Preparing Applications for *Mira*, a 10 PetaFLOPS IBM Blue Gene/Q System

Timothy J. Williams

Argonne Leadership Computing Facility

Argonne National Laboratory

ANL Booth presentation at SC11

11/16/2011

- **ALCF**
- ***Mira***
- **Applications**
  - Design of Blue Gene/Q
  - Early Science Program
  - Tools & Libraries

Argonne Leadership Computing Facility

- **Established 2006 at Argonne National Lab**
- **One of two DOE national Leadership Computing Facilities (OLCF is other)**

Computer Time Grants

- **DOE INCITE Program – 60% of available cycles**
  - Annual CFP solicits large, computationally-intensive projects requiring LCF capability
    - Open to all disciplines and institutions worldwide
- **2011 INCITE at ALCF**
  - **30 projects**
  - 732**M core hours**
- **ALCC Program – 30% of cycles**
  - **2012 ALCC: 5 projects**
- **Director's Discretionary – 10%**
  - https://accounts.alcf.anl.gov

Today's ALCF Hardware

- ***Intrepid* - IBM Blue Gene/P System**
  - 40K nodes / 160K PPC cores
    - 40 racks of 1024 nodes
  - 80 terabytes of memory
  - 557 teraFLOPS
  - 8 petabytes disk

Next ALCF System: *Mira*

- **Coming in 2012**
  - 48K nodes, over 750K cores
    - 48 racks
  - 800 TB of memory
  - 10 petaFLOPS
  - 70 PB of disk
- **Next Blue Gene/Q architecture**
  - 16 cores/node
  - 16 GB of memory/node
  - water cooled
- **BG/P applications should run immediately on the BG/Q**
  - Better performance expected with higher levels of on-node parallelism

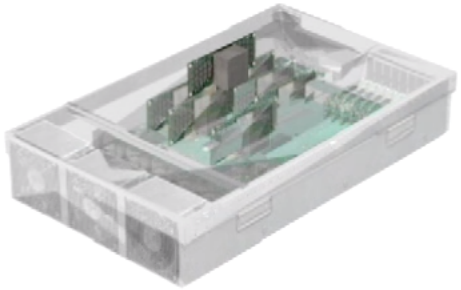
Next ALCF System: *Mira*

- **Coming in 2012**
  - 48K nodes, over 750K cores
    - 48 racks
  - 800 TB of memory
  - 10 petaFLOPS
  - 70 PB of disk
- **Next Blue Gene/Q architecture**
  - 16 cores/node
  - 16 GB of memory/node
  - water cooled
- **BG/P applications should run immediately on the BG/Q**
  - Better performance expected with higher levels of on-node parallelism

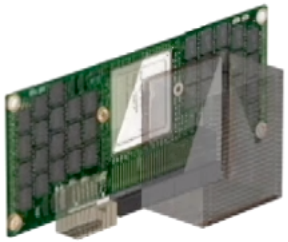
Threads

Blue Gene/Q Packaging *(cont'd)*

Blue Gene/Q Packaging *(cont'd)*



I/O Nodes



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 (neighbor)	2.6 us	2.2 us	--
Concurrency / Rack	4,096	65,536	16x
Network Interconnect	3D torus	5D torus	Smaller diameter
GFlops/Watt	0.77	2.10	3x
Cooling	Air	Water	~30% savings/W

Guide to Machines

- ***Mira***
  - Blue Gene/Q
- ***Intrepid***
  - Blue Gene/P
  - Current ALCF production machine
- ***EAS* – Early Access System**
  - Blue Gene/Q
  - ANL's to use
    - IBM T.J. Watson
  - 128 nodes
  - Latest compilers

Applications and Design of IBM Blue Gene/Q

- **Partnership of IBM, LLNL, ANL**
  - Detailed discussions of hardware/software requirements
    - Quarterly Executive Review meetings
    - Bi-weekly working-group conference calls
- **Applications and kernels specified in contracts with IBM**
  - Expectations of
    - Functionality
    - Correctness
    - Performance
  - Applications of key importance to labs

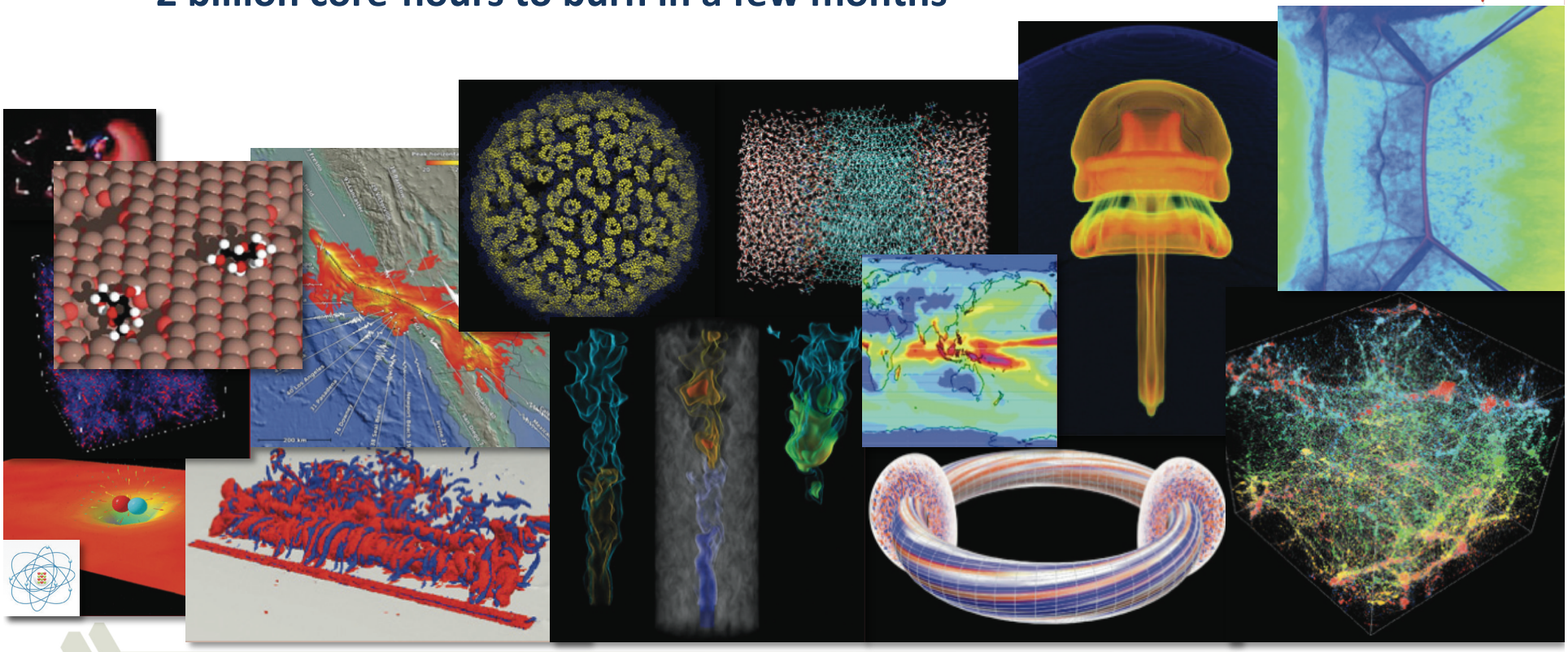
Applications and Design of IBM Blue Gene/Q

- **Partnership of IBM, LLNL, ANL**
  - Detailed discussions of hardware/software requirements
    - Quarterly Executive Review meetings
    - Bi-weekly working-group conference calls
- **<u>Applications</u> and kernels specified in contracts with IBM**
  - Expectations of
    - Functionality
    - Correctness
    - Performance
  - Applications of key importance to labs

First in *Mira* Queue: Early Science Program

- **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**

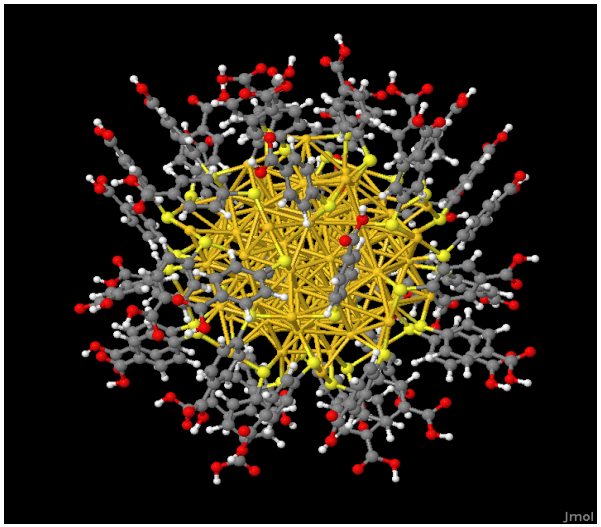
http://esp.alcf.anl.gov



Materials Design and Discovery: Catalysis and Energy Storage

Larry Curtiss (Argonne National Laboratory)
Nick Romero (ALCF)
Anouar Benali (ESP postdoc, ALCF)

- **Electronic Structure Codes: QMCPACK, CPMD**
  - Quantum Monte Carlo (QMC)
  - Density functional theory (DFT)
- **Address catalysis and electric energy storage in 4 areas**
  - Biomass conversion: structure of nanobowls on metal oxide surfaces
  - Electrical energy interfaces
  - Lithium-air batteries
  - Catalysis with transition metal nanoparticles
    - Simulated nanoparticles of up to 1415 atoms using 40% of *Intrepid*



Materials Design and Discovery: Catalysis and Energy Storage *(cont'd)*

- **Quantum Monte Carlo for electronic structure**
  - Operations depend on type of wave function: LCAO, real-space, PWs.
    - Spline interpolation
    - Small DGEMM and DGEMV
- **Current performance**
  - Mixture of compute and bandwidth-limited kernels
    - 5-10% of per core peak performance on IBM Blue Gene/P
    - 20-30% of per core peak performance on x86
  - Heavily rely on C++ compiler optimizations
  - OpenMP 2.5 compliance
- **Paths forward**
  - Reformulate loops to use BLAS2+3 (in progress)
  - Hand tune the SIMD kernels
  - Add nested parallelism to MCWalker evaluation
    - Requires OpenMP 3.0
  - IBM Zurich is optimizing CPMD for Blue Gene/Q

Accurate Simulation of Chemistry in Energy Production & Storage

Robert Harrison (Oak Ridge National Laboratory)
Jeff Hammond (ALCF)
Alvaro Vazquez-Mayagoitia (ESP postdoc, ALCF)

- **Codes: MADNESS & MPQC**
- **Catalysis (chemical processes on metal-oxide surfaces)**
  - MADNESS: Model 500-2000 atom lithium oxide clusters
  - MPQC: 50-200 atom models of organic and surface catalysis
    - Run without an eigensolver
- **Heavy element chemistry for fuel reprocessing**
  - Molecular interfacial partitioning
  - Ligand design
  - *Ab initio* dynamics to include finite temperature and entropy

Accurate Simulation of Chemistry in Energy Production & Storage

Robert Harrison (Oak Ridge National Laboratory)
Jeff Hammond (ALCF)
Alvaro Vazquez-Mayagoitia (ESP postdoc)

MADNESS

- **Codes: MADNESS & MPQC**
- **Catalysis (chemical processes on metal-oxide surfaces)**
  - MADNESS: Model 500-2000 atom lithium oxide clusters
  - MPQC: 50-200 atom models of organic and surface catalysis
    - Run without an eigensolver
- **Heavy element chemistry for fuel reprocessing**
  - Molecular interfacial partitioning
  - Ligand design
  - *Ab initio* dynamics to include finite temperature and entropy

Accurate Simulation of Chemistry in Energy Production & Storage *(cont'd)*

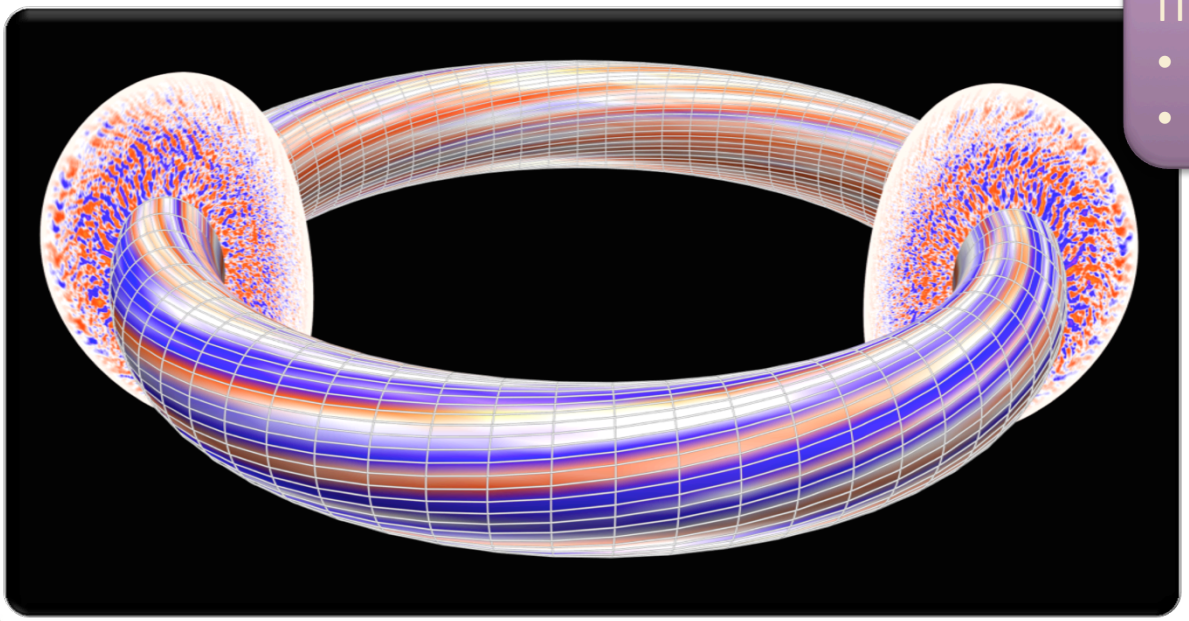
- **Replaced LAPACK with Eigen**
  - fits with C++ OO/template design of MADNESS
- **Tuning assembly implementation of key kernel (mtxm)**
- **Thread Building Blocks (TBB) port for BGP, POWER7 and BGQ**
- **Pthread + OpenMP interoperability and affinity optimizations underway**
- **Exploring native active-message implementation instead of MPI+polling**

- **MADNESS Running on Blue Gene/Q EAS**
  - 64 nodes x {15,30,45,60} compute threads/node
  - Scaling surprisingly good given no tuning and early versions of components.

Global Simulation of Plasma Microturbulence at the Petascale & Beyond

William Tang (Princeton Plasma Physics Laboratory)
Stephane Ethier (PPPL)
Bei Wang (Princeton U.)

- **Codes: GTC, GTS**
- **Particle-in-cell simulation of plasma**
  - Study energy loss through turbulence
  - Trying to validate key assumption about scaling in ITER



Long-duration simulation of ITER plasmas
- $O(10^{10})$ particles
- $O(10^{8})$ grid cells

Global Simulation of Plasma Microturbulence at the Petascale & Beyond *(cont'd)*

- **Parallelism: MPI plus loop-level OpenMP**
  - To do: mapping ranks to nodes optimizing for 5D network topology
- **10.7x better performance per node (Blue Gene/Q versus P)**
  - BG/Q: 16 MPI ranks/node, 4 threads/rank
  - BG/P: 4 MPI ranks/node

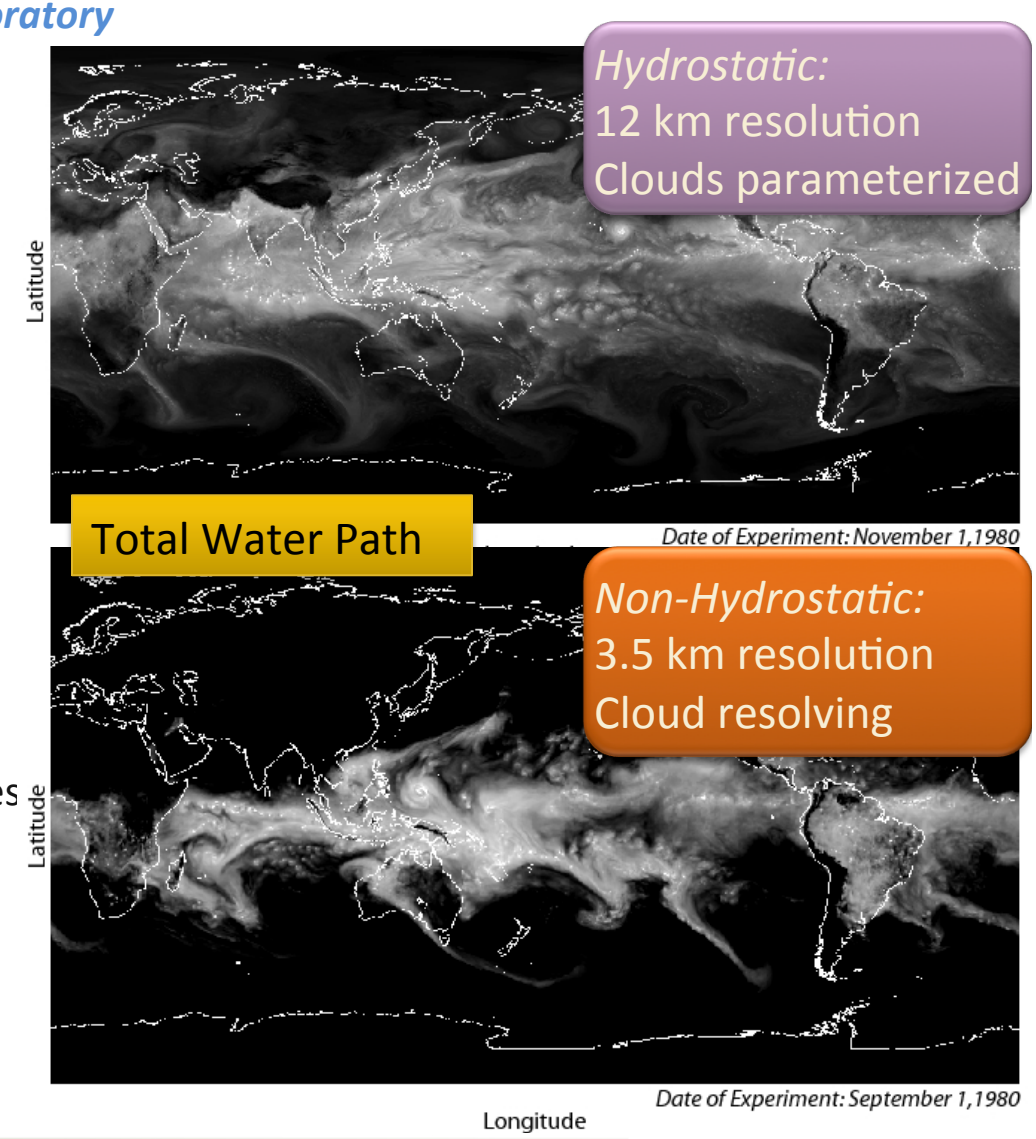
Vitali Morozov (ALCF)

Development of a Prototype Global Cloud-Resolving Climate-Weather Model

Chris Kerr, Geophysical Fluid Dynamics Laboratory
S-J Lin, Isaac Held (GFDL)
V. Balaji, Princeton Univ.

Mira will enable:

- **High-Resolution Atmospheric Model (HIRAM)**
  - Cubed sphere dynamical core
  - Science: understand role of clouds
- **High-Resolution Climate Model (CM4+)**
  - Fully coupled ocean, atmospheric, land, … climate model
  - Science: Fully resolved ocean eddies coupled feedbacks



Development of a Prototype Global Cloud-Resolving Climate-Weather Model *(cont'd)*

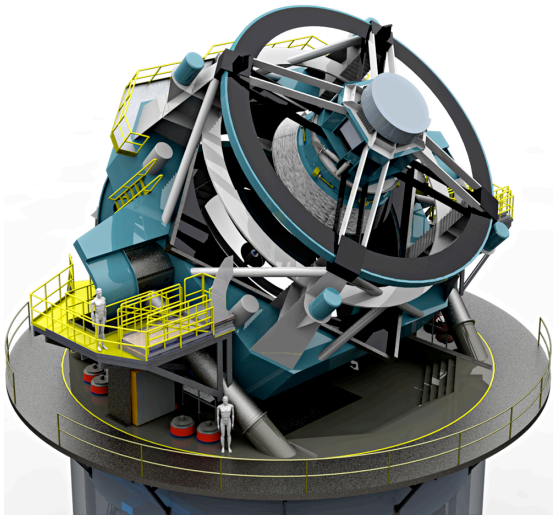
- **Flexible Modeling System (FMS) infrastructure**
  - Supports MPI and high-level OpenMP
- **Coarse-grained threads on high-level tasks**
  - Atmospheric physics and dynamics packages
  - Land package
- **Running on Blue Gene/Q**
  - ANL early access machine (dynamics package with Held-Suarez)
  - Other Q hardware at IBM T. J. Watson
- **10.7x better performance per node (Blue Gene/Q versus P)**
  - Full atmospheric physics
    - BG/Q: 8 MPI ranks, 8 OpenMP threads
    - BG/P: 4 MPI ranks, 4 OpenMP threads
  - Dynamics package only: 8.2x better
    - 2K cores, 8 ranks/node, 8 OpenMP threads/rank, -O2

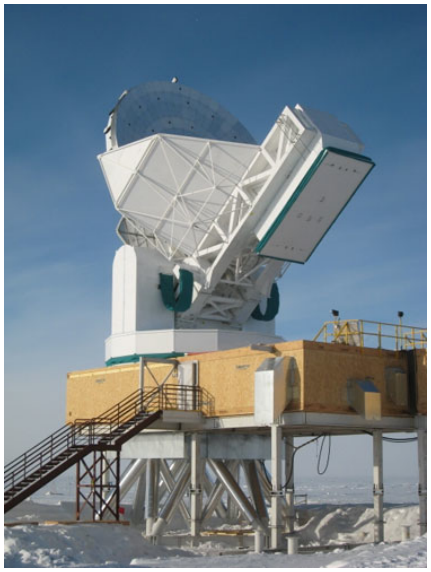
Vitali Morozov (ALCF)

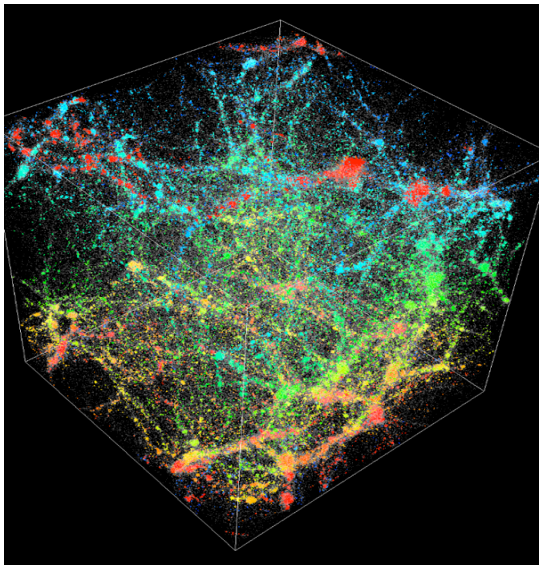
Cosmic Structure Probes of the Dark Universe

Salman Habib (Argonne National Laboratory)
Hal Finkel (ESP postdoc, ALCF)

- **Code: Hardware/Hybrid Accelerated Cosmology Code (HACC) framework**
- **Formation and evolution of large-scale structure in the Universe**
  - Characterize dark energy & dark matter by predicting observational signatures for a variety of new/existing experimental cosmological probes
    - 1st simulations resolving galaxy-scale mass concentration at size scale of state-of-the-art sky surveys
    - Precision predictions for many 'sky survey' observables
  - Study primordial fluctuations by predicting the effects on cosmic structures today







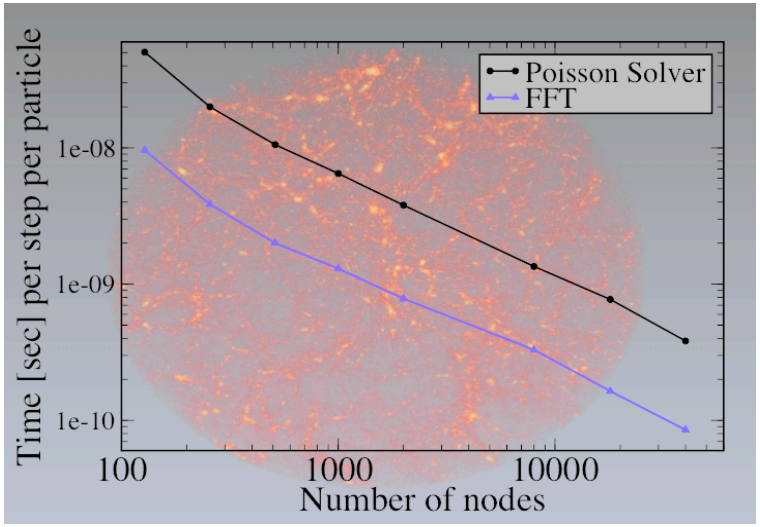
Cosmic Structure Probes of the Dark Universe *(cont'd)*

Two-layer HACC framework designed to run on a variety of architectures

- **No change needed for grid layer of code (long-range forces)**
  - Performance of MPI with new 2D-decomposed parallel FFT tested on Intrepid (full machine)
- **Node-level 'plug-ins' (short-range forces)**
  - Particle-Particle algorithm for Cell and GPU-based nodes
  - Tree algorithm for BG/P and BG/Q implemented
- **Hydrodynamics capability**
  - Particle-based methods under investigation (hydro-PIC as alternative to SPH)

$O(10^{11}\text{-}10^{12})$ grid cells
$O(10^{11}\text{-}10^{12})$ particles

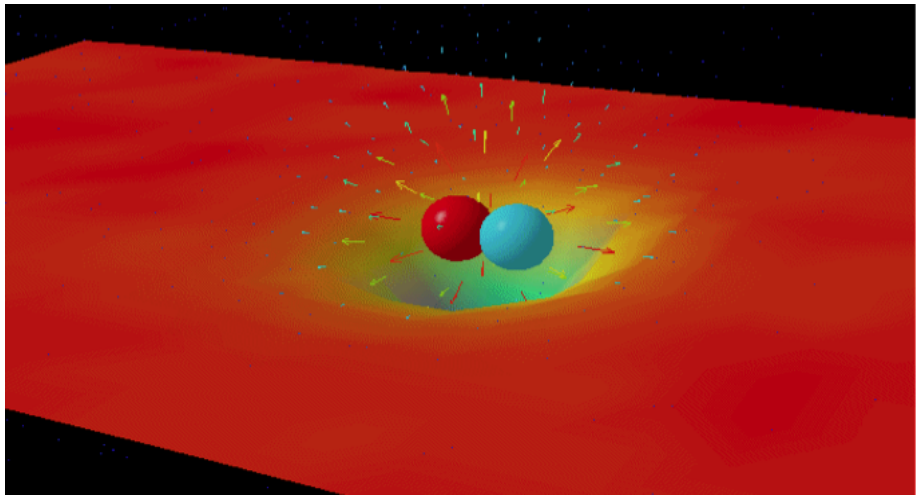
Blue Gene	Threads/ Node	Seconds/particle per timestep
Q	64	4.13e-5
Q	32	4.67e-5
Q	16	7.30e-5
Q	8	1.35e-4
Q	4	2.56e-4
P	4	2.50e-4



Lattice Quantum Chromodynamics

Paul Mackenzie (Fermilab) and the USQCD consortium
James Osborn (ALCF)
Heechang Na (ESP postdoc, ALCF)
4-years with IBM on BG/Q: {Brookhaven, Columbia U., U. Edinburgh}

- **Code: MILC, Chroma, CPS**
- **4D space-time lattice solving theory of quarks and gluons**
  - Determine basic parameters of standard model of particle physics
  - Compute masses, decay properties, internal structure of strongly interacting particles
  - Quantitative understanding of strongly interacting matter under extreme conditions of temperature, density
  - Strongly interacting theories we may need to explain electroweak symmetry breaking under study at LHC



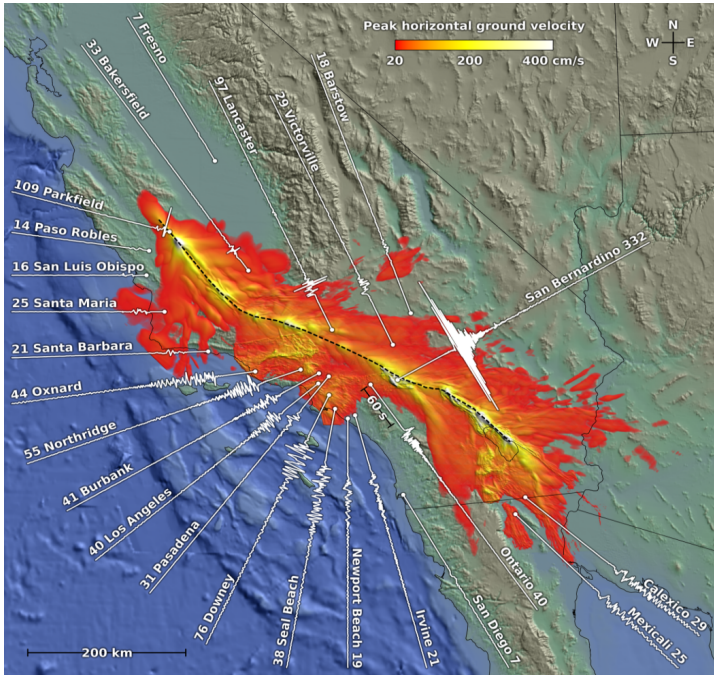
Lattice Quantum Chromodynamics *(cont'd)*

- **Working on Early Access System and other Blue Gene/Q hardware**
  - 512 node prototype at IBM Yorktown
  - Code fragments in Dirac solver getting > 80% of peak communication bandwidth
    - Using low level "SPI" communications
- **Node-level optimizations for Blue Gene/Q**
  - SIMD optimizations – quad FPU on BG/Q
  - Designed prefetching interface between CPU and L2 cache on BG/Q (Boyle/Christ/Kim)
  - Important feedback from LQCD performance to memory system design
  - Dirac matrix solver kernel now gets 60% of theoretical peak on BG/Q chip
  - Full hybrid Monte Carlo evolution now running on BG/Q
    - Running on 128 BG/Q nodes
- **Threading**
  - Dirac solver kernel using pthreads
  - OpenMP used in rest of code
- **Development/porting of improved algorithms**
  - Force-gradient integrator, multi-grid

Using Multi-scale Dynamic Rupture Models to Improve Ground Motion Estimates

Thomas Jordan (USC)
Geoffrey Ely (ESP postdoc, ALCF)

- **Earthquake dynamic rupture modeling, seismic hazard analysis**
- **Implemented hybrid OpenMP/MPI version of the SORD code**
  - Dynamic rupture application
- **Testing mixed C/Python skeleton version of SORD**
  - Threading
  - Optimizing kernels for BG/Q quad FPU.
- **KAUST-IBM collaboration**
  - Includes optimizing SORD



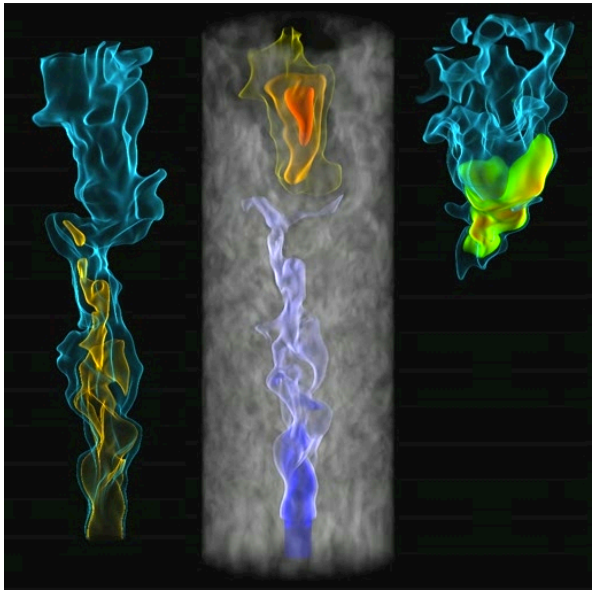
Direct Numerical Simulation of Autoignition in a Jet in a Cross-Flow

Christos Frouzakis(ETH Zürich)
Paul Fischer (Argonne National Laboratory)
Scott Parker (ALCF)
December: Fabrice Schlegal (ESP postdoc, ALCF)

- **Auto-ignition of fuel-air mix related to lean combustion gas turbines**
  - Goal: avoid autoignition for safer, cleaner lean combustion
  - First simulation in lab-scale jet
  - Never studied, experimentally or computationally
- **Code: Nek5000**
  - Spectral element
- **Running on Blue Gene/Q EAS**
  - MPI everywhere
  - Good performance up to 4 MPI ranks/core (*64 ranks/node*)
  - 6.7x better performance/node than Blue Gene/P
    - 1024 cores
    - 32 ranks/node
    - *No QPX yet*

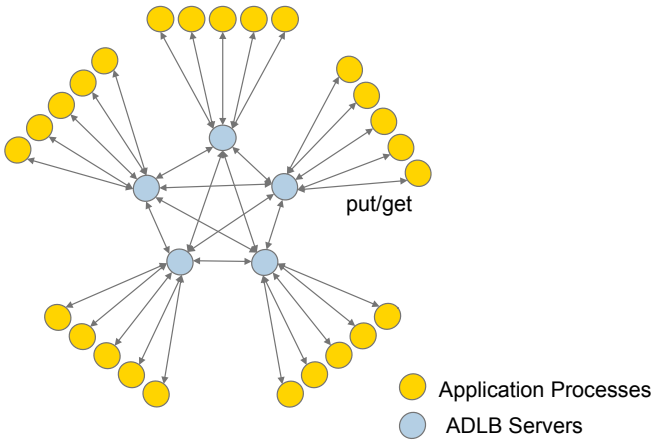
Vitali Morozov (ALCF)

$O(10^{10})$ gridpoints
100k timesteps



Ab-initio Reaction Calculations for Carbon-12

Steven Pieper (Argonne National Laboratory)
Rusty Lusk (ANL)
James Osborn (ALCF)



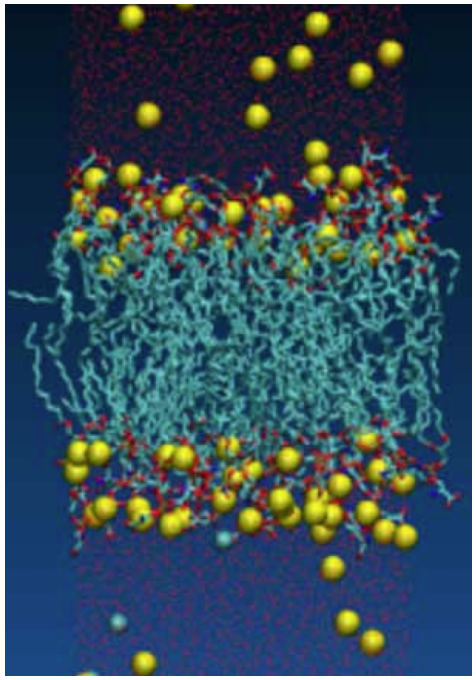
- **Calculate fundamental ^{12}C nuclear properties including density matrix**
  - Green's function monte carlo method
- **Code: GFMC**
  - Uses ADLB dynamic load balancing library
  - MPI plus OpenMP
- **Running on Blue Gene/Q EAS**
  - 10.5x better performance per node on Blue Gene/Q versus P
    - 2K cores, 8 ranks/node
  - Changed the format for matrices from dense to compressed
    - Universal improvement on BG architecture
    - Hiding latency with a number of contiguous streams: more benefits than hiding latency of individual loads.
  - Separated threading part from serial vectorizable part of kernels
    - working toward SIMD instructions within threads - a long standing issue for XL compiler
  - Tuned Q-related parameters for data sizes

Vitali Morozov (ALCF)

NAMD - The Engine for Large-Scale Classical MD Simulations of Biomolecular Systems Based on a Polarizable Force Field

Benoit Roux (U. Chicago)
Yun Luo (ESP postdoc, ALCF)
NAMD developers (Univ. of Illinois at Urbana-Champaign)

- **Next level of methods and problems incorporating new force field.**
- **Running on Blue Gene/Q**
  - Standard (non-threaded) NAMD
    - ported by researchers at IBM
- **Threading for Blue Gene/Q**
  - Threaded NAMD version developed
    - Theoretical and Computational Biophysics Group of the Beckman Institute for Advanced Science and Technology at the UIUC
  - Performance analysis underway on Blue Gene/P
    - To continue on Blue Gene/Q on available hardware

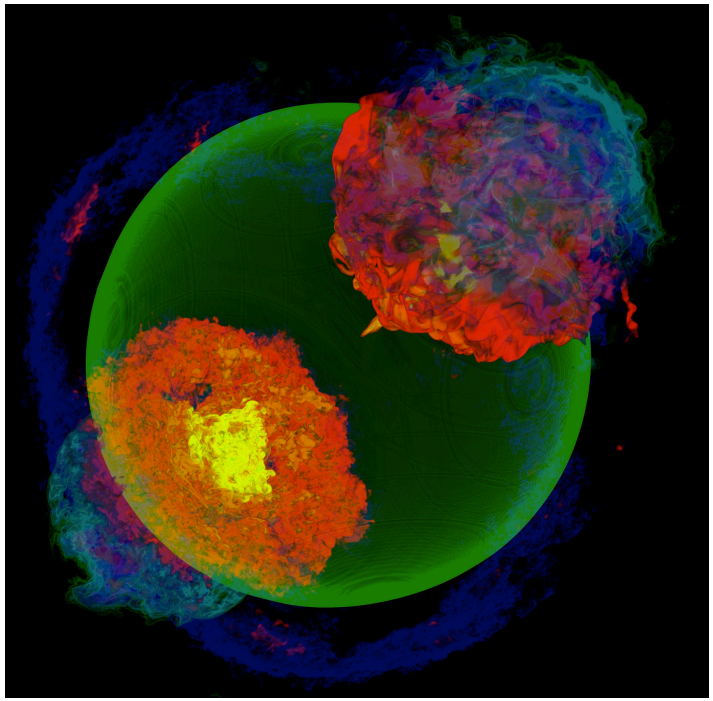


Petascale Simulations of Turbulent Nuclear Combustion

Donald Lamb (U. Chicago: Flash Center for Computational Science)
Chris Daley (U. Chicago)
Katherine Riley (ALCF)

- **Simulation of thermonuclear-powered Type Ia supernovae**
  - Deflagration
  - Buoyancy-driven turbulent burn
  - Detonation
- **Code: FLASH**
- **Development for Blue Gene/Q**
  - Threading relevant physics solvers using OpenMP
  - Performance studies of coarse (block) level threading and fine (sub-block) level threading
    - Introducing Chombo mesh package, which is threaded
  - Current FLASH version running on EAS system

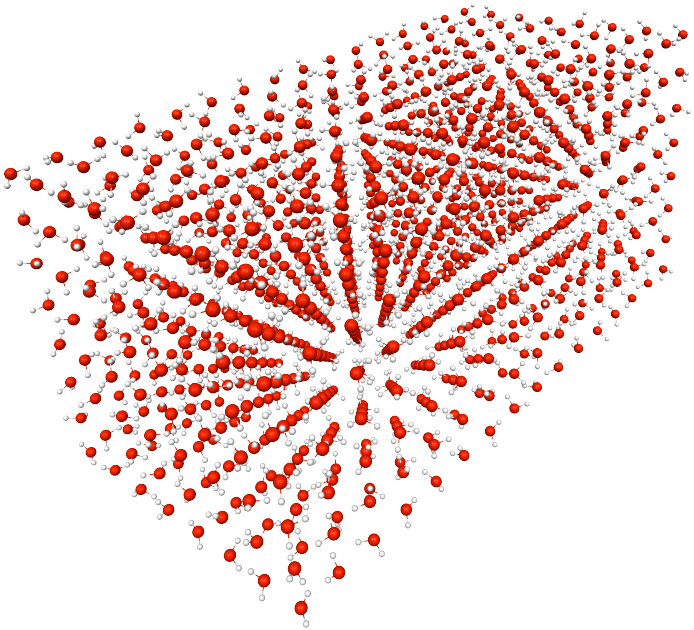
Vitali Morozov (ALCF)



High Accuracy Predictions of the Bulk Properties of Water

Mari Gordon (Iowa State U.)
Maricris Mayes (ESP postdoc, ALCF)
Graham Fletcher (ALCF)

- **Calculate bulk properties of liquid water with *ab initio* cluster simulations**
- **GAMESS code Running on Blue Gene/Q EAS**
  - Performance testing underway
    - Test case: glycine/MP2 gradient/6-31G(d,p) - 100 AO's
    - Varying processes & MPI ranks per node
    - MPI-only so far

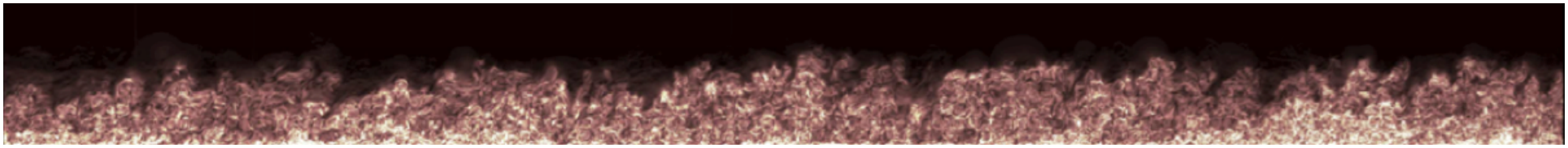


	BG/P		BG/Q					
cores	nodes	4 ranks/node	nodes	32 ranks/node	P->Q speedup/node	nodes	64 ranks/node	P->Q speedup/node
32	8	453.6	2	402	4.513432836	2	321	5.652336449
64	16	234.2	4	216.9	4.319041033	4	178	5.262921348
128	32	131.7	8	123.7	4.25869038			

Petascale Direct Numerical Simulations of Turbulent Channel Flow

Robert Moser (U. Texas)
Ramesh Balikrishnan (ALCF)

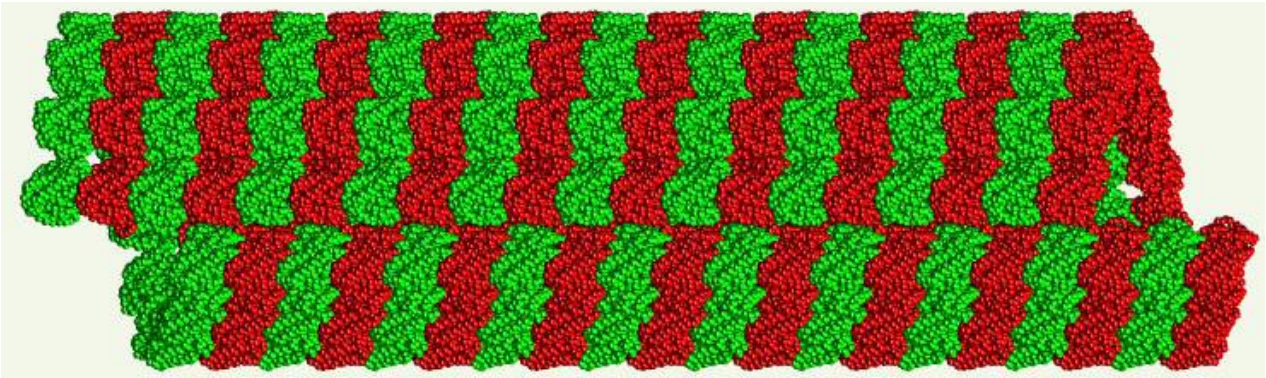
- **DNS of wall-bounded turbulence relevant to energy losses in transportation**
  - Design pipes that reduce turbulent skin friction – dominant cause of energy loss in fluid transport in pipelines
    - Proposed hi-resolution runs will resolve interaction of viscous near-wall and outer-layer turbulence with Reynolds number = 5000
  - Running code kernels on Blue Gene/Q EAS
    - - exercising XL C++ compiler with Boost
  - Tuning kernels
    - introducing multithreading with OpenMP
    - SIMD optimization



Multiscale Molecular Simulations at the Petascale

Gregory Voth (U. Chicago)
Anatole von Lilienfeld (ALCF)

- **Cellular-scale molecular modeling of biological processes.**
  - For the coarse-grained interactions studied, experimental data insufficient
  - Must be refined via proposed atomistic simulations that *Mira* enables
- **Running preliminary studies needed to set up problem to be run on Mira**
  - Multiscale coarse-graining modeling of actin filaments requires underlying atomistic simulation
    - ~100 ns all-atom trajectories calculated with NAMD
    - Leveraging NAMD Blue Gene/Q development



Tools and Libraries Project

Kalyan Kumaran (ALCF) + 32 co-PIs

- **Performance Tools**
  - PAPI
  - HPCToolkit
  - TAU
  - Scalasca
  - Open|Speedshop
  - FPMPI2
- **Debuggers**
  - DDT (Allinea)
  - TotalView (Rogue Wave)
- **Libraries**
  - FFTW, BLAS
  - PETSc
  - Parallel I/O
    - pNetCDF
    - HDF5
  - Chombo (AMR)
- **Programming Model Implementations**
  - Charm++, AMPI
  - GA Toolkit
  - CoArray Fortran
  - UPC
  - GASnet
  - MPI
- **Visualization Tools**
  - VisIt
  - ParaView

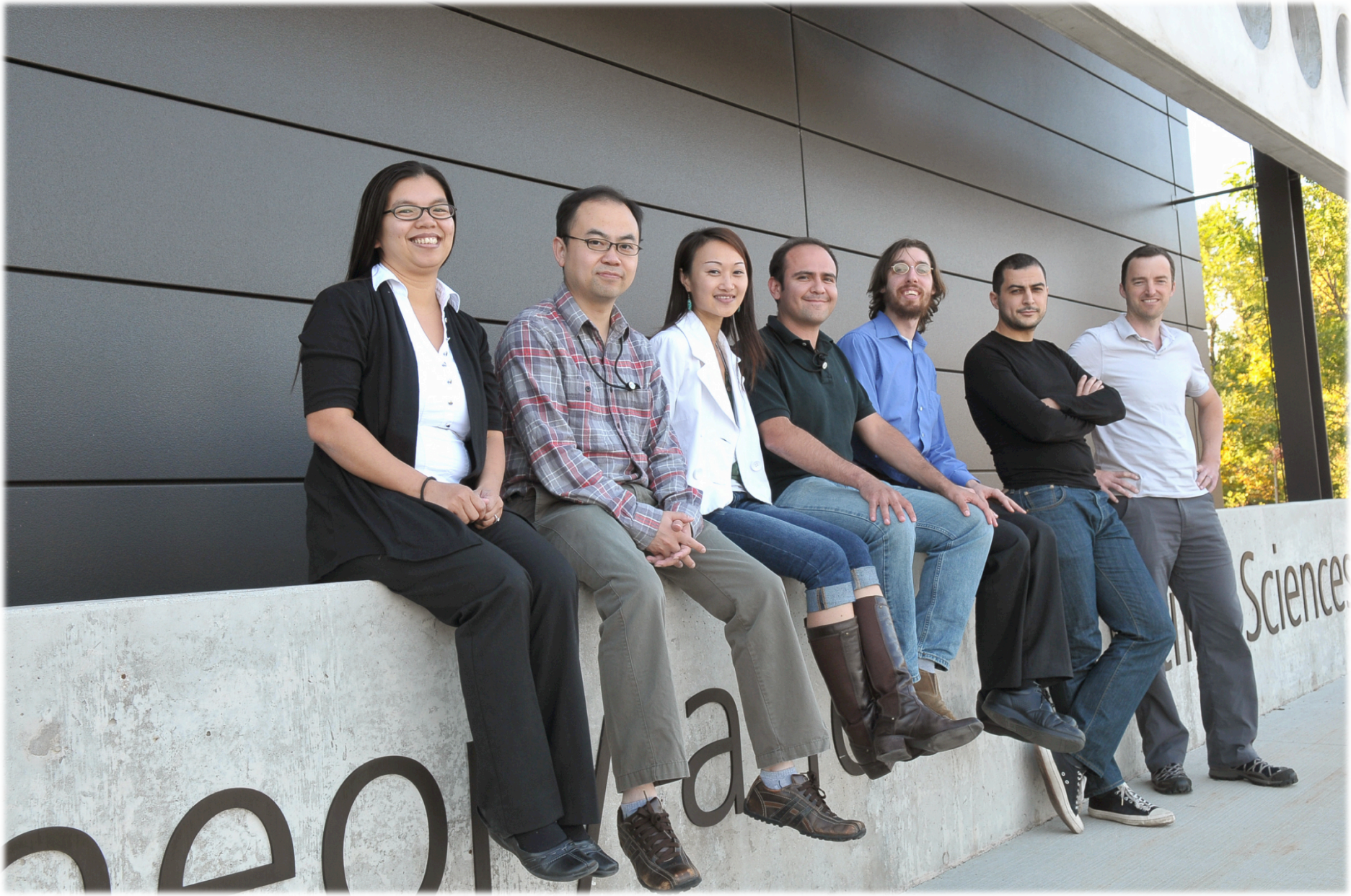
Managed by ALCF, in parallel with ESP projects

Tools and Libraries Project *(cont'd)*

- **Global Array (GA) Toolkit**
  - Worked with IBM to provide optimal support for one-sided programming models in PAMI
  - Developing new one-sided communication runtime called OSPRI (One-Sided PRImitives) as replacement for ARMCI on state-of-the-art interconnects (BGQ, PERCS, Gemini)
    - OSPRI aligned with Argonne-led MPI-3 and Unistack efforts, supports a richer set of consistency semantics oriented at application needs and optimal hardware support
    - OSPRI follows prescription for "MPI on a Million Processors," that is, eliminating O(N) algorithms and data structures
  - Reimplementing Global Arrays for hybrid programming models (thread-safety without global lock, internal multithreading, NUMA optimizations)
  - New Global Arrays will support ScaLAPACK as well as 21st-century math libraries (Elemental, PLASMA, MAGMA)

BG/Q Performance 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	Development planned
FPMPI2	UIUC	MPI	Development planned
DynInst	UMD/Wisc/IBM	Binary rewriter	In development
ValGrind	ValGrind/IBM	Memory & Thread Error Check	In development



Early Science Program Postdocs

Extra Blue Gene/Q Hardware Slides

PowerPC A2 Core Overview

- Full PowerPC compliant 64-bit CPU
- Clocked at 1600 MHz
- In order execution
- 4 hardware threads
- Dual-issue, 1 instruction per cycle from different threads
  - one instruction must be integer/load/store
  - one instruction must be floating point
- 4-wide SIMD floating point unit with complete set of parallel instructions
  - 4 FMA's @ 1600 MHz = 12.8 Gflops/core
- 16 KB L1 data cache, shared between hardware threads
- 2 KB L1P buffer, private to core, prefetch and coherency engine

40

Overview of BG/Q: Faster cores

Design Parameters	BG/P	BG/Q	Improvement
Clock Speed (GHz)	0.85	1.6	1.9x
Threads	1	4	4x
Flop / Clock / Core	4	8	2x
Instructions / Clock	2	2	--
Address width	32	64	2x
FP Registers	2x32	4x4x32	8x
INT Registers	32	4x32	4x
Flops (GF)	3.4	12.8	3.8x
In order	no	yes	--
Advanced Features	no	SE, TM	--

41