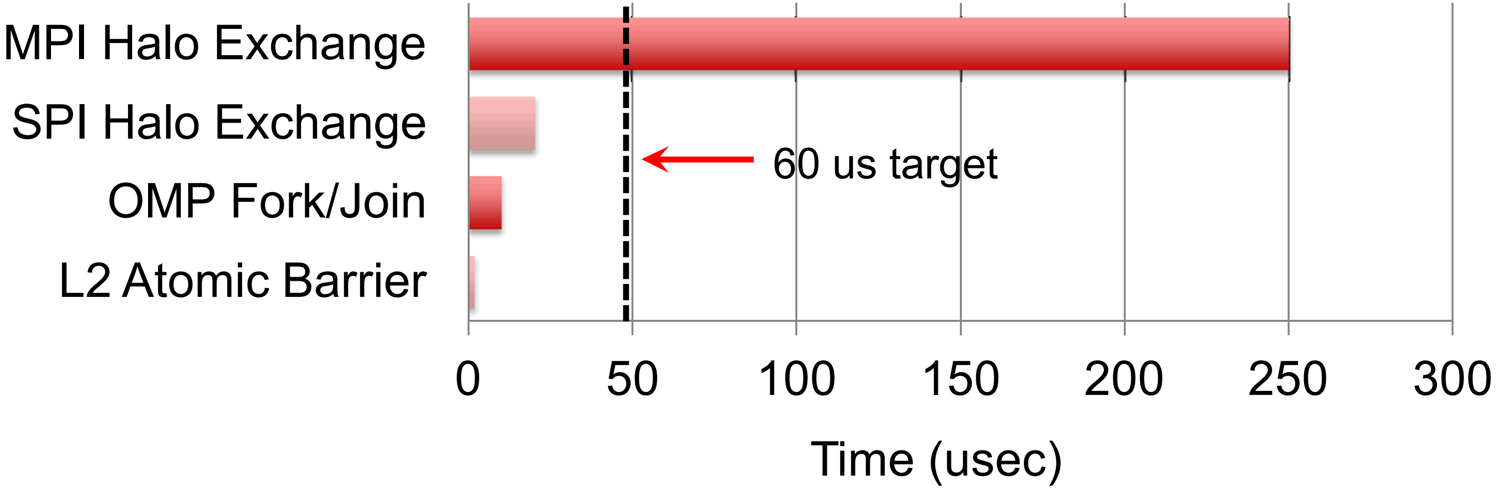


370 Million Cells

1.6 Million Cores

1600 Flops/cell

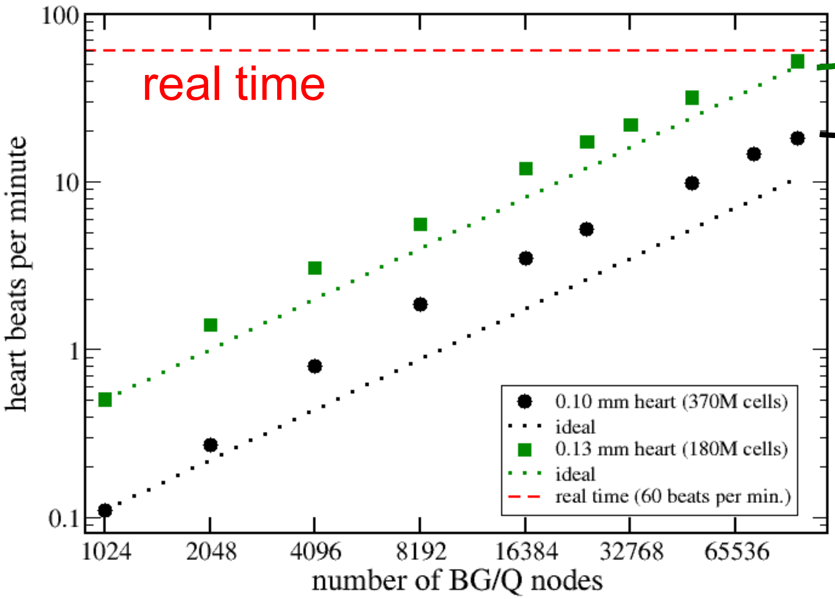
60 us per iteration



Peak Performance and Scalability

Measured peak performance: **11.84 PFlop/s (58.8% of peak)**

- 0.05 mm resolution heart (3B tissue cells)
- Ten million iterations, dt = 4 usec
- Performance of full simulation loop, including I/O, measured with HPM.



60 beats in 67.2 seconds

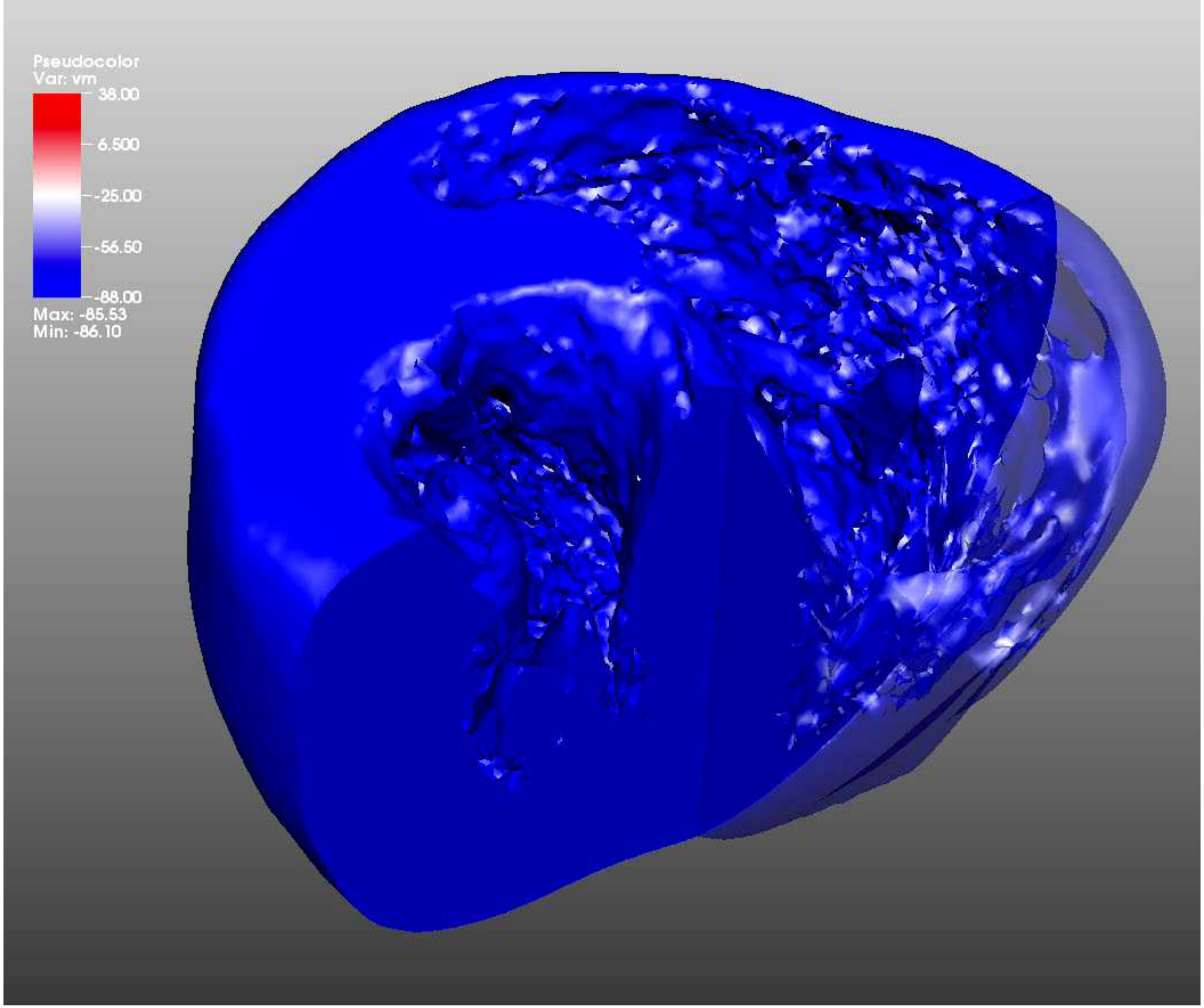
60 beats in 197.4 seconds

Extreme strong scaling limit:

0.10 mm: 236 tissue cells/core.

0.13 mm: 114 tissue cells/core.

Activation in the 0.1 mm Heart



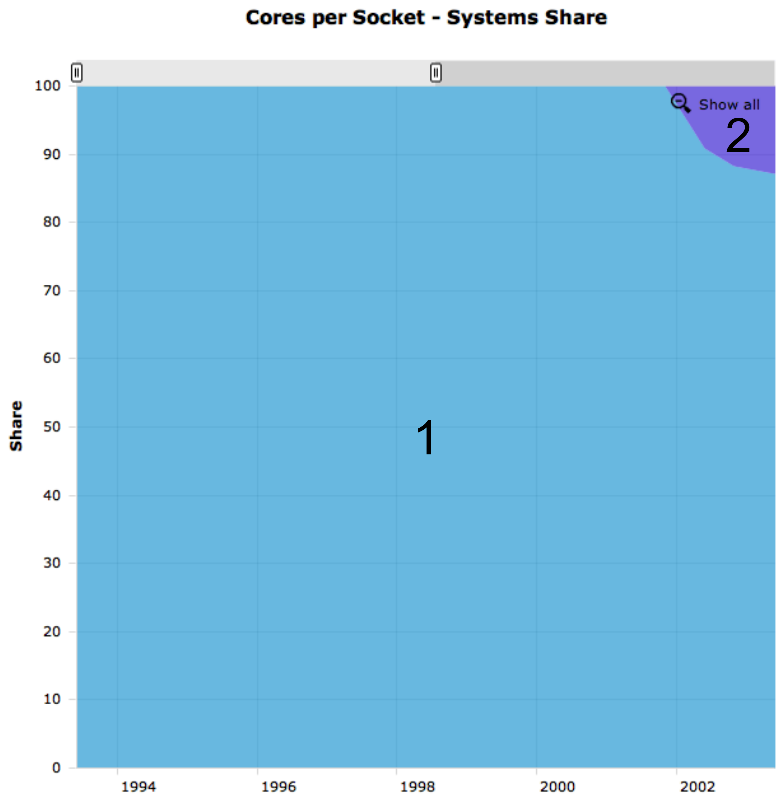
Talk Overview

- The DOE labs: Who we are and what we do
- DOE supercomputers and their capabilities
- Present challenges, including multicore/manycore

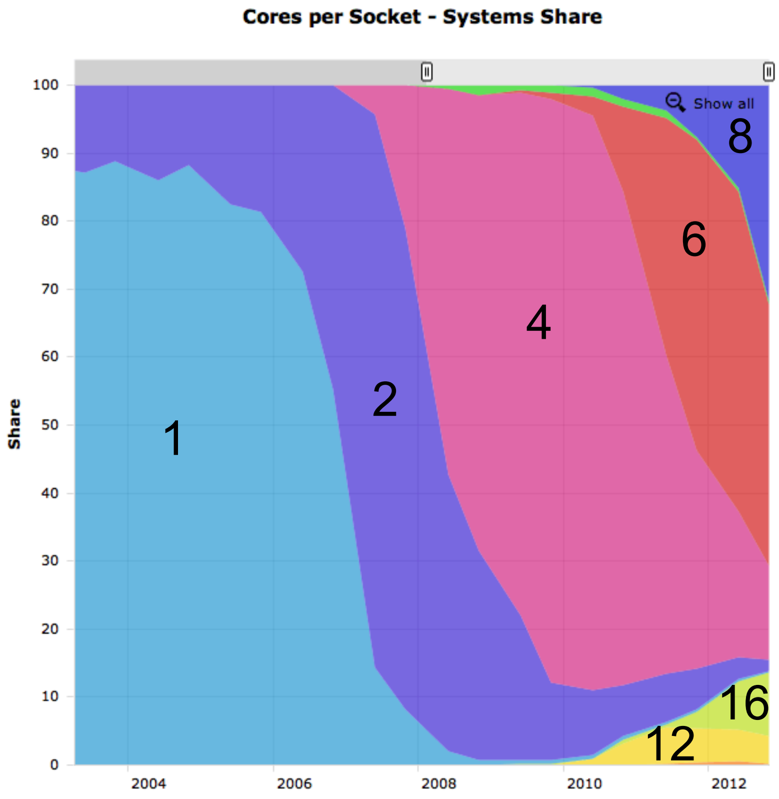
Challenges for Exascale

- Sequoia: 20 PetaFlops Max performance
  - ~10MW of power
  - ~40km (25 miles) of cable
- Goal is to reach 1 ExaFlops (1000 PetaFlops) by end of decade
  - 2X-3X power is okay … 20X-50X is not!
    - Need efficiency improvements in every system component
  - Must tolerate very low Mean Time Between Failure (MTBF)
    - At any point in time, some part has failed
  - Node-level computational performance must improve
    - Manycore chips → More on-node parallelism
  - Key question: How will we program this thing?

Multicore in Supercomputers

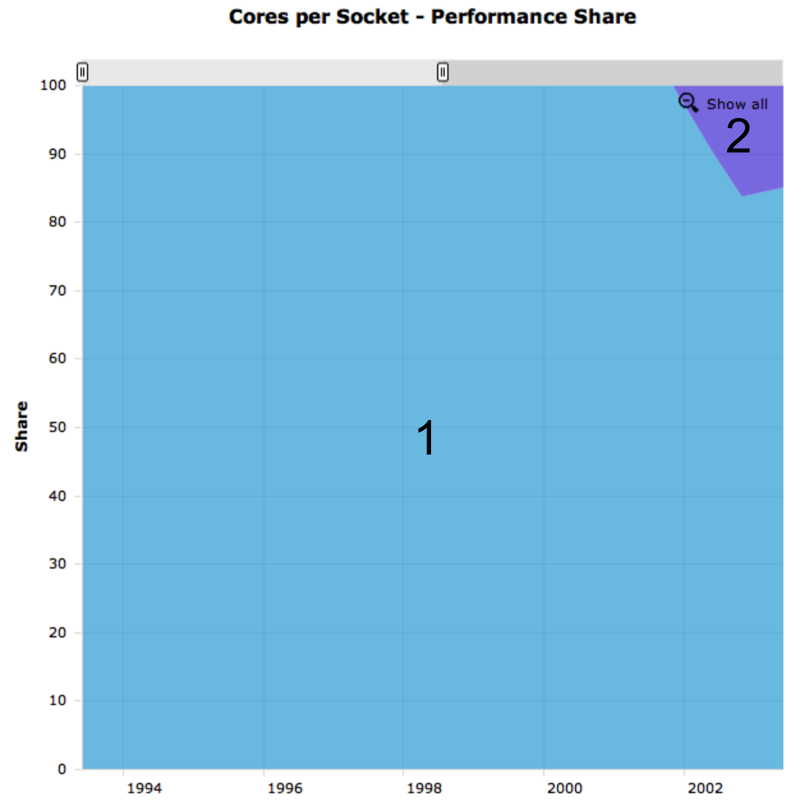


10 Years



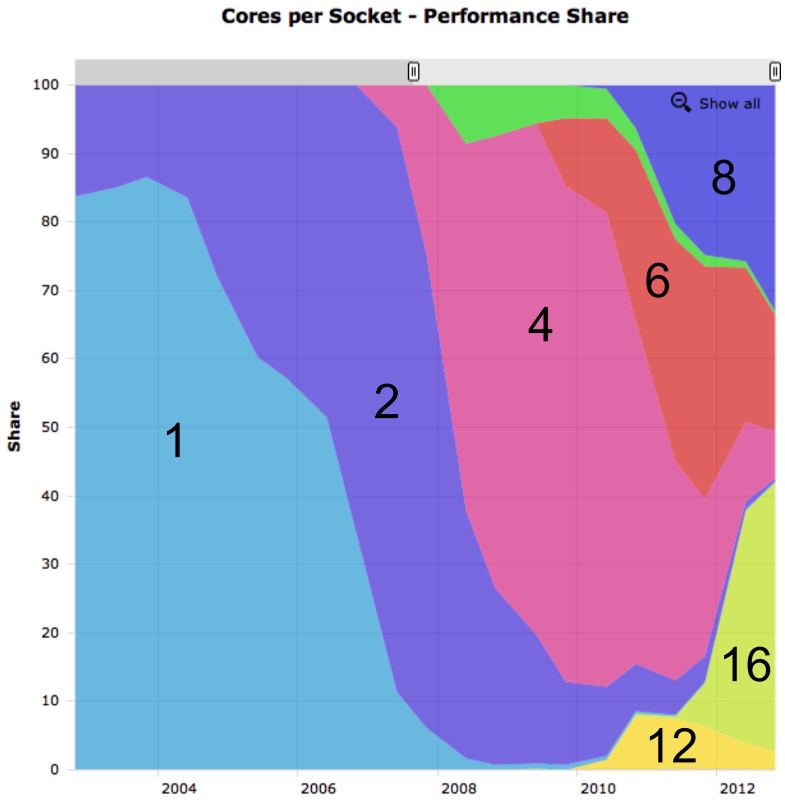
9.5 Years

Multicore in Supercomputers



Cores per Socket - Performance Share

10 Years



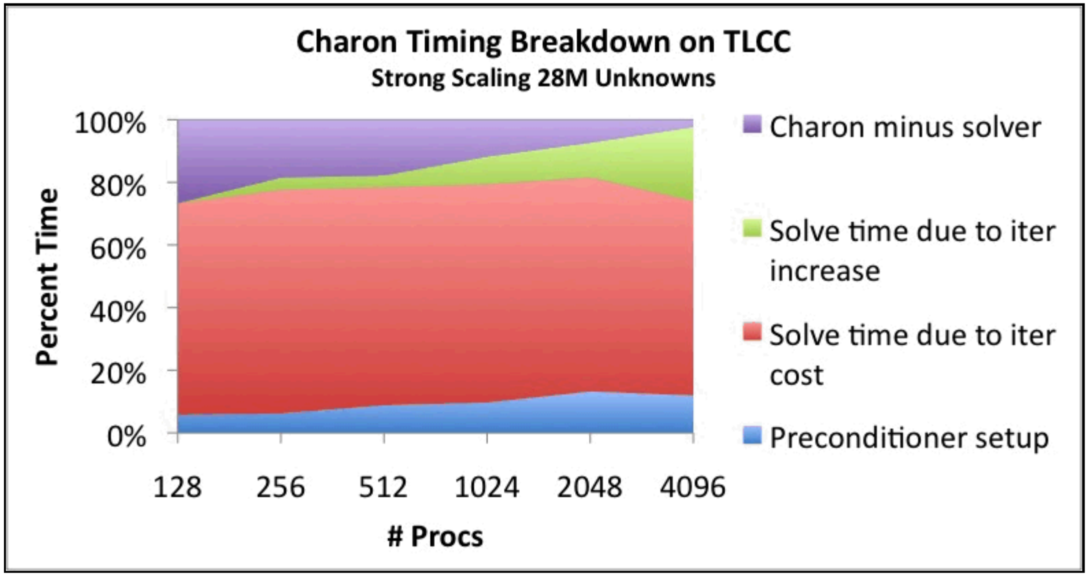
Cores per Socket - Performance Share

9.5 Years

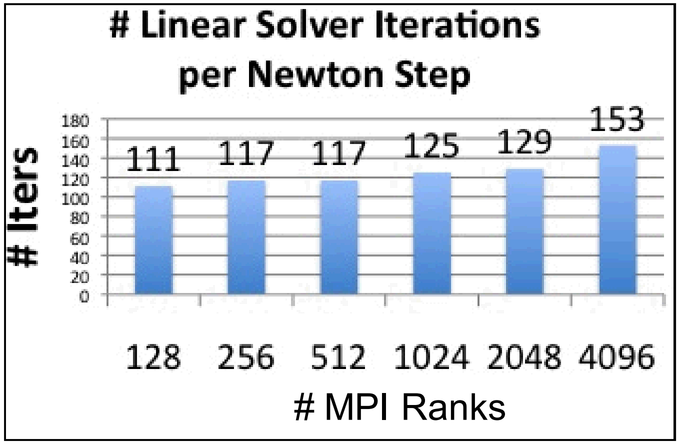
Coping with Multicore

- Many LANL/SNL apps still use message passing everywhere
  - One MPI process per core
- More use of shared memory programming at LLNL
  - Began when ASCI White (circa 2001) lacked the network bandwidth to support one MPI process per core
- Even with enough network bandwidth, algorithmic limitations of some applications require a shift to multithreading…

Scaling Preconditioners for Multicore



Strong scaling of Charon on TLCC (P. Lin, J. Shadid 2009)



- Observe: Iteration count increases with number of subdomains.
- With scalable threaded smoothers (LU, ILU, Gauss-Seidel):
  - Solve with fewer, larger subdomains.
  - Better kernel scaling (threads vs. MPI processes).
  - Better convergence, More robust.
- Exascale Potential: Tiled, pipelined implementation.
- Three efforts:
  - Level-scheduled triangular sweeps (ILU solve, Gauss-Seidel).
  - Decomposition by partitioning
  - Multithreaded direct factorization

MPI Tasks	Threads	Iterations
4096	1	153
2048	2	129
1024	4	125
512	8	117
256	16	117
128	32	111

Factors Impacting Performance of Multithreaded Sparse Triangular Solve, Michael M. Wolf and Michael A. Heroux and Erik G. Boman, VECPAR 2010.

The Programming Challenge

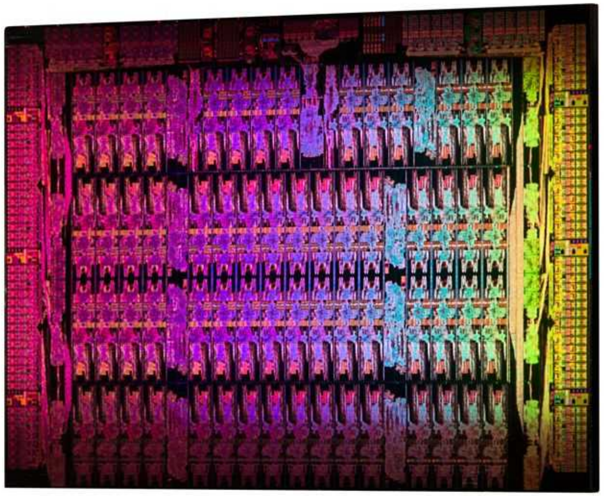
- Scientific codes often O(million) lines
- Diverse development teams
  - Physicists / Chemists
  - Mathematicians
  - Computer scientists
  - Project managers
- Many have a useful life of 30+ years
  - Compare to ~5 year useful life of a machine
- Want to transition between machines as painlessly as possible
  - Minimal porting effort desirable
  - Unlikely to commit resources for complete overhaul

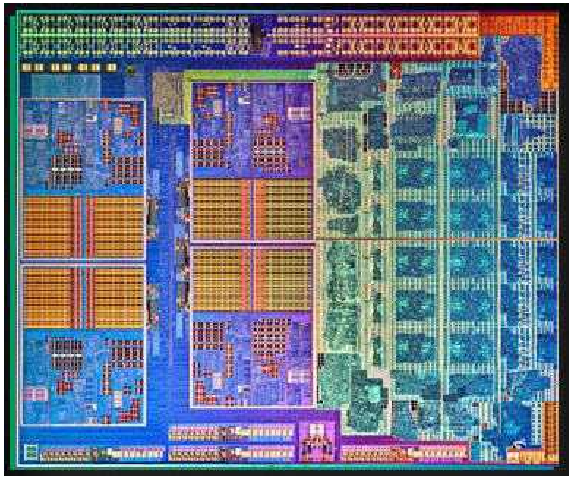
Exascale Testbeds

- Expectations: These systems are needed for exploratory R&D and pathfinding explorations of:
  - Alternative programming models
  - Architecture-aware algorithms
  - Low-energy runtime & system software
  - Advanced memory sub-system development
- It is more important to explore a diverse set of architectural alternatives than to push large scale.

Testbed Platforms

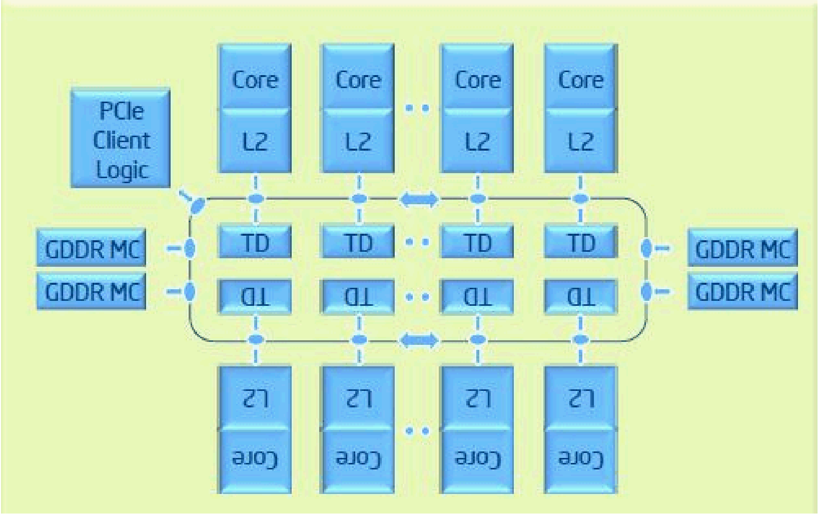
- Intel Many Integrated Core (MIC) / “Phi” co-processors
- AMD Fusion heterogeneous CPU/GPU
- Convey HC-1ex
- Tilera TILE-Gx36 processors





Intel MIC / "Phi" Architecture

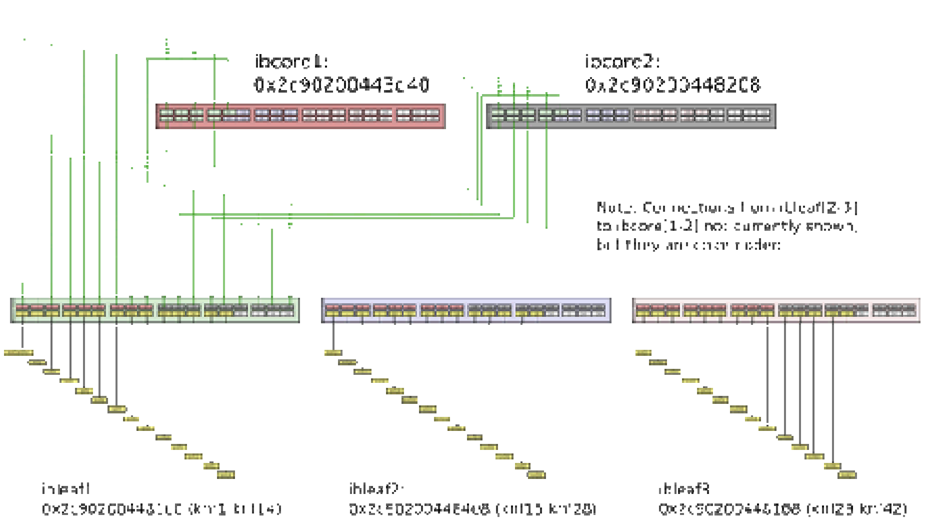
- 57 or 60 x86 cores on chip, each with 32KB L1 instruction cache, 32KB L1 data cache, 512KB L2 cache
- Each X86 core has a 512-bit vector unit, allowing 16 single precision or 8 double precision floating point ops / instruction
- Individual L2 data caches kept coherent by ring network (64B each way), also connected to the on-chip memory controllers
- 4 hardware threads / core, enough to stay busy while filling an L1 cache miss
- Implemented on a PCI card, with its own memory, connected to host memory via PCI using DMA operations



42-Node Testbed (84 MICs)

Arthur Many Integrated Cores (MIC) Exascale Test Machine

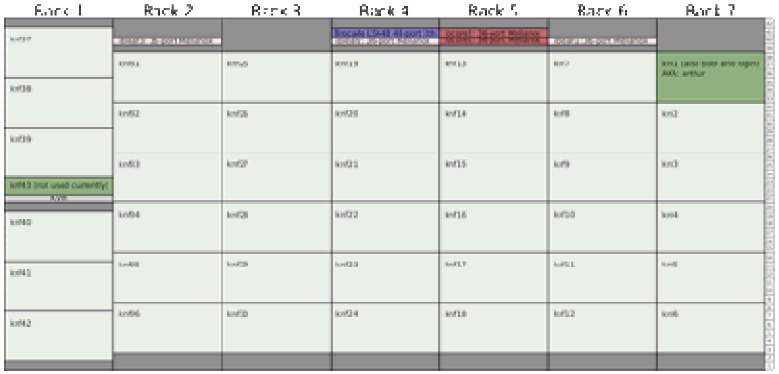
InfiniBand Diagram - QDR Fat Tree with 5 36-port Switches


ibcore1:
0x2c90200443c40
ibcore2:
0x2c90200448208
ibleaf1
0x2c90200443ec0 (kn1 kn14)
ibleaf2:
0x2c90200448408 (kn15 kn28)
ibleaf3
0x2c90200446108 (kn29 kn42)

Note: Connections from ibleaf[2-3] to ibcore[1-2] not currently shown, but they are connected.

Arthur Many Integrated Cores (MIC) Exascale Test Machine

Rack Diagram



Rack 1	Rack 2	Rack 3	Rack 4	Rack 5	Rack 6	Rack 7
knf37	[illegible]		[illegible] [illegible]	ibcore1 - 36-port Mellanox [illegible]	[illegible]	
knf38	knf31	knf25	knf19	knf13	kn7	kn1 (also boot and login) AKA arthur
knf39	knf32	knf26	knf20	knf14	kn8	kn2
knf43 (not used currently) kvm	knf33	knf27	knf21	knf15	kn9	kn3
knf40	knf34	knf28	knf22	knf16	kn10	kn4
knf41	knf35	knf29	knf23	knf17	kn11	kn5
knf42	knf36	knf30	knf24	knf18	kn12	kn6

Arthur Diagrams

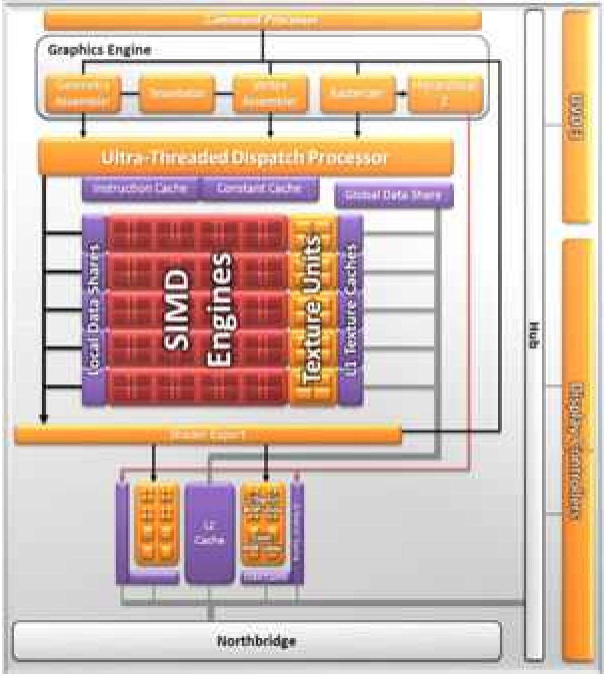
– **Matt Bohnsack**

Photographs

– **Victor Kuhns**

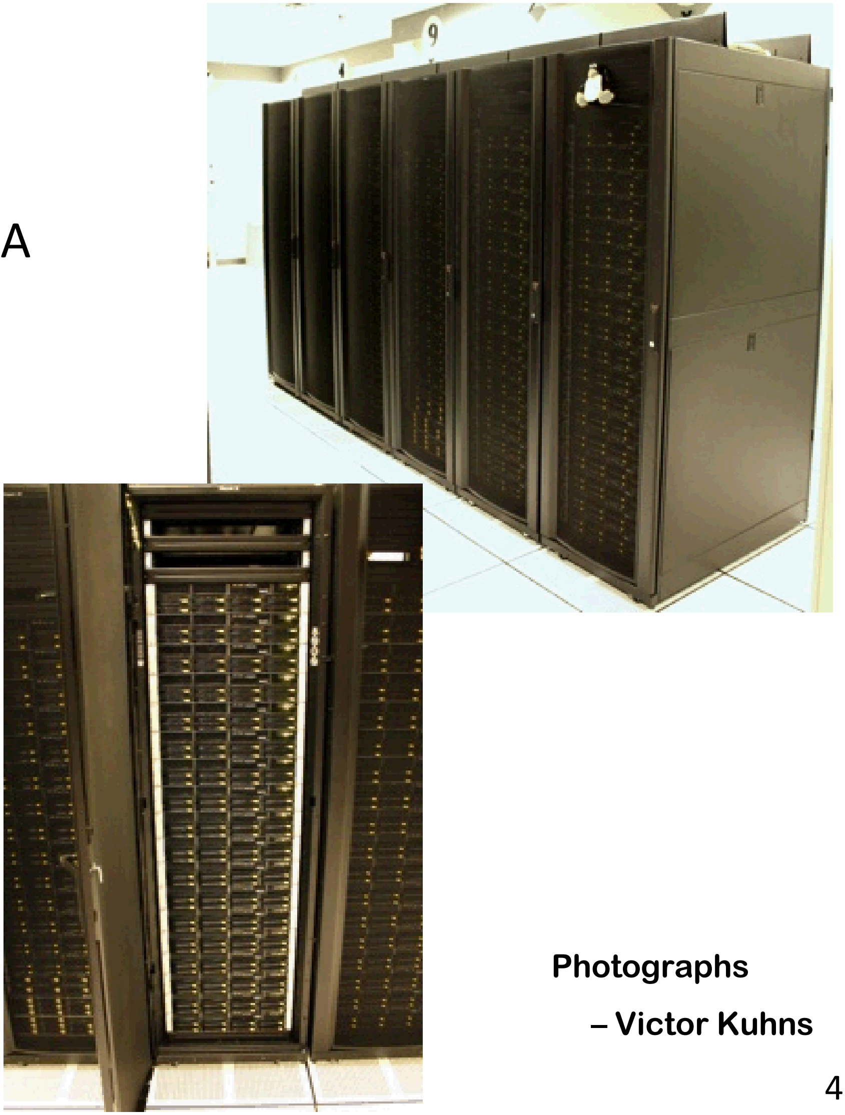
AMD Fusion CPU/GPU Architecture

- The Llano Fusion has four K10 x86 cores on chip, each with 64KB L1 instruction cache, 64KB L1 data cache, 1MB L2 cache
- The Radeon HD 6550D has 400 shader cores @ 600Mhz
  - 5 SIMDs
  - 20 Texture Units
  - 2 render backends
  - 32 Z/Stencil ROPs
  - 8 color ROPs
  - 600 MHz GPU clock rate



104-Node AMD Fusion Testbed

- Each Node has
  - One Llano Fusion APU
  - One 256GB Micron C400 SSD SATA 6Gb/s, MLC NAND Flash drive
- 100 nodes have 16GB DDR3-1600MHz
- 4 nodes have 8GB DDR3-1866MHz
- Interconnection Network: Qlogic QSFP QDR Infiniband

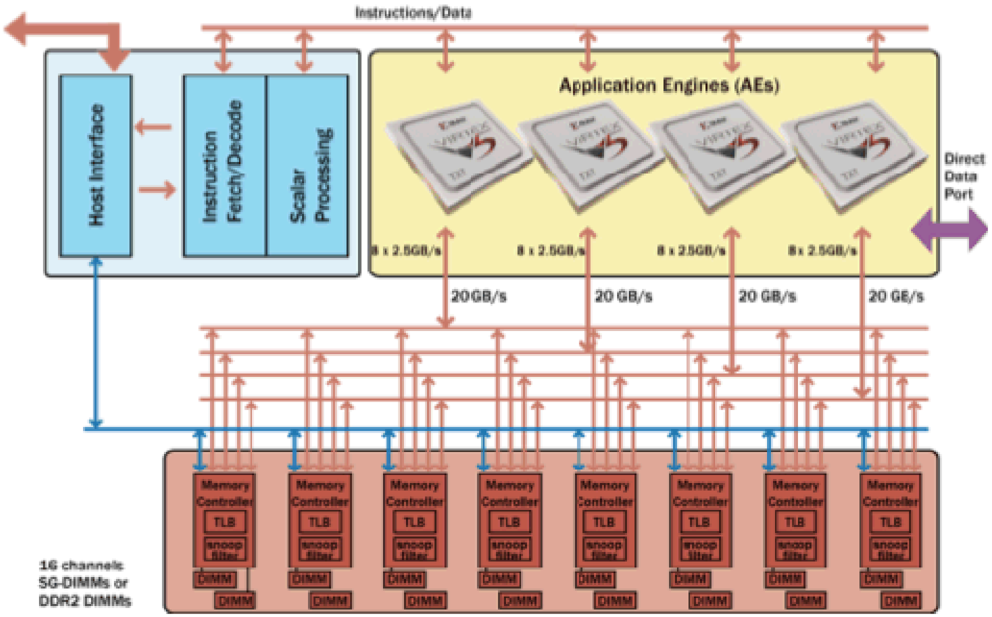


Photographs

– Victor Kuhns

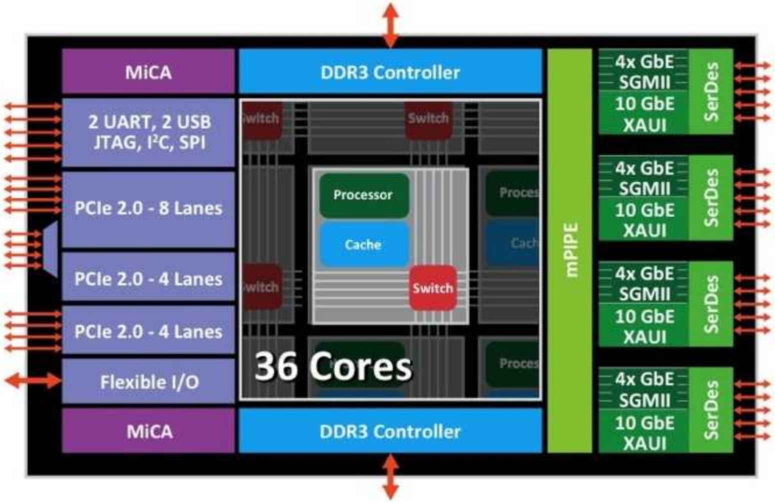
Convey Advanced Architecture

- Convey GV HC-1EX Board
  - Intel Nehalem Quad-core X86 @ 2.13GHz
  - 4 Xilinx Vertex6 LX760 FPGA Co-processor
  - 8 Xilinx FPGAs for programmable memory controllers that support 16 channels of Convey-designed Scatter-Gather DIMMs



Tilera Streaming Architecture

- Liberty Chassis – Four Tile-Gx8036 processor platform
- Each Tile-Gx8036 Processor has 36 cores @ 1.2GHz
  - 16GB DDR3-1333MHz, 9MB coherent L3 cache
  - 256 KB L2 cache/core
  - 32KB L1 instruction cache/core & 32KB L1 data cache/core
- Tilera's iMesh on-chip network



Testbed Experiments

- Run Mantevo proxy applications (http://mantevo.org)
  - Initial testing with miniFE, miniMD, and miniGhost
  - Evaluate programming models (e.g., OpenMP, OpenACC, OpenCL)
- Validation of SST Architectural simulation results (http://code.google.com/p/sst-simulator/)
- System Software R&D
  - Portals4 network stack
  - Kitten LWK OS
  - Qthreads multithreading runtime
  - Power management
  - I/O experiments with SSDs

Summary

- DOE National labs such as Sandia solve complex problems in the national interest.
  - Strategic scientific and technological capabilities
- Supercomputer simulations play a large role in the science and engineering work of the labs.
  - Scale has progressed from the first 1 Tflops machine in 1997 to a 20 Pflops machine today
- Reaching the goal of exascale computing will require significant innovation.
  - New advances in hardware, software, and algorithms

Acknowledgments

- Doug Doerfler, Cielo Chief Architect, SNL
- Mike Heroux, Distinguished Member of Technical Staff, SNL
- Richard Barrett, Principal Member of Technical Staff, SNL

- Bronis de Supinski, CTO of Livermore Computing, LLNL
- Dave Richards, Condensed Matter & Materials Group, LLNL