

Tools and Environments for Large-scale Systems

Jiayuan Meng
Performance Engineering Group
Argonne Leadership Computing Facility

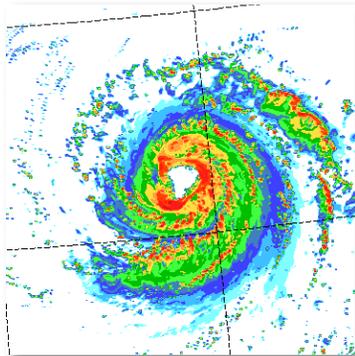


Argonne Leadership Computing Facility

- **One of two DOE national Leadership Computing Facilities**
- **To provide a leading-edge computing capability dedicated to breakthrough science and engineering**
- **To discover, develop, and deploy the computational and networking tools that enable researchers in the scientific disciplines to analyze, model, simulate, and predict complex phenomena important to DOE.**

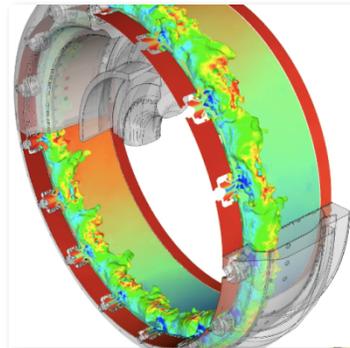


ALCF Scientific Accomplishments



Climate

Used leadership class, vortex-following calculation to more accurately predict hurricane track, to better mitigate risks.

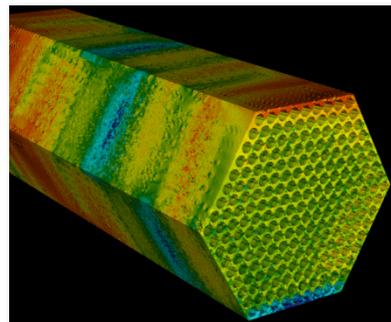


Gas Turbines

Two-phase flow and combustion modeling identified instability mechanisms that reduce efficiency, leading to design of more efficient aircraft engines.

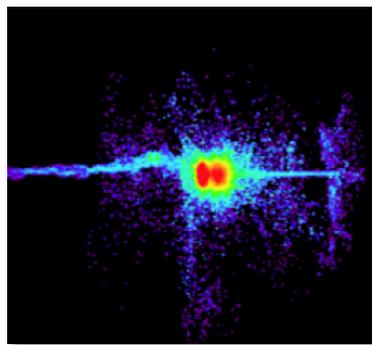
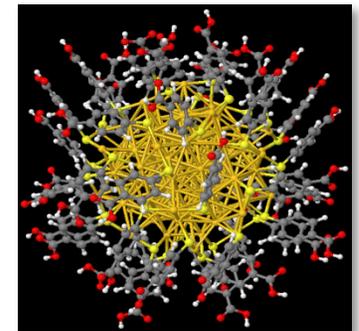
Nuclear Energy

High-fidelity fluid flow and heat transfer simulation of next-generation reactor designs, aiming to reduce the need for costly experimental facilities.



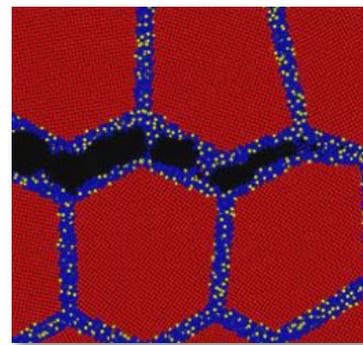
Nano Catalysts

Mapped out properties of a wide range of gold nanoparticles to design catalysts for fuel cells and methane conversion.



Fusion Energy

New hybrid algorithm allowed study of physics in Fast Ignition inertial confinement fusion over a much greater density range than planned.



Materials Science

Molecular dynamics simulation explained how a minute sulfur impurity embrittles nickel—relevant to next-generation nuclear reactor design.





Overview

- **Hardware Systems**
- **Allocations and Applications**
- **Tools and Libraries**

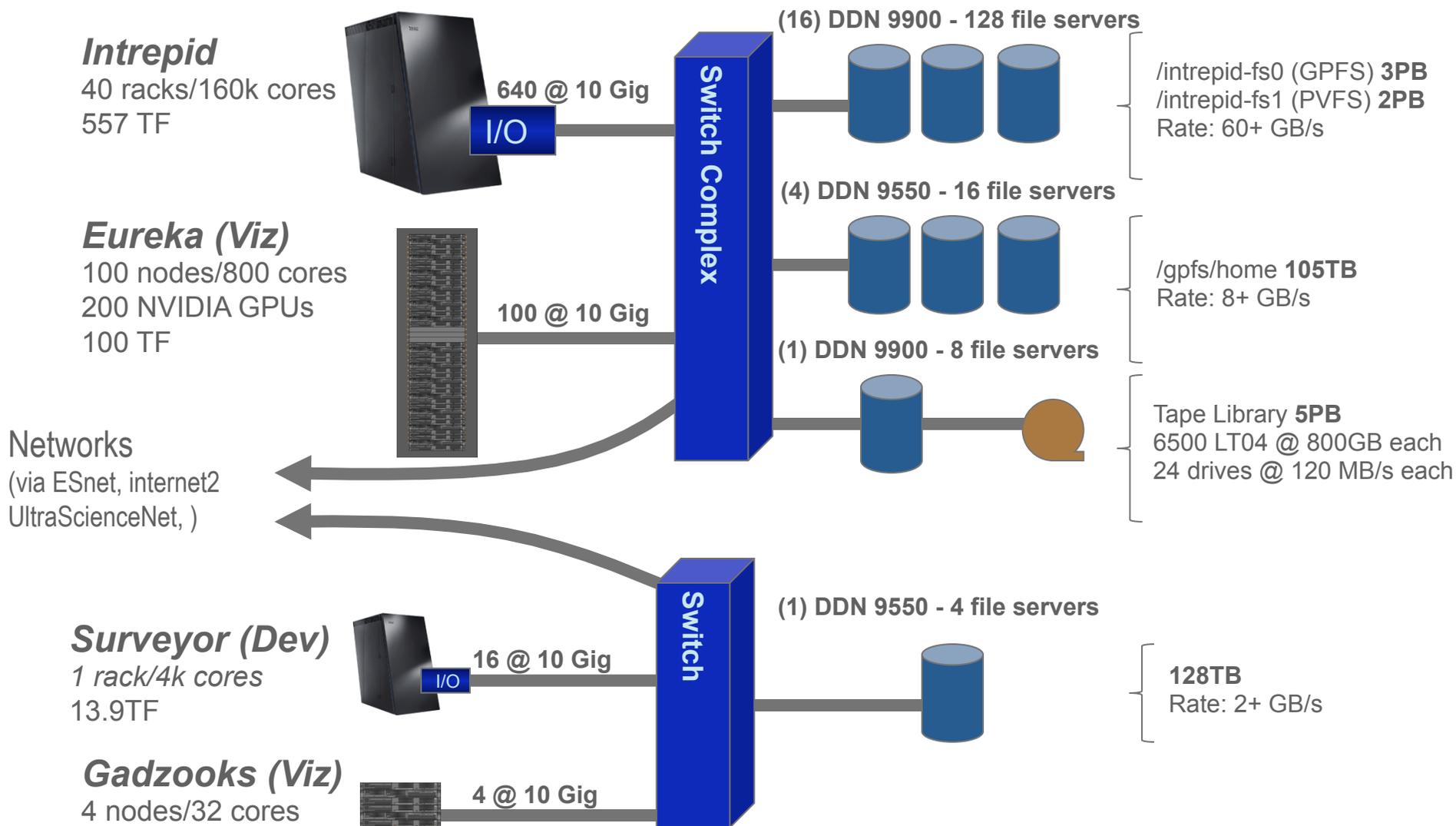


Hardware Systems

- 7 stories
- 25,000 ft² computing center
- 18,000 ft² library
- 10,000 ft² advanced digital laboratory
- 7,000 ft² conference center
- 30 conference rooms
- 3 computational labs
- 700 employees from 6 divisions



ALCF Resources - Overview



ALCF-2: BG/Q System

Mira: A 10PF Computational Science Platform



ALCF-2: BG/Q System

Mira: A 10PF Computational Science Platform

Mira



BG/Q Compute

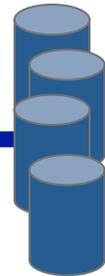
Mira: Latin: to wonder at, wonderful; causing one to smile



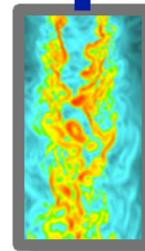
BG/Q IO



IB Switch



Data Storage



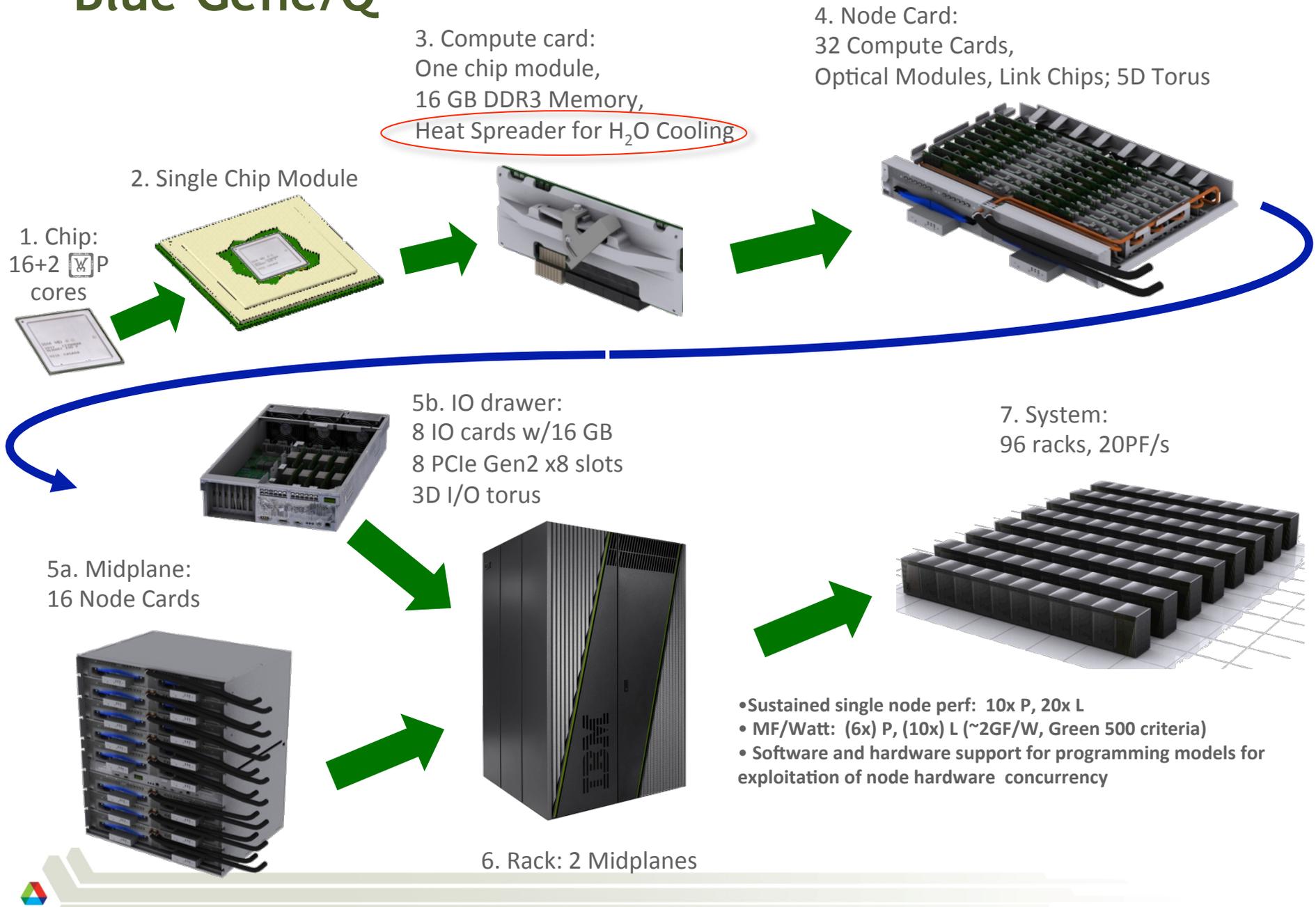
Viz & Data Analytics

Configuration

- BG/Q
 - 48 racks, 10 PF Peak Flop Rate
 - 48K 1.6 GHz nodes
 - 768K cores & 786TB RAM
 - 384 I/O nodes
- Storage
 - 240 GB/s, 35 PB



Blue Gene/Q



Preparing for Mira - Chilled Water Plant

U.S. DEPARTMENT OF ENERGY
Office of Science

Argonne NATIONAL LABORATORY UChicago Argonne

Site of Chilled Water Plant
Campus Map

Blue Gene Q Supercomputer

Centralized Chilled Water Plant (CCWP)

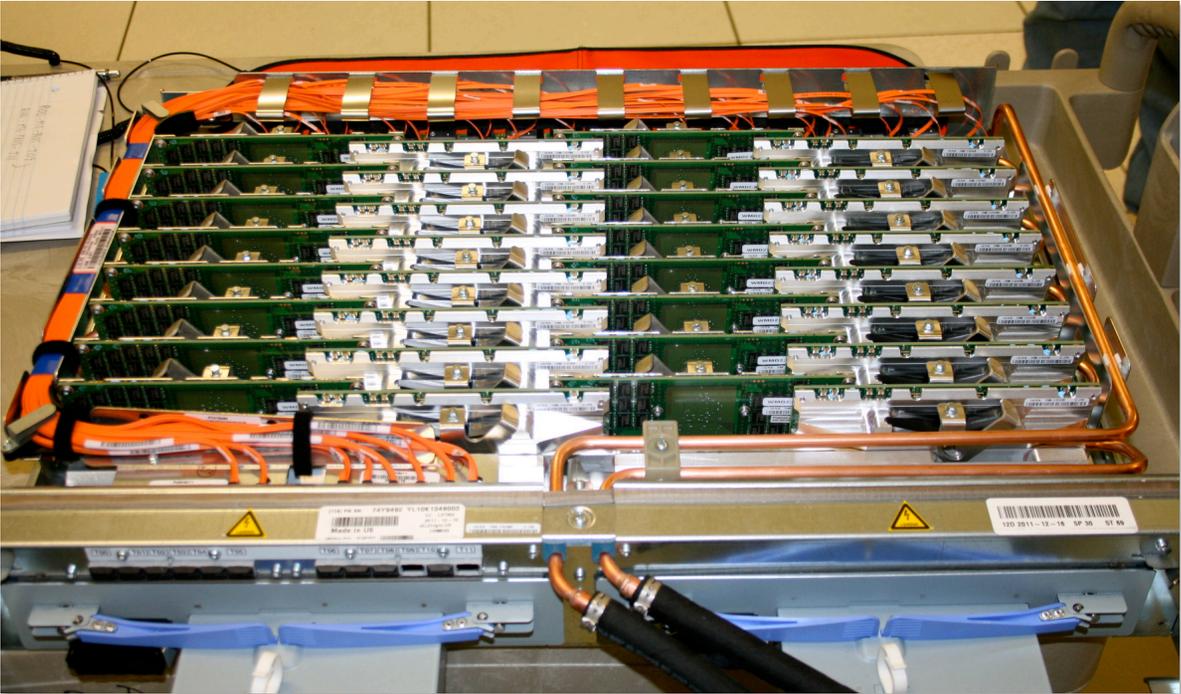
Argonne is completing the construction of a new centralized chilled water plant for the 200 area, west of the Theory and Computing Science (TCS) Center. In support of Argonne's campus modernization plan, the CCWP will provide necessary cooling for both current and future needs of the TCS (including IBM's next-generation Blue Gene Q supercomputer), and the laboratory's new Materials Design Laboratory, and Energy Sciences and Biosciences Buildings. Additionally, it will support existing 200-area building comfort cooling needs.

The CCWP initial capacity will be 2,600 tons of chilled water; its maximum future expanded capacity will support 17,000 tons of cooling. The facility has been designed to be Leadership in Energy and Environmental (LEED)-certified and will include sustainable features such as high-efficiency chillers and cooling towers, and free cooling and blended operations modes—which minimize energy use and maximize Mother Nature's cooling capabilities.

HDR

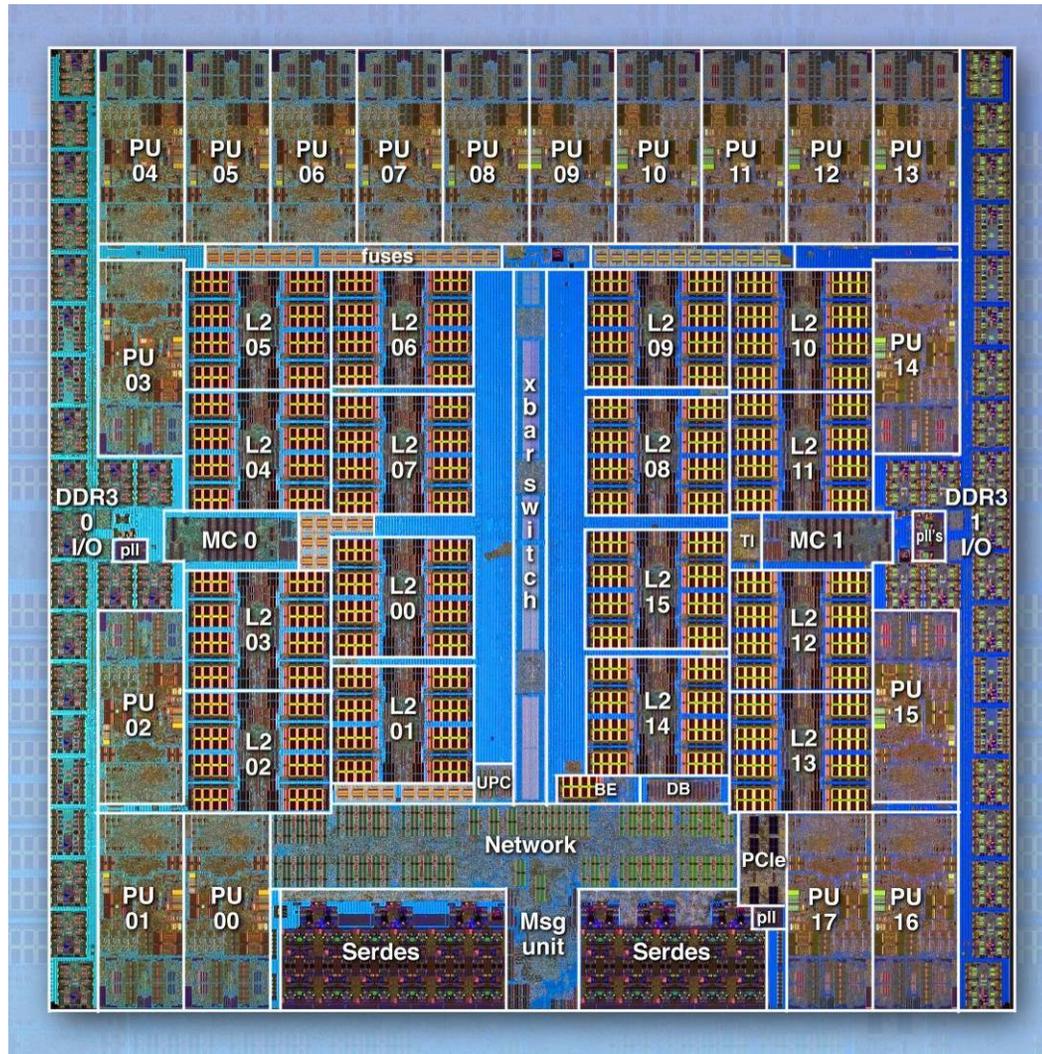


Water Cooling



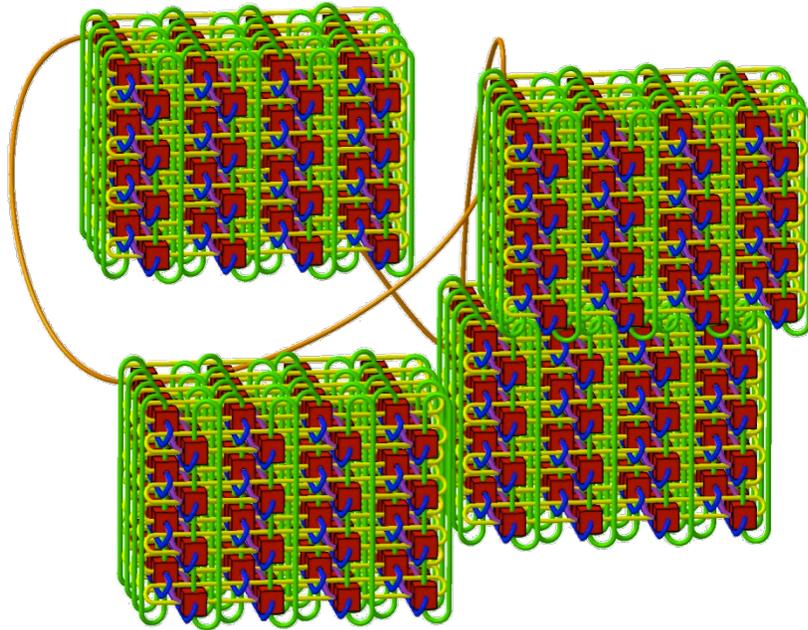
BlueGene/Q Compute chip

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



- **360 mm² Cu-45 technology (SOI)**
 - ~ 1.47 B transistors
- **16 user + 1 service processors**
 - plus 1 redundant processor
 - all processors are symmetric
 - each 4-way multi-threaded
 - 64 bits PowerISA™
 - 1.6 GHz
 - L1 I/D cache = 16kB/16kB
 - L1 prefetch engines
 - each processor has Quad FPU (4-wide double precision, SIMD)
 - peak performance 204.8 GFLOPS@55W
- **Central shared L2 cache: 32 MB**
 - eDRAM
 - multiversioned cache will support transactional memory, speculative execution.
 - supports atomic ops
- **Dual memory controller**
 - 16 GB external DDR3 memory
 - 42.6 GB/s
 - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
 - Router logic integrated into BQC chip.
- **External IO**
 - PCIe Gen2 interface

Inter-Processor Communication



Network Performance

- All-to-all: 97% of peak
- Bisection: > 93% of peak
- Nearest-neighbor: 98% of peak
- Collective: FP reductions at 94.6% of peak

- **Integrated 5D torus**
 - Virtual Cut-Through routing
 - Hardware assists for collective & barrier functions
 - FP addition support in network
 - RDMA
 - Integrated on-chip Message Unit
- **2 GB/s raw bandwidth on all 10 links**
 - each direction -- i.e. 4 GB/s bidi
 - 1.8 GB/s user bandwidth
 - protocol overhead
- **5D nearest neighbor exchange measured at 1.76 GB/s per link (98% efficiency)**
- **Hardware latency**
 - Nearest: 80ns
 - Farthest: 3us
(96-rack 20PF system, 31 hops)
- **Additional 11th link for communication to IO nodes**
 - BQC chips in separate enclosure
 - IO nodes run Linux, mount file system
 - IO nodes drive PCIe Gen2 x8 (4+4 GB/s)
↔ IB/10G Ethernet ↔ file system & world



Overview of BG/Q: Another step forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
Nodes / Rack	1,024	1,024	--
RAM / core (GB)	0.5	1	2x
Flops / Node (GF)	13.6	204.8	15x
Mem. BW/Node (GB/sec)	13.6	42.6	3x
Latency (MPI zero-length, nearest-neighbor node)	2.6 μ s	2.2 μ s	~15% less
Bisection BW (32 racks)	1.39TB/s	13.1TB/s	9.42x
Network Interconnect	3D torus	5D torus	Smaller diameter
Concurrency / Rack	4,096	65,536	16x
GFlops/Watt	0.77	2.10	3x





Hints for Programming on Mira

- **Vectorization**
- **Hybrid Programming**
- **Multi-threading**



Hybrid Programming (OpenMP + MPI)

- Utilize the computation power
- Compared to P
 - Less memory per thread
 - Less bandwidth per thread
 - Less cache per thread



Bad



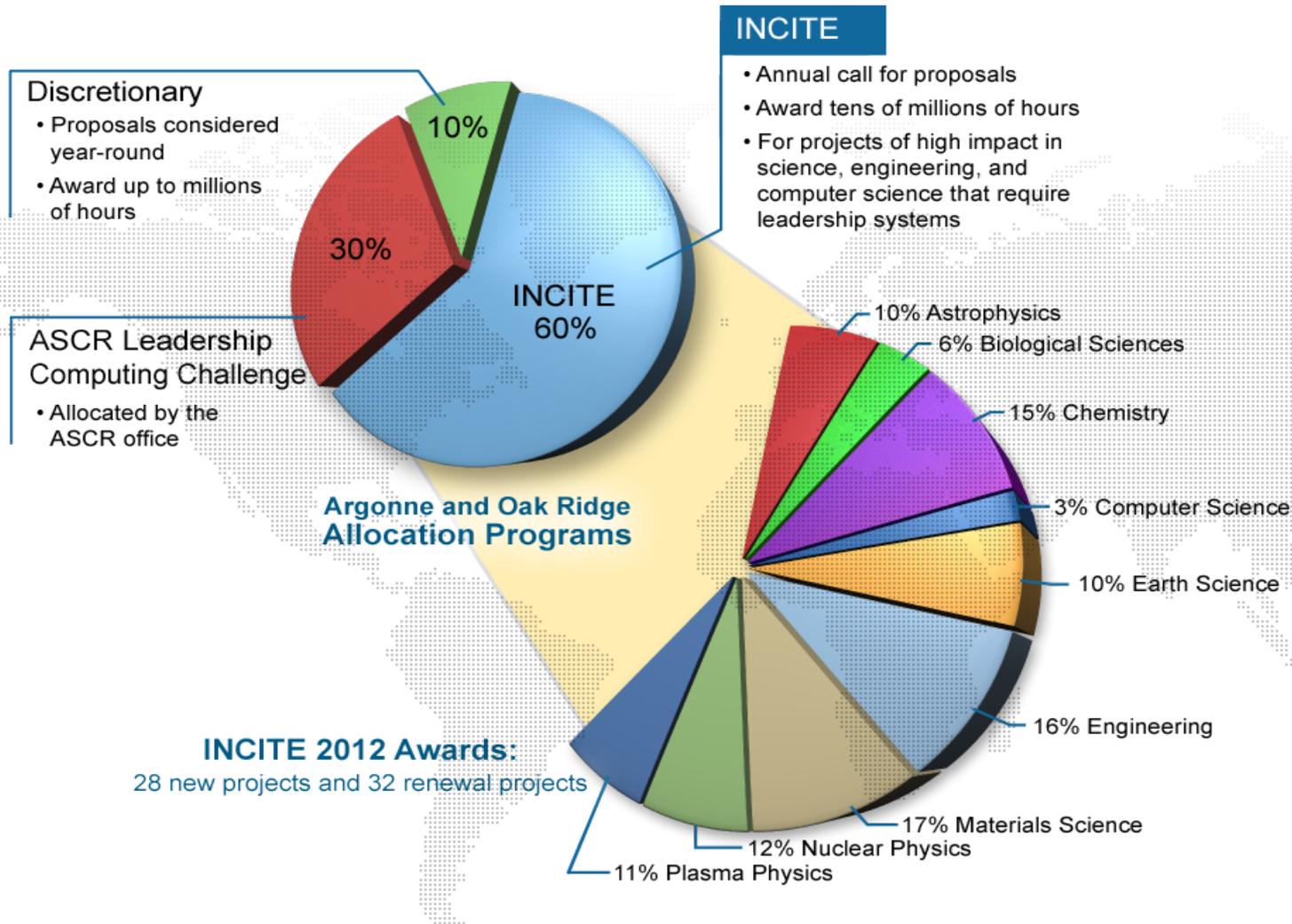
Good

Latency Hiding

- **Use multi-threading to hide latency caused by**
 - Instruction dependency
 - Memory accesses



Allocations and Applications





DOE INCITE Program

Innovative and Novel Computational Impact on Theory and Experiment

- Solicits large computationally intensive research projects
 - To enable high-impact scientific advances
 - <http://hpc.science.doe.gov/> (2012 call closes 6/30/2011)
- Open to all scientific researchers and organizations
 - Scientific Discipline Peer Review
 - Computational Readiness Review
- Provides large computer time & data storage allocations
 - To a small number of projects for 1-3 years
 - Academic, Federal Lab and Industry, with DOE or other support
- ***60% of time at Leadership Facilities***
- In 2010, 35 INCITE projects allocated more than 600M CPU hours at the ALCF





DOE ALCC Program

ASCR Leadership Computing Challenge

- **Areas of special interest to DOE**
- **An emphasis on high risk, high payoff simulations**
- **30% of the core hours at Leadership Facilities**
- **<http://science.energy.gov/ascr/facilities/alcc/>**
- **10 awards at ALCF in 2010 for 300+ million core hours**





Discretionary Allocations

- **ALCF Discretionary allocations provide time for:**
 - Porting, scaling, and tuning applications
 - Benchmarking codes and preparing INCITE proposals
 - Preliminary science runs prior to an INCITE award
 - Early Science Program
- **To apply go to the ALCF allocations page**
 - www.alcf.anl.gov/support/gettingstarted



Early Science Program

- **Goals**

- Help us shake-out the system and software stack using real applications
- Develops community and ALCF expertise on the system
- A stable and well- documented system moving into production
- Exemplar applications over a broad range of fields
- At least 2 billion core-hours to science

- **16 projects**

- Large target allocations
- Postdoc

- **Proposed runs between *Mira* acceptance and start of production**

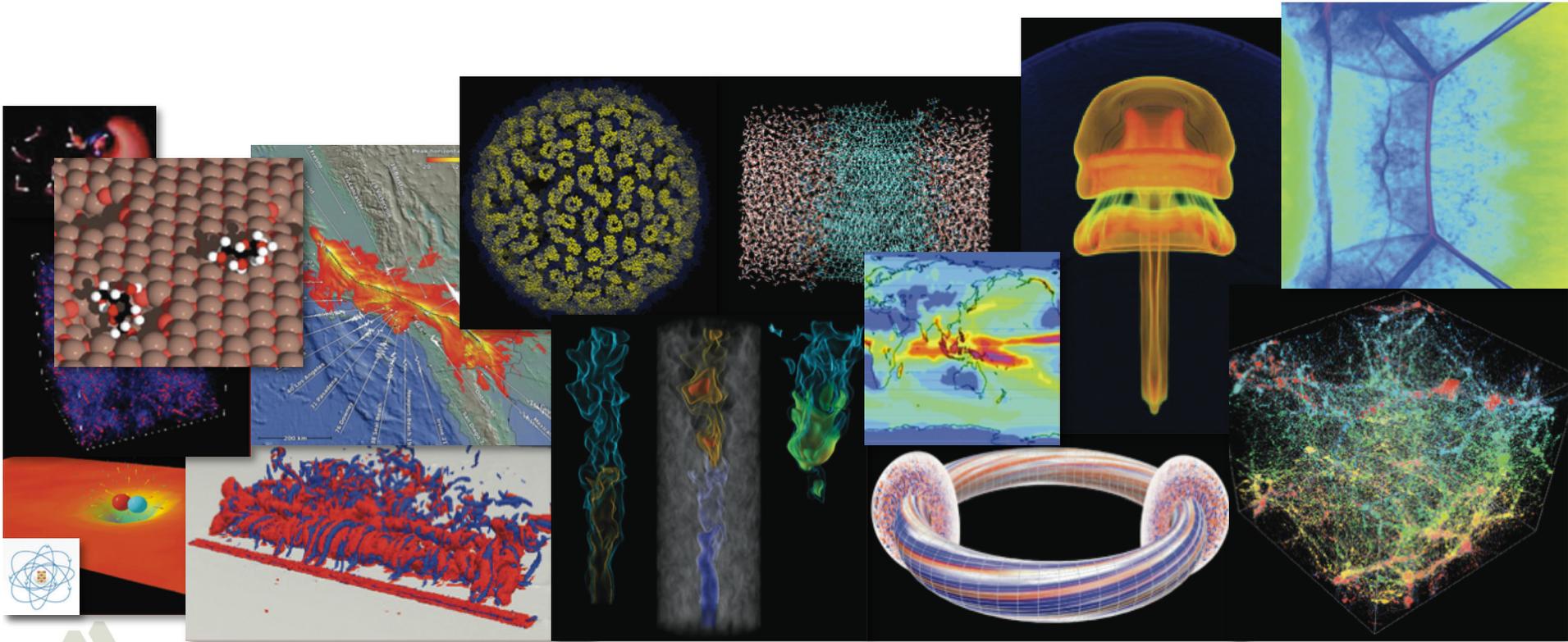
- **2 billion core-hours to burn in a few months**



First in *Mira* Queue: Early Science Program

- Chemistry
- Nuclear Structure
- Biology
- Materials
- Astro/Cosmology
- Climate
- Fusion
- Geophysics
- Combustion
- CFD/Aero
- Energy

<http://esp.alcf.anl.gov>



Tools and Libraries



BG/Q Performance Tools

- **Collaboration between tool providers, IBM, and ANL**
- **BG/Q provides a hardware & software environment that supports many standard performance tools:**
 - Software:
 - Environment similar to 64 bit PowerPC Linux
 - provides standard GNU binutils
 - New performance counter API bgpm
 - Performance Counter Hardware:
 - 424 64-bit counters in a central node counting unit
 - Counter for all cores, prefetchers, L2 cache, memory, network, message unit, PCIe, DevBus, and CNK events
 - Provides support for hardware threads and counters can be controlled at the core level
 - Countable events include: instruction counts, flop counts, cache events, and many more



BG/Q Tools

Tool Name	Source	Provides	Q Status
gprof	GNU/IBM	Timing (sample)	In development
TAU	Unv. Oregon	Timing (inst), MPI	Development pending
Rice HPCToolkit	Rice Unv.	Timing (sample), HPC (sample)	In development & testing
IBM HPCT	IBM	MPI, HPC	In development
mpiP	LLNL	MPI	In development & testing
PAPI	UTK	HPC API	In development & testing
Darshan	ANL	IO	In development & testing
Open Speedshop	Krell	Timing (sample), HCP, MPI, IO	In development
Scalasca	Juelich	Timing (inst), MPI	In development & testing
FPMPI2	UIUC	MPI	Development planned
DynInst	UMD/Wisc/IBM	Binary rewriter	In development
ValGrind	ValGrind/IBM	Memory & Thread Error Check	Development planned



Parallel Debuggers

- **IBM CDTI (Code Development and Tools Interface)**
- **Rogue Wave TotalView**
- **Allinea DDT**
 - Preparation via ANL scalability research contract on BG/P to address I/O node bottlenecks



Libraries

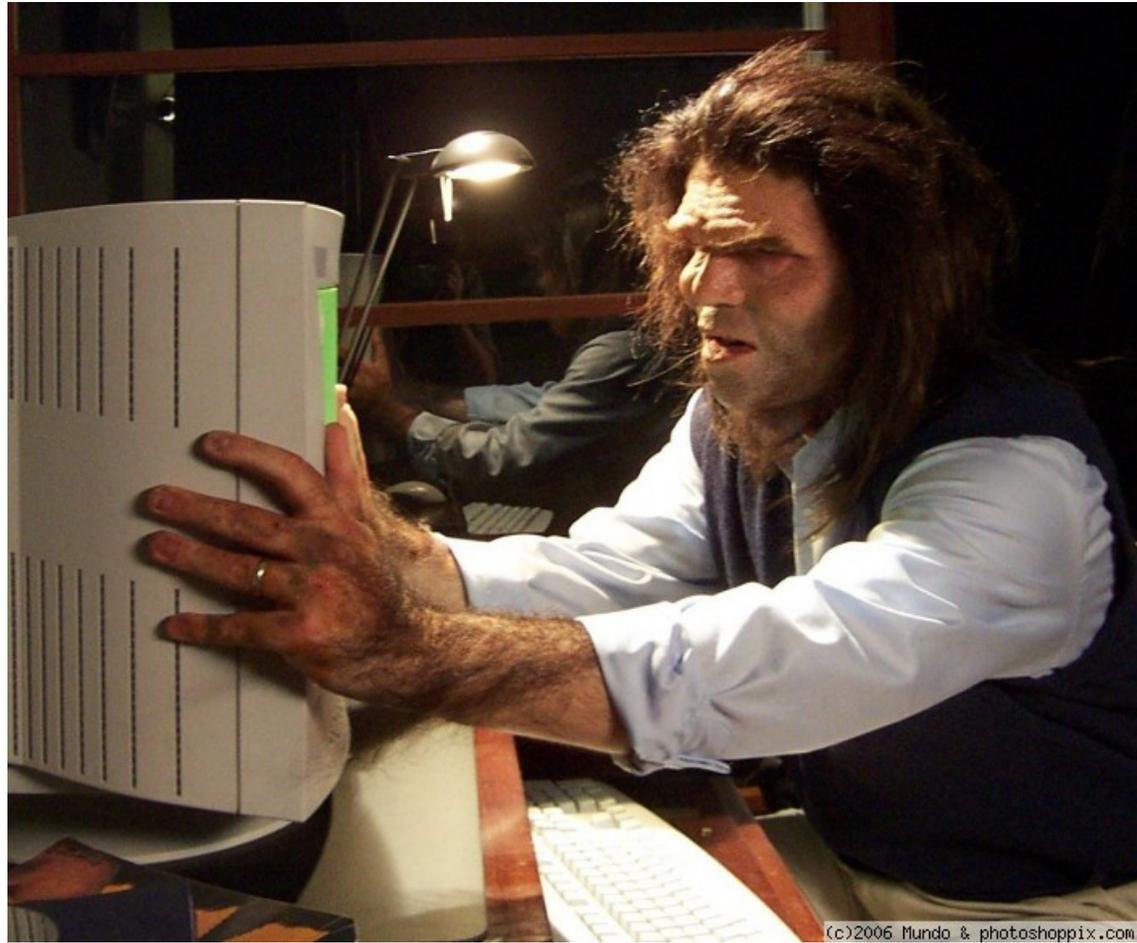
- ESSL
- PETSc
- FFTW
- BLAS
- LAPACK
- ScaLAPACK
- P3DFFT



Performance Modeling & Projection

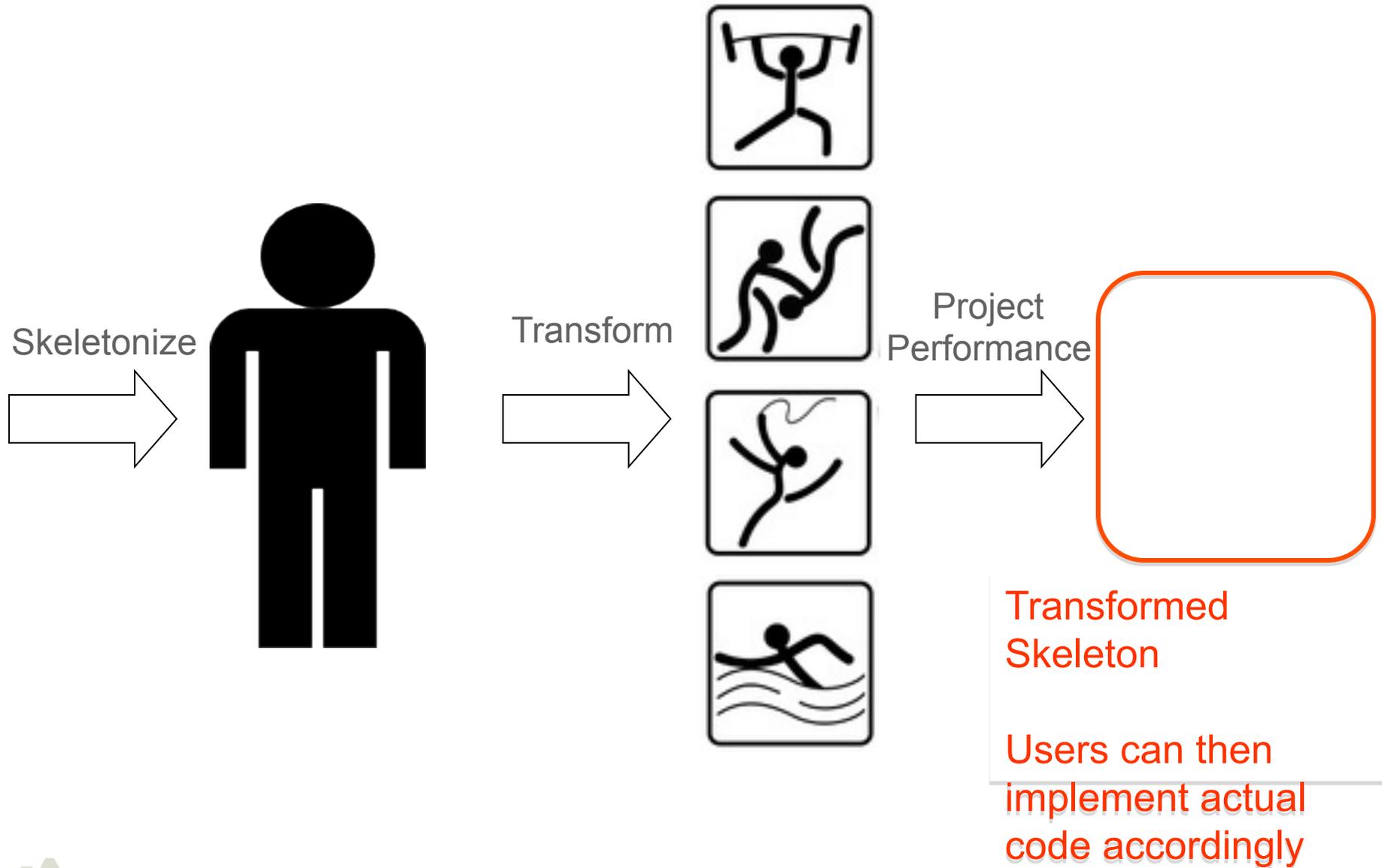








Approach



Acknowledgement



- **Kalyan Kumaran**
- **Charles Bacon**
- **Ray Loy**
- **Tim Williams**
- **Vitali Morozov**
- **Pete Beckman**
- **Michael Papka**

- **Argonne Leadership Computing Facility**



Thank You!



Overview

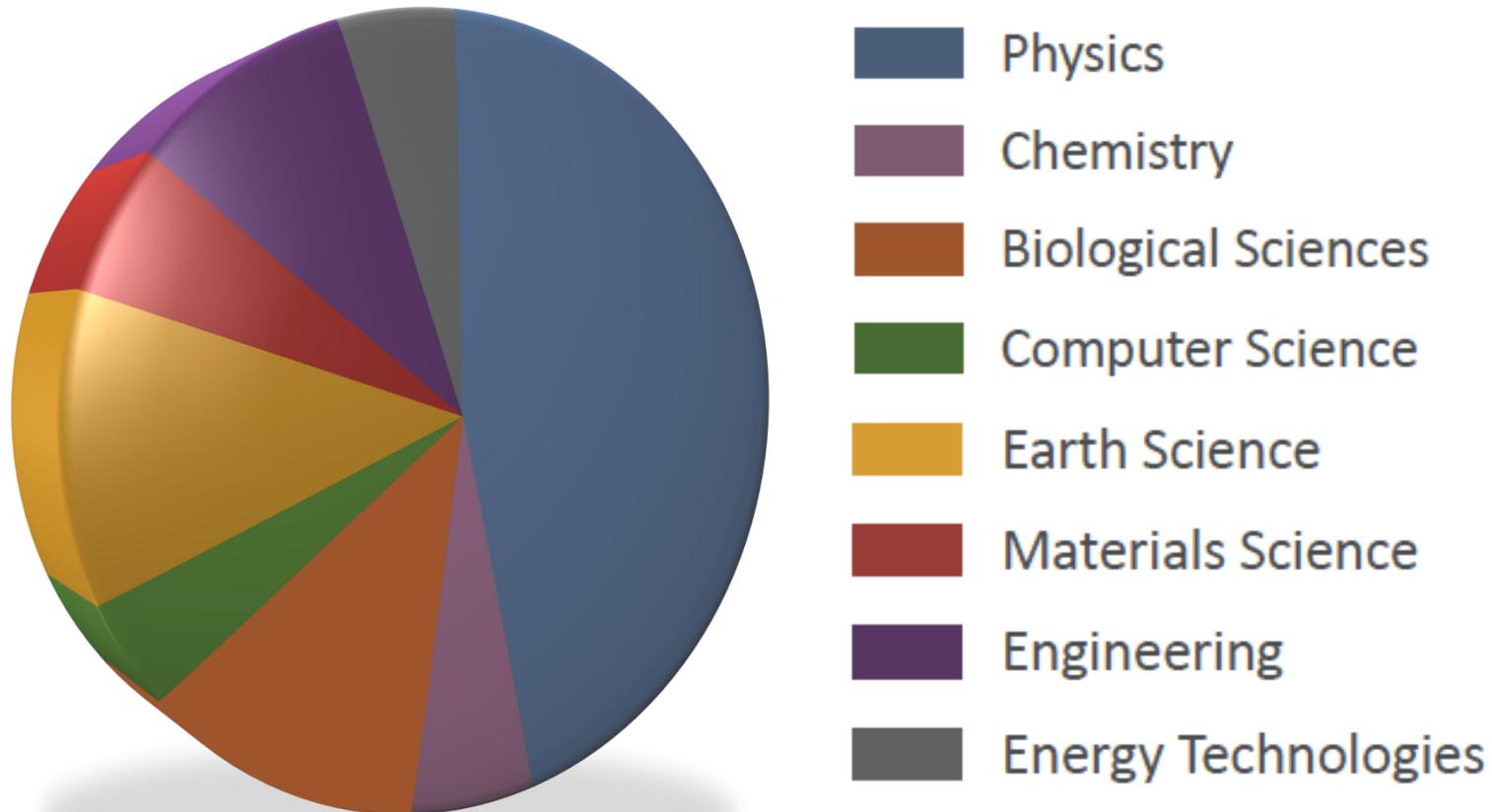
- **Hardware Systems**
- **Allocations and Applications**
- **Libraries and Tools**



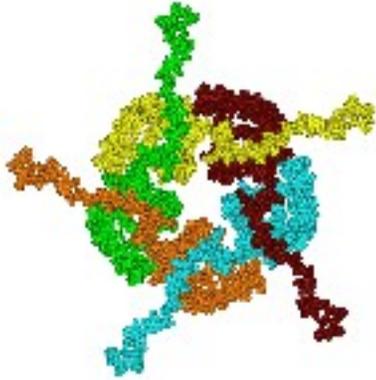
Percentage of compute hours used by scientific discipline

January–October 2010

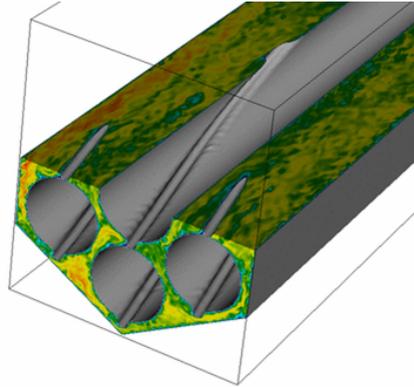
Total hours used: 849 Million



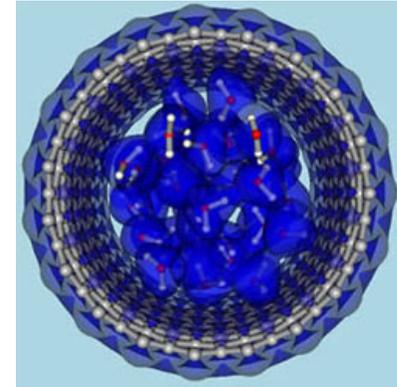
ALCF Projects Span Many Domains



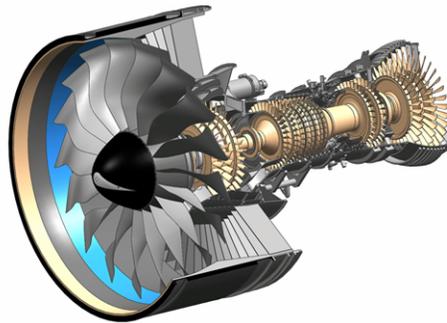
Life Sciences
U CA-San Diego



Applied Math
Argonne Nat'l Lab

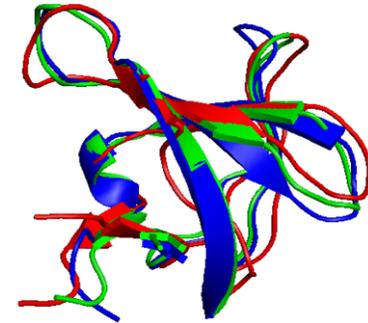


Physical Chemistry
U CA-Davis



Nanoscience
Northwestern U

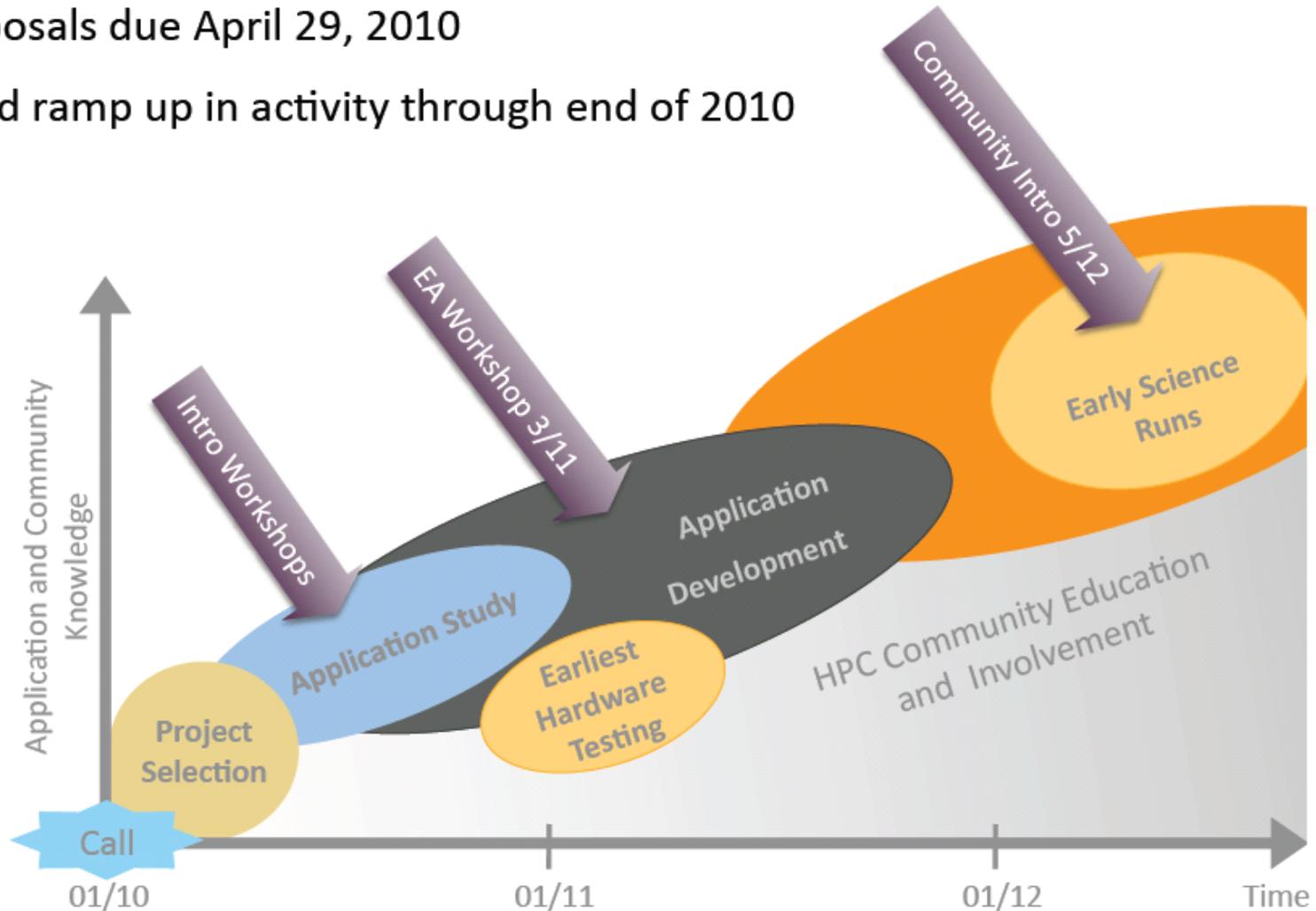
Engineering Physics
Pratt & Whitney



Biology
U Washington

Early Science Program Timeline

- Call Opens January 2010
- Proposals due April 29, 2010
- Rapid ramp up in activity through end of 2010



Argonne Leadership Computing Facility

- ***Intrepid* - ALCF Blue Gene/P System:**

- 40,960 nodes / 163,840 PPC cores
- 80 Terabytes of memory
- Peak flop rate: 557 Teraflops
- Linpack flop rate: 450.3

- ***Eureka* - ALCF Visualization System:**

- 100 nodes / 800 2.0 GHz Xeon cores
- 3.2 Terabytes of memory
- 200 NVIDIA FX5600 GPUs
- Peak flop rate: 100 Teraflops

- **Storage:**

- 6+ Petabytes of disk storage with an I/O rate of 80 GB/s
- 5+ Petabytes of archival storage (10,000 volume tape archive)



New Resources Coming CY2012

- **Mira - Blue Gene/Q System**

- 48K nodes / 768K cores
- 786 TB of memory
- Peak flop rate: 10 PF

- **Storage**

- ~35 PB capacity, 240GB/s bandwidth (GPFS)

- **New Visualization Systems**

- Initial system in 2012
- Advanced visualization system in 2014
 - State-of-the-art server cluster with latest GPU accelerators
 - Provisioned with the best available parallel analysis and visualization software



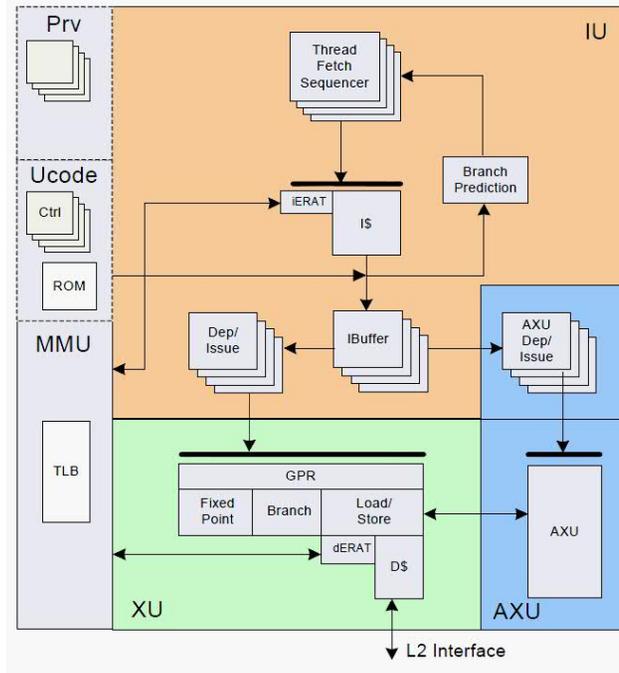
BG/Q Processor Unit

■ A2 processor core

- Mostly same design as in PowerEN™ chip
- Implements 64-bit PowerISA™
- Optimized for aggregate throughput:
 - 4-way simultaneously multi-threaded (SMT)
 - 2-way concurrent issue 1 XU (br/int/l/s) + 1 FPU
 - in-order dispatch, execution, completion
- L1 I/D cache = 16kB/16kB
- 32x4x64-bit GPR
- Dynamic branch prediction
- 1.6 GHz @ 0.8V

■ Quad FPU

- 4 double precision pipelines, usable as:
 - scalar FPU
 - 4-wide FPU SIMD
 - 2-wide complex arithmetic SIMD
- Instruction extensions to PowerISA
- 6 stage pipeline
- 2W4R register file (2 * 2W2R) per pipe
- 8 concurrent floating point ops (FMA)
 - + load + store
- Permute instructions to reorganize vector data
 - supports a multitude of data alignments



QPU: Quad FPU

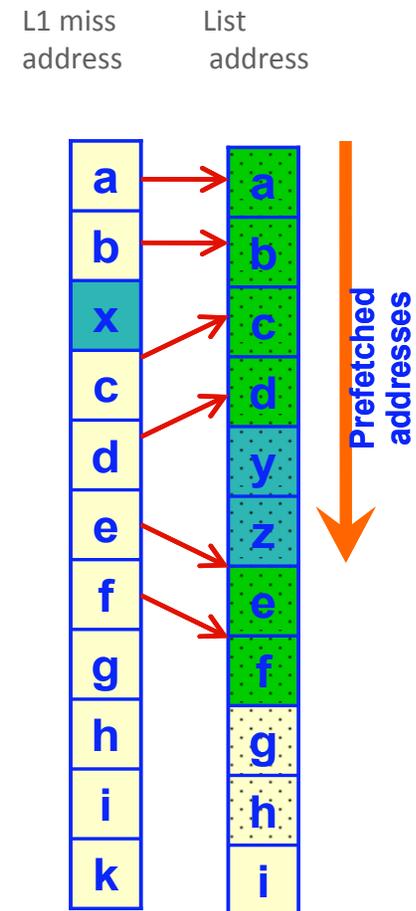
BlueGene/Q PUnit - ct.

▪ L1 prefetcher

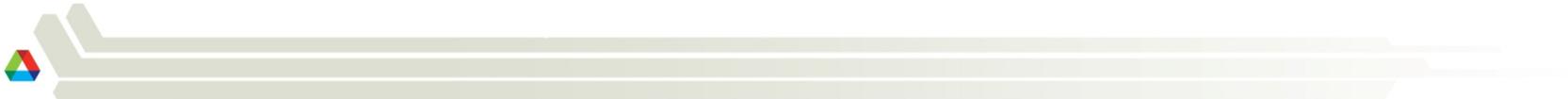
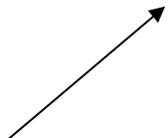
- Normal mode: [Stream Prefetching](#)
 - in response to observed memory traffic, adaptively balances resources to prefetch L2 cache lines (@ 128 B wide)
 - from 16 streams x 2 deep through 4 streams x 8 deep
- Additional: 4 [List-based Prefetching](#) engines:
 - One per thread
 - Activated by program directives, e.g. bracketing complex set of loops
 - Used for repeated memory reference patterns in arbitrarily long code segments
 - Record pattern on first iteration of loop; playback for subsequent iterations
 - On subsequent passes, list is adaptively refined for missing or extra cache misses (async events)

▪ Wake-up unit

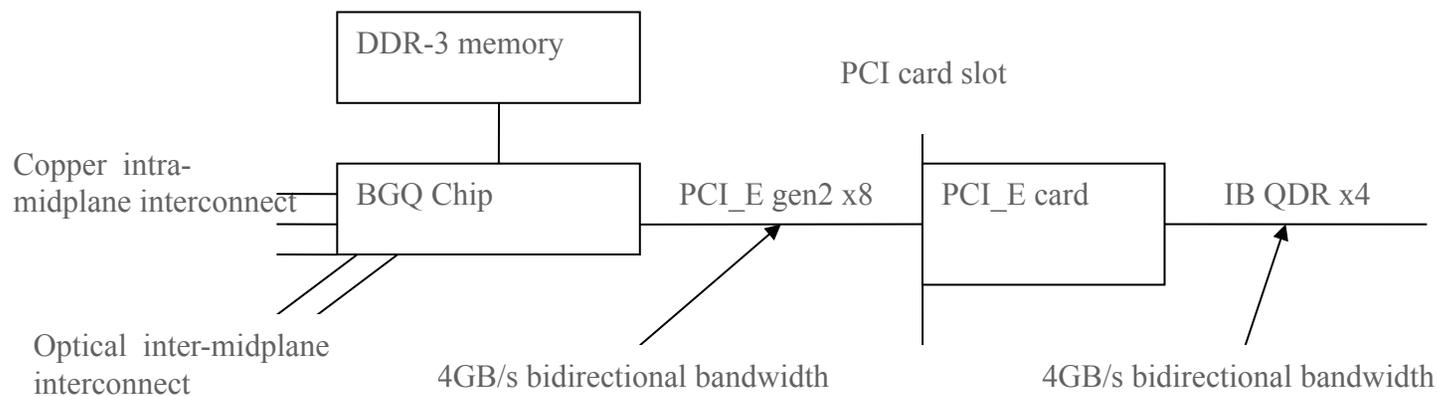
- Will allow SMT threads to be suspended, while waiting for an event
- Lighter weight than wake-up-on-interrupt -- no context switching
- Improves power efficiency and resource utilization



List-based “perfect” prefetching has tolerance for missing or extra cache misses



Blue Gene/Q I/O node



Alternatives:

- PCI_E to IB QDR x4 (shown)
- PCI_E to (dual) 10 Gb ethernet card (log in nodes)
- PCI_E to single 10GbE + IB QDR
- PCI_E to SATA for direct disk attach



BG I/O Max Bandwidth

	BG/L	BG/P	BG/Q
Type	1GbE	10GbE	PCI-e
BW/node	1Gb/s x2 250MB/s	10Gb/sx2 2.5GB/s	4GB/sx2
# of I/O nodes	128	64	8-128
BW/rack in	16GB/s	80GB/s	512GB/s@128
BW/rack out	16GB/s	80GB/s	512GB/s@128
I/O byte/flop	0.0056	0.011	0.0048





Libraries in /soft/libraries/unsupported

- Not actively maintained by ALCF (at least for now)
- Provided as a convenience
- Boost 1.49.0
- HDF5 1.8.8
- NETCDF 4.1.3
- P3DFFT 2.4 (patched)
- Tcl 8.4.14
- zlib 1.2.6





Math Libraries in /soft/libraries/alcf

- Maintained in-house, frequently updated
- GCC and XL built versions of each library
- BLAS
- LAPACK 3.3.1
- ScaLAPACK 2.0.2
- FFTW 2.1.5
- FFTW 3.3.1
- PARPACK



BG/Q A2 Core - Quick Overview for the Programmer

- Full PowerPC compliant 64-bit CPU (BG/P PowerPC 450d was 32-bit)
- 1.6GHz, in-order execution, 4 hardware threads/core, 16 cores/node, 16GB/node
- At most one instruction can be completed per cycle per thread
- At most 2 instructions can be completed per cycle per core:
 - one instruction must be integer/load/store (XU)
 - one instruction must be floating point (AXU)
- 4-wide SIMD floating point unit with complete set of parallel instructions
 - 4 FMA's @ 1.6GHz = 12.8 Gflops/core
- Cache:
 - 16 KB L1 data cache, 64 byte lines, shared between 4 hardware threads
 - L1 Prefetch buffer, 32 lines, 128 bytes each
 - 32 MB shared L2 cache



Comments on using all hardware threads

- Speed up with hardware threads will be limited if the issue rate is already high with 1 thread/core (NEK is an example).
- Speed-up with hardware threads will be limited if the problem is already near the scaling limit at 1 thread/core. Using all threads will require 4x more threads.
- Speed-up can be limited if there is contention for L1-D and L1P resources.
- In some cases using OpenMP or Pthreads instead of MPI might reduce L1 contention.

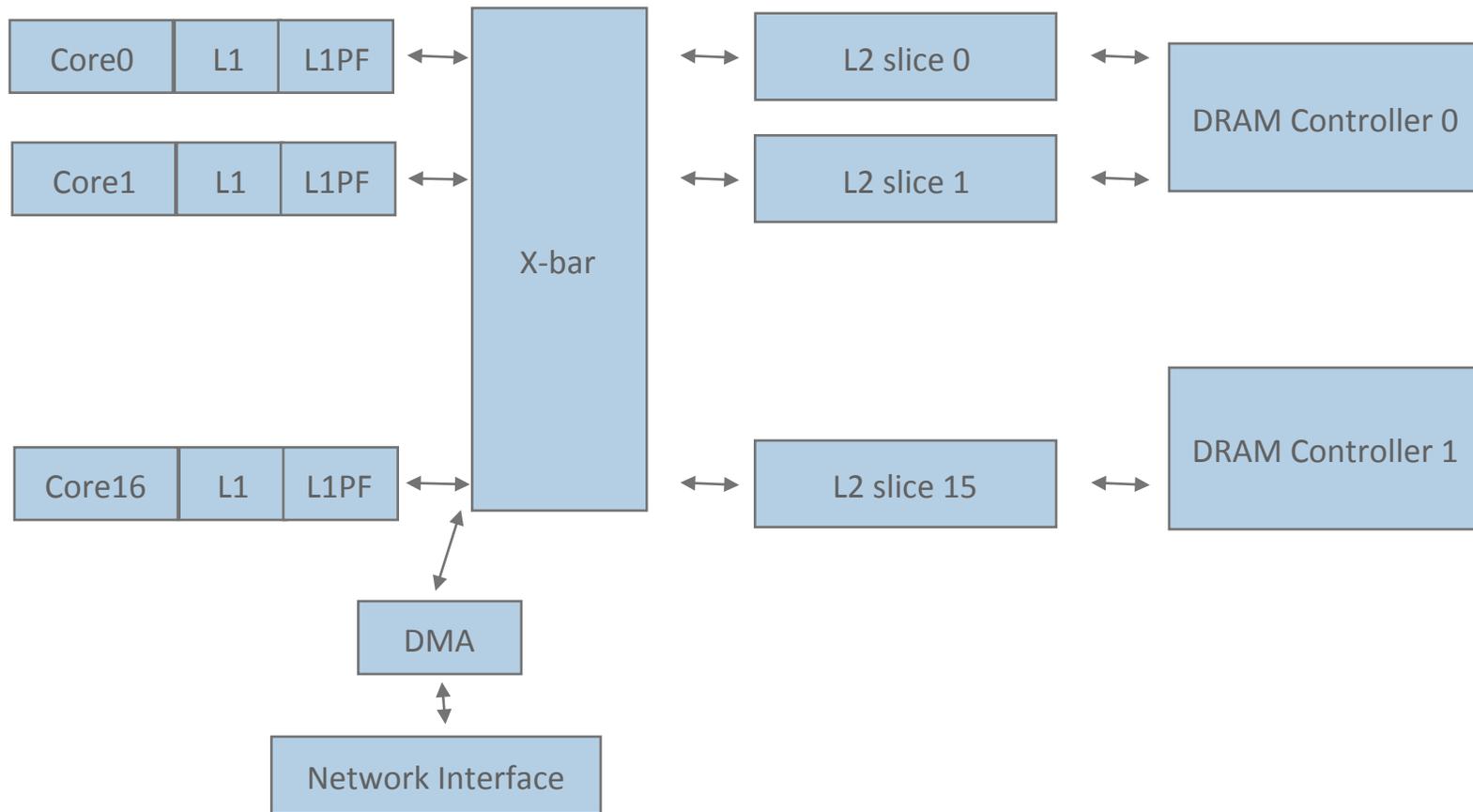


Libraries

- ESSL available through IBM
- PETSc is being optimized as part of BG/Q Tools ESP project
- Will port and tune 3rd party libraries (FFTW, BLAS, LAPACK, ScaLAPACK, ParMetis, P3DFFT, ...) using compiler optimizations
- Collecting actual library usage data; libraries will be stamped with a detectable string id.
- Collaborating with Robert van de Geijn's group on rewriting Goto-BLAS so that it can be easily ported and tuned to new architectures like BG/Q (BLIS)
- Exploring an optimized FFT library with Spiral Gen



BG/Q Memory Structure



BG/Q Performance Tools

- A variety of tools providers are currently working with IBM and Argonne to port and test tools on the Q
- **BG/Q provides a hardware & software environment that supports many standard performance tools:**
 - Software:
 - Environment similar to 64 bit PowerPC Linux
 - provides standard GNU binutils
 - New performance counter API bgpm
 - Performance Counter Hardware:
 - BG/Q provides 424 64-bit counters in a central node counting unit
 - Counter for all cores, prefetchers, L2 cache, memory, network, message unit, PCIe, DevBus, and CNK events
 - Provides support for hardware threads and counters can be controlled at the core level
 - Countable events include: instruction counts, flop counts, cache events, and many more



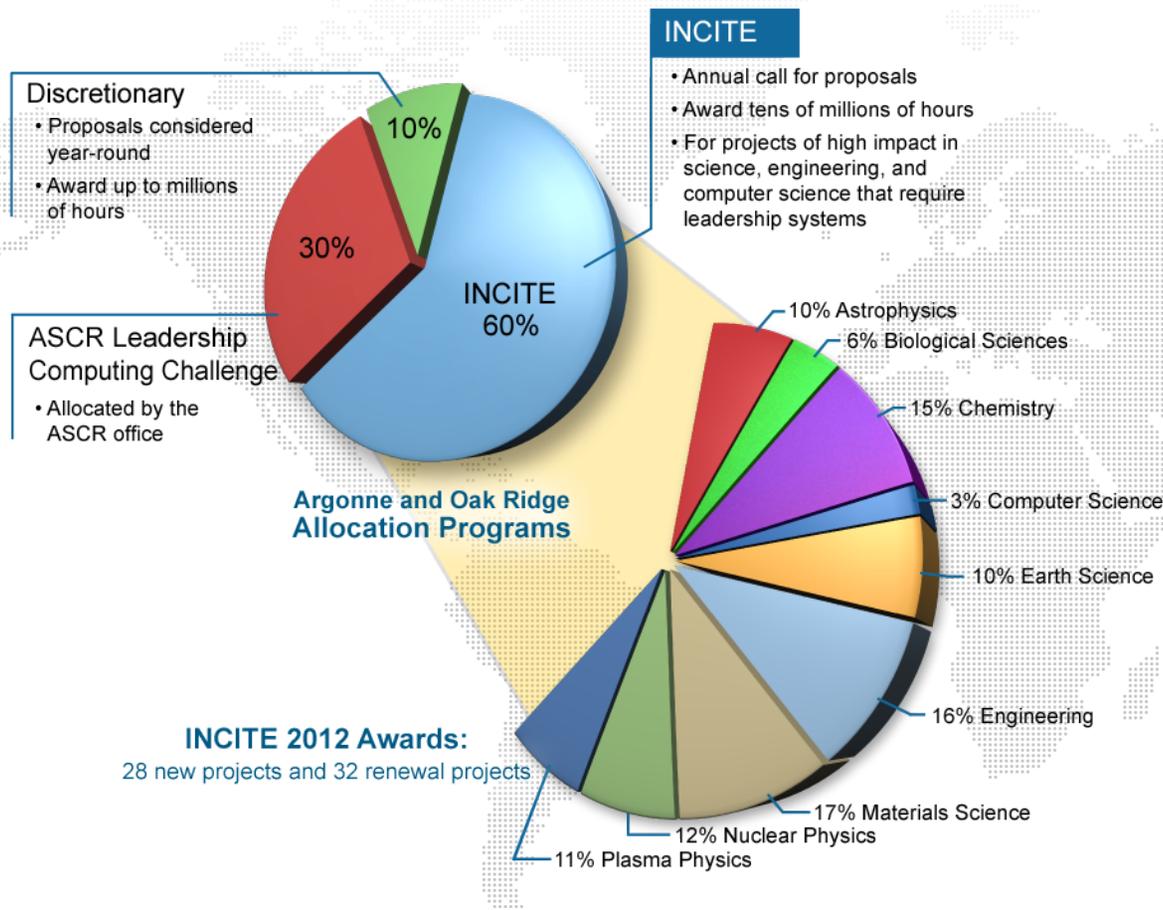
Parallel Debuggers

- **IBM CDTI (Code Development and Tools Interface)**
 - Collaboration of IBM/LLNL/ANL resulted in update v1.7 (August 2011)
 - Refined interface for multiple tool support, breakpoint handling, stepping, and signal handling
- **Rogue Wave TotalView**
 - Ported to BG/Q (Q32 at IBM) with basic functionality in August 2011
 - Pre-release testing by LLNL December 2011
 - Status
 - Tested working: basic ops (step, breakpoint, stack), QPX instructions, fast conditional breakpoints, job control for C/C++/Fortran with MPI/OMP/threads.
 - Still testing: **scalability**, fast conditional watchpoints, debugging in TM/SE
- **Allinea DDT**
 - Preparation via ANL scalability research contract on BG/P to address I/O node bottlenecks
 - Multiplexed debug daemons – complete and tested (Nov 2011)
 - Multiplexed gdbserver processes – complete and tested for single threading (Dec 2011)
 - Still testing: multiplexed gdbserver with multiple threads/process.
 - Status
 - Expected BG/P Beta release Jan 2012.
 - BG/Q port to begin on ANL T&D Feb 2012 as part of Early Science project (ESP).



Innovative and Novel Computational Impact on Theory and Experiment (INCITE)

INCITE provides awards of time on the Argonne and Oak Ridge Leadership Computing Facility (ALCF and OLCF) systems for researchers to pursue transformational advances in science and technology: **1.7 billion core hours** were awarded in 2012.



Call for Proposals

The INCITE program seeks proposals for high-impact science and technology research challenges that require the power of the leadership-class systems. Allocations will be for calendar year 2013.

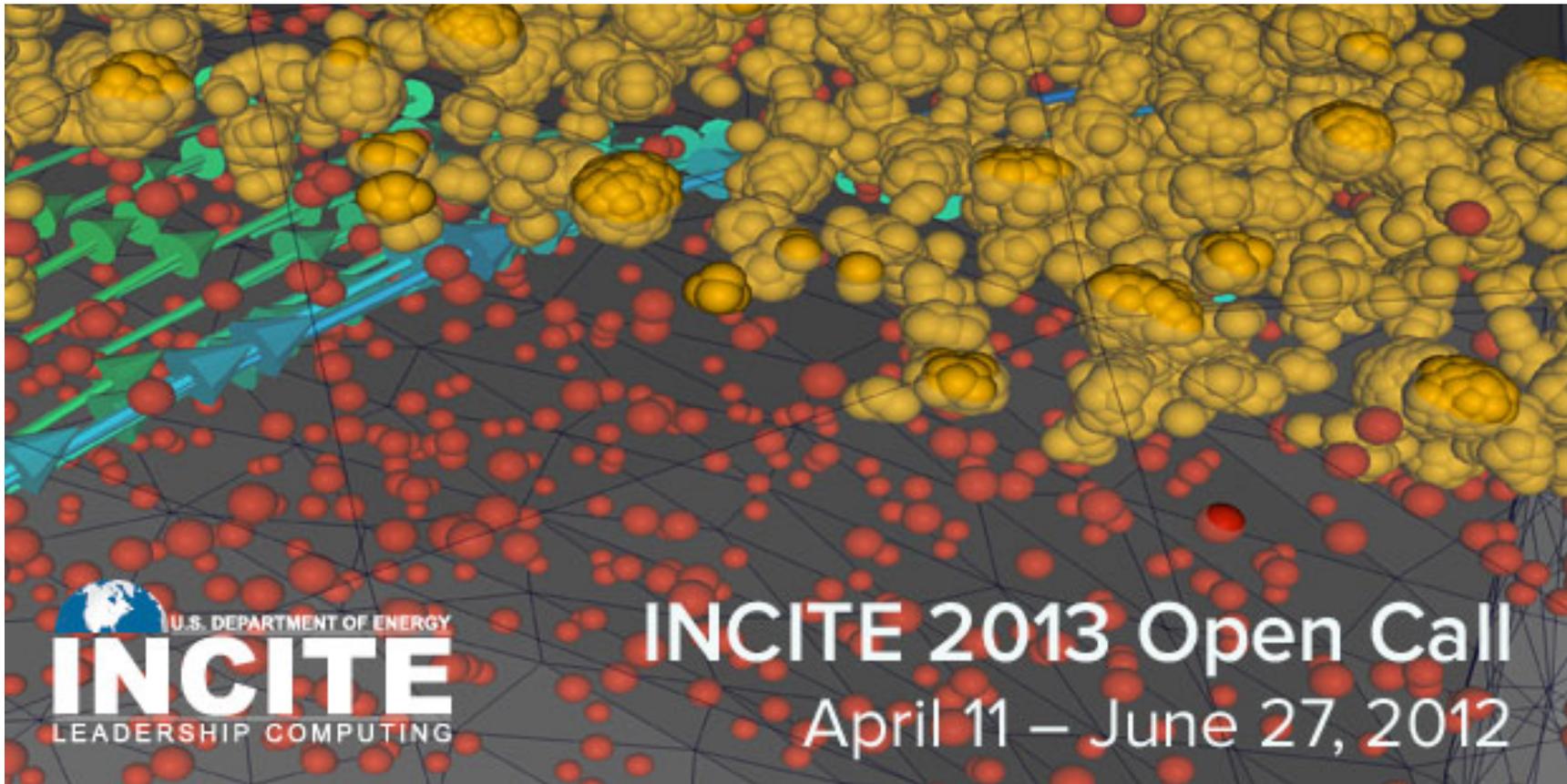
April 11 - June 27, 2012

Contact information:

Julia C. White, INCITE Manager
whitejc@DOEleadershipcomputing.org



Allocations and Applications



 U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

INCITE 2013 Open Call
April 11 – June 27, 2012



Math Library Future Plans

- LAPACK 3.4.1 port (with LAPACKE)
- CBLAS
- METIS/ParMETIS
- Goto-BLAS ported and tuned on BG/Q
- New kernel infrastructure codenamed “BLIS”, designed in collaboration with Univ. of Texas
- Right now ESSL GEMM routines are extracted into the ALCF BLAS library
- Tune FFTW 3.x, time permitting

