Early Science at the Argonne Leadership Computing Facility

Paul Messina
Director of Science
Argonne Leadership Computing Facility

and scores of early science participants and ALCF staff
Outline

- Overview of Mira and of BG/Q hardware, programming models
- BG/Q Tools and Libraries Project
- Status of Mira, T&D system
- Mira application benchmarks
- Overview of ESP and BG/Q Tools and Libraries Project
- Selected applications
  - Performance, enhancements, insights
- Summary
Production ALCF Hardware

- **Intrepid – IBM Blue Gene/P**
  - 40K nodes / 160K PPC cores
    - 40 racks of 1024 nodes
  - 80 terabytes of memory
  - Peak speed: 557 teraFLOPS
  - 6 PB of disk

- **Node**
  - PowerPC 450, 850 Mhz
    - 4 cores
    - 1 HW thread/core
    - 2 GB memory
    - Double-hummer
    - Peak 13.6 GFLOPS

- **Max hardware parallelism:**
  - 163,840
New ALCF Hardware: Mira

- **Mira** – IBM Blue Gene/Q system
  - 48K nodes, 768K cores
    - 48 racks of 1024 nodes
  - 786 TB of memory
  - 10 petaFLOPS
  - 35 PB of disk
  - Water cooled

- #3 Top500, #1 Graph 500, #1 Green 500 (tied), all June 2012

- **Node**
  - PowerPC A2, 1.6 GHz
    - 16 cores
    - 4 HW thread/core
    - 16 GB memory
    - QPX, dual-instruction issue
    - Peak 205 GFLOPS

- **Max hardware parallelism:**
  - 3,145,728
Mira: A 10PF Computational Science Platform

**Mira**

**BG/Q Compute**
Mira: Latin: to wonder at, wonderful; causing one to smile

**BG/Q IO**

**Data Storage**

**Viz & Data Analytics**

**Configuration**
- **BG/Q**
  - 48 racks
  - 48K 1.6 GHz nodes
  - 768K cores & 786TB RAM
  - 384 I/O nodes
- **Storage**
  - 240 GB/s, 35 PB
## Overview of BG/Q vs. BG/P

<table>
<thead>
<tr>
<th>Design Parameters</th>
<th>BG/P</th>
<th>BG/Q</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores / Node</td>
<td>4</td>
<td>16</td>
<td>4x</td>
</tr>
<tr>
<td>Clock Speed (GHz)</td>
<td>0.85</td>
<td>1.6</td>
<td>1.9x</td>
</tr>
<tr>
<td>Flop / Clock / Core</td>
<td>4</td>
<td>8</td>
<td>2x</td>
</tr>
<tr>
<td>Nodes / Rack</td>
<td>1,024</td>
<td>1,024</td>
<td>--</td>
</tr>
<tr>
<td>RAM / core (GB)</td>
<td>0.5</td>
<td>1</td>
<td>2x</td>
</tr>
<tr>
<td>Flops / Node (GF)</td>
<td>13.6</td>
<td>204.8</td>
<td>15x</td>
</tr>
<tr>
<td>Memory BW/Node (GB/sec)</td>
<td>13.6</td>
<td>42.6</td>
<td>3x</td>
</tr>
<tr>
<td>Latency (MPI zero-length, nearest-neighbor node)</td>
<td>2.6 μs</td>
<td>2.2 μs</td>
<td>~15% less</td>
</tr>
<tr>
<td>Latency (MPI 64 KB message)</td>
<td>181μs</td>
<td>48 μs</td>
<td>~74% less</td>
</tr>
<tr>
<td>Network Interconnect</td>
<td>3D torus</td>
<td>5D torus</td>
<td>Smaller diameter</td>
</tr>
<tr>
<td>Concurrency / Rack</td>
<td>4,096</td>
<td>65,536</td>
<td>16x</td>
</tr>
<tr>
<td>GFlops/Watt</td>
<td>0.77</td>
<td>2.10</td>
<td>3x</td>
</tr>
</tbody>
</table>
Programming and Running on BG/Q

- **MPI**
- **Threads**: OpenMP, PTHREADS
- **QPX intrinsics**: vec_ld, vec_add, vec_madd, ....
- **Topology interfaces**
  - E.g. MPIX_* functions
- **Run modes: combinations of**
  - MPI ranks/node = \{1,2,4,...,64\}
  - Threads/node = \{1,2,4,...,64\}
BG/P applications should run, *unchanged*, on BG/Q — *faster*
Ten Mira Science Application Benchmarks used for acceptance

- Applications cannot be manually tuned; only compiler optimizations are allowed.
- 3 of the applications are threaded – i.e., use both OpenMP and MPI (GTC, GFMC, GFDL).
- The remainder are 100% MPI applications (DNS3D, FLASH, GPAW, LS3DF, MILC, NAMD & NEK 5000).
- For 100% MPI applications, we tested multiple MPI ranks per core (max of 4 ranks per core).
- For MPI + OpenMP applications, we tested 1 MPI rank per core and multiple OpenMP threads per core (max of 4 threads per core).
## Node/node comparison details

<table>
<thead>
<tr>
<th></th>
<th>BG/P</th>
<th>BG/Q</th>
<th>RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores / Node</td>
<td>4</td>
<td>16</td>
<td>4x</td>
</tr>
<tr>
<td>Clock Speed (GHz)</td>
<td>0.85</td>
<td>1.6</td>
<td>1.9x</td>
</tr>
<tr>
<td>Flop / Clock / Core</td>
<td>4</td>
<td>8</td>
<td>2x</td>
</tr>
<tr>
<td>Flops/core</td>
<td>3.4 GF</td>
<td>12.8 GF</td>
<td>3.8x</td>
</tr>
<tr>
<td>Flops/node</td>
<td>13.6</td>
<td>204.8</td>
<td>15.1x</td>
</tr>
<tr>
<td>Nodes / Rack</td>
<td>1,024</td>
<td>1,024</td>
<td>1x</td>
</tr>
<tr>
<td>Flops / Rack</td>
<td>13.9 TF</td>
<td>210 TF</td>
<td>15.1x</td>
</tr>
</tbody>
</table>
# Mira Science Applications

**BG/P version as is on BG/Q**

<table>
<thead>
<tr>
<th>Apps</th>
<th>BQ/P Ratio</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNS3D</td>
<td>11.8</td>
<td>2048(^3) grid, 16K cores, 64 ranks/node</td>
</tr>
<tr>
<td>FLASH</td>
<td>5.9 (9.1)</td>
<td>rtflame, 2K cores, 64 ranks/node, 8 threads/rank, no MPI-IO</td>
</tr>
<tr>
<td>GFMC</td>
<td>10.5</td>
<td>c12-test, 2K cores, 8 ranks/node, 8 thrds/rank</td>
</tr>
<tr>
<td>GTC</td>
<td>10.8</td>
<td>M0360, 8K cores, 16 ranks/node, 4 thrds/rank</td>
</tr>
<tr>
<td>GFDL</td>
<td>11.9</td>
<td>Atm, 2K cores, 16 ranks/node, 4 thrds/rank</td>
</tr>
<tr>
<td>MILC</td>
<td>6.1</td>
<td>32(^3)x64 lattice, 2K cores, 64 ranks/node, no QPX</td>
</tr>
<tr>
<td>NEK</td>
<td>7.3</td>
<td>med case, 1K cores, 32 ranks/node, no QPX</td>
</tr>
<tr>
<td>NAMD</td>
<td>9.7</td>
<td>ATPase bmk, 2K cores, 16 ranks/node</td>
</tr>
<tr>
<td>GPAW</td>
<td>7.6</td>
<td>Au(_{\text{bulk}})5x5x5, 2K cores, 16 ranks/node</td>
</tr>
<tr>
<td>LS3DF</td>
<td>8.1</td>
<td>ZnO(_{\text{Te}}), 8K cores, ESSLsmp, I/O sensitive</td>
</tr>
</tbody>
</table>
## Additional BG/Q Programming Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Source</th>
<th>Provides</th>
<th>Q Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Arrays Toolkit</td>
<td>PNNL</td>
<td>One sided communication and high level library for array computation</td>
<td>Available (MPI), PAMI pending</td>
</tr>
<tr>
<td>Berkeley UPC</td>
<td>Berkeley</td>
<td>PGAS variant of C</td>
<td>Ported with XLC and gcc</td>
</tr>
<tr>
<td>GASnet</td>
<td>Berkeley</td>
<td>Active message communication</td>
<td>Available, PAMI and MPI ports</td>
</tr>
<tr>
<td>Charm++</td>
<td>UIUC</td>
<td>Dynamic load balancing and task virtualization</td>
<td>Available, PAMI and MPI ports</td>
</tr>
<tr>
<td>CoArray Fortran</td>
<td>Rice U.</td>
<td>PGAS variant of Fortran</td>
<td>Debugging. Uses GASnet</td>
</tr>
</tbody>
</table>
## Mira Performance Tools

<table>
<thead>
<tr>
<th>Tool Name</th>
<th>Source</th>
<th>Provides</th>
<th>Mira Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGPM</td>
<td>IBM</td>
<td>HPC</td>
<td>Available</td>
</tr>
<tr>
<td>gprof</td>
<td>GNU/IBM</td>
<td>Timing (sample)</td>
<td>Available</td>
</tr>
<tr>
<td>TAU</td>
<td>Unv. Oregon</td>
<td>Timing (inst), MPI</td>
<td>Available</td>
</tr>
<tr>
<td>Rice HPCToolkit</td>
<td>Rice Univ.</td>
<td>Timing (sample), HPC (sample)</td>
<td>Available</td>
</tr>
<tr>
<td>mpiP</td>
<td>LLNL</td>
<td>MPI</td>
<td>Available</td>
</tr>
<tr>
<td>PAPI</td>
<td>UTK</td>
<td>HPC API</td>
<td>Available</td>
</tr>
<tr>
<td>Darshan</td>
<td>ANL</td>
<td>IO</td>
<td>Available</td>
</tr>
<tr>
<td>Open</td>
<td>Speedshop</td>
<td>Krell</td>
<td>Available</td>
</tr>
<tr>
<td>Scalasca</td>
<td>Juelich</td>
<td>Timing (inst), MPI</td>
<td>Available</td>
</tr>
<tr>
<td>Jumpshot</td>
<td>ANL</td>
<td>MPI</td>
<td>Available</td>
</tr>
<tr>
<td>DynInst</td>
<td>UMD/Wisc/IBM</td>
<td>Binary rewriter</td>
<td>In development</td>
</tr>
<tr>
<td>ValGrind</td>
<td>ValGrind/IBM</td>
<td>Memory &amp; Thread Error Check</td>
<td>In development</td>
</tr>
</tbody>
</table>
## Mira Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Source</th>
<th>Provides</th>
<th>Mira Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSL</td>
<td>IBM</td>
<td>Dense Linear Algebra &amp; FFT Kernels</td>
<td>5.1.1-0 beta version</td>
</tr>
<tr>
<td>BLAS</td>
<td>NETLIB (UTK) &amp; ESSL</td>
<td>Dense linear algebra kernels</td>
<td>Available, based on ESSL GEMM</td>
</tr>
<tr>
<td>BLIS</td>
<td>U. Texas &amp; ANL</td>
<td>Framework for GotoBLAS successor</td>
<td>In development</td>
</tr>
<tr>
<td>CBLAS</td>
<td>UTK</td>
<td>C wrappers to BLAS</td>
<td>Available</td>
</tr>
<tr>
<td>LAPACK</td>
<td>UTK</td>
<td>Dense linear algebra solver</td>
<td>3.4.1</td>
</tr>
<tr>
<td>SCALAPACK</td>
<td>UTK</td>
<td>Parallel dense linear algebra solver</td>
<td>2.0.2</td>
</tr>
<tr>
<td>ARPACK &amp; PARPACK</td>
<td>Rice U.</td>
<td>Eigenvalues &amp; eigenvectors</td>
<td>2.1</td>
</tr>
<tr>
<td>FFTW</td>
<td>MIT</td>
<td>Fast Fourier transform</td>
<td>2.1.5, 3.3.1, no Q hand tuning yet</td>
</tr>
<tr>
<td>METIS</td>
<td>UMN</td>
<td>Graph partitioning (for meshes)</td>
<td>5.0.2</td>
</tr>
<tr>
<td>ParMETIS</td>
<td>UMN</td>
<td>Graph partitioning (for meshes)</td>
<td>4.0.2</td>
</tr>
<tr>
<td>Boost</td>
<td>OSS community</td>
<td>C++ data structures &amp; algorithms</td>
<td>1.49.0</td>
</tr>
<tr>
<td>HDF5</td>
<td>NCSA</td>
<td>High-level I/O</td>
<td>1.8.8</td>
</tr>
<tr>
<td>NetCDF</td>
<td>UCAR</td>
<td>High-level I/O</td>
<td>4.1.3</td>
</tr>
<tr>
<td>P3DFFT</td>
<td>SDSC</td>
<td>Parallel FFT</td>
<td>2.4 (patched by ANL)</td>
</tr>
<tr>
<td>PETSc</td>
<td>ANL</td>
<td>Parallel solvers</td>
<td>Available, 3.3-p1, initial PTHREADs</td>
</tr>
</tbody>
</table>
First in *Mira* Queue: Early Science Program

- 2 billion core-hours on Mira for science runs between *Mira* acceptance and start of production

http://esp.alcf.anl.gov
The ALCF Early Science Program (ESP)

- The ESP was launched to
  - prepare key applications for the architecture and scale of Mira and
  - solidify libraries and infrastructure

- 16 Early Science projects were chosen based on computational and scientific reviews.

- The projects have a running start for delivery of exciting new science
  - Postdocs were assigned to the projects
  - ALCF staff has held three workshops for the ESP projects
  - Before Mira was available, development was done through
    - Time on Intrepid
    - Some access to BG/Q systems at IBM
    - Time on Vesta, a 1-rack BG/Q Test & Development system at ALCF
    - Some access to Sequoia BG/Q system at LLNL
## Profile of the ESP Projects

### Science Areas
- Astrophysics
- Biology
- CFD/Aerodynamics
- Chemistry
- Climate
- Combustion
- Cosmology
- Energy
- Fusion Plasma
- Geophysics
- Materials
- Nuclear Structure

### Algorithms/Methods
- Structured Grids
- Unstructured Grids
- FFT
- Dense Linear Algebra
- Sparse Linear Algebra
- Particles/N-Body
- Monte Carlo

### National Lab PIs
- 7 National Lab PIs

### University PIs
- 9 University PIs

---

**Leap to Petascale**
Making the Move Towards Mira
May 22-25 - Argonne National Laboratory
Timeline for Mira availability

ESP access to Mira

- ESP projects will have access to Mira late summer 2012
  - access to 1-rack BG/Q system since March, some access to racks at IBM
- An ESP project completed a 16-rack science run late July, and a full-machine performance run on August 3rd
- Currently in preparation for acceptance phase
- Early science runs likely to start second half CY2012

Plan for 2013 INCITE time on Mira

- Mira is committed to go live for INCITE on October 1, 2013 with 768M core-hours for allocation
- Start date of production INCITE time is likely to happen earlier in CY 2013
- Guidance for INCITE 2013
  - Propose science based on a 3B core-hour pool, 100-300M per project
  - 2013 INCITE Allocation scenarios
    - 768M
    - 2B core-hours
    - 3B core-hours

2012

Q1 2012: VEAS and T&D

2013

Q4 2013: INCITE
### All ESP Projects are running on BG/Q (1)

<table>
<thead>
<tr>
<th>Project Title</th>
<th>PI/affiliation</th>
<th>Code(s)</th>
<th>Runs on BG/Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate-Weather Modeling Studies Using a Prototype Global Cloud-System Resolving Model</td>
<td>Balaji/GFDL</td>
<td>HIRAM</td>
<td>✔</td>
</tr>
<tr>
<td>First Principles Calculations of Interfaces in Electrical Energy Storage Systems</td>
<td>Curtiss/ANL</td>
<td>QMCPACK</td>
<td>✔</td>
</tr>
<tr>
<td>Direct Numerical Simulation of Autoignition in a Jet in a Cross-Flow</td>
<td>Frouzakis/Swiss Fed Inst of Technology</td>
<td>Nek5000</td>
<td>✔</td>
</tr>
<tr>
<td>High Accuracy Predictions of the Bulk Properties of Water</td>
<td>Gordon/Iowa State</td>
<td>GAMESS</td>
<td>✔</td>
</tr>
<tr>
<td>Cosmic Structure Probes of the Dark Universe</td>
<td>Habib/ANL, UC</td>
<td>HACC</td>
<td>✔</td>
</tr>
<tr>
<td>Accurate Numerical Simulations Of Chemical Phenomena Involved in Energy Production and Storage with MADNESS and MPQC</td>
<td>Harrison/ORNL</td>
<td>MADNESS &amp; MPQC</td>
<td>✔</td>
</tr>
<tr>
<td>Petascale, Adaptive CFD</td>
<td>Jansen/U Colorado</td>
<td>PHASTA</td>
<td>✔</td>
</tr>
<tr>
<td>Using Multi-scale Dynamic Rupture Models to Improve Ground Motion Estimates</td>
<td>Jordan/USC</td>
<td>AWP-ODC, SORD</td>
<td>✔</td>
</tr>
</tbody>
</table>
All ESP Projects are running on BG/Q (2)

<table>
<thead>
<tr>
<th>Project Title</th>
<th>PI/affiliation</th>
<th>Code(s)</th>
<th>Runs on BG/Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-Speed Combustion and Detonation</td>
<td>Khoklov/UC</td>
<td>HSCD</td>
<td>✔</td>
</tr>
<tr>
<td>Petascale Simulations of Turbulent Nuclear Combustion</td>
<td>Lamb/UC</td>
<td>FLASH/RTFlame</td>
<td>✔</td>
</tr>
<tr>
<td>Lattice Quantum Chromodynamics</td>
<td>Mackenzie/Fermilab</td>
<td>MILC, Chroma, CPS</td>
<td>✔</td>
</tr>
<tr>
<td>Petascale Direct Numerical Simulations of Turbulent Channel Flow</td>
<td>Moser/UTexas</td>
<td>PSDNS</td>
<td>✔</td>
</tr>
<tr>
<td>Ab-initio Reaction Calculations for Carbon-12</td>
<td>Pieper/ANL</td>
<td>GFMC</td>
<td>✔</td>
</tr>
<tr>
<td>NAMD - The Engine for Large-Scale Classical MD Simulations of Biomolecular Systems Based on a Polarizable Force Field</td>
<td>Roux/UC</td>
<td>NAMD, Charm++</td>
<td>✔</td>
</tr>
<tr>
<td>Global Simulation of Plasma Microturbulence at the Petascale &amp; Beyond</td>
<td>Tang/Princeton</td>
<td>GTC-P</td>
<td>✔</td>
</tr>
<tr>
<td>Multiscale Molecular Simulations at the Petascale</td>
<td>Voth/UC, ANL</td>
<td>NAMD, LAMMPS, Raptor</td>
<td>✔</td>
</tr>
</tbody>
</table>
All ESP applications are running on BG/Q --

How much effort did it take?

In the next two slides we use S = small, M = medium, L = Large as measures, characterized as

- S = zero - few days of effort, modifications to 0 - 3% of existing lines of code
- M = a few weeks of effort, modifications to 3% - 10% of existing lines of code
- L = a few months of effort, changes to beyond 10% of existing lines of code

The ranking is based on estimates by the people who actually did the work
## All ESP Projects are running on BG/Q (1)

<table>
<thead>
<tr>
<th>PI/affiliation</th>
<th>Code(s)</th>
<th>Runs on BG/Q</th>
<th>Magnitude of changes</th>
<th>Nature of Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balaji/GFDL</td>
<td>HIRAM</td>
<td>✔</td>
<td>L</td>
<td>Improve OpenMP implementation, reformulate divergence-damping</td>
</tr>
<tr>
<td>Curtiss/ANL</td>
<td>QMCPACK</td>
<td>✔</td>
<td>M</td>
<td>S to port, L to use QPX in key kernels; plan: nested OpenMP</td>
</tr>
<tr>
<td>Frouzakis/Swiss Fed Inst of Technology</td>
<td>Nek5000</td>
<td>✔</td>
<td>S</td>
<td>Optimized small matrix-matrix multiply using QPX</td>
</tr>
<tr>
<td>Gordon/Iowa State</td>
<td>GAMESS</td>
<td>✔</td>
<td>M</td>
<td>64-bit addressing, thread integral kernels with OpenMP</td>
</tr>
<tr>
<td>Habib/ANL, UC</td>
<td>HACC</td>
<td>✔</td>
<td>M</td>
<td>Short-range-force only: tree code</td>
</tr>
<tr>
<td>Harrison/ORNL</td>
<td>MADNESS</td>
<td>✔</td>
<td>S</td>
<td>Threading runtime tuning Kernel tuning to use QPX</td>
</tr>
<tr>
<td>Jansen/U Colorado</td>
<td>PHASTA</td>
<td>✔</td>
<td>S</td>
<td>Unchanged MPI-only performs well; OpenMP threaded in testing</td>
</tr>
<tr>
<td>Jordan/USC</td>
<td>AWP-ODC, SORD</td>
<td>✔</td>
<td>S, M</td>
<td>None, Threading</td>
</tr>
</tbody>
</table>
## All ESP Projects are running on BG/Q (2)

<table>
<thead>
<tr>
<th>PI/affiliation</th>
<th>Code(s)</th>
<th>Runs on BG/Q</th>
<th>Magnitude of changes</th>
<th>Nature of Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khoklov/UC</td>
<td>HSCD</td>
<td>✔</td>
<td>S</td>
<td>Tune OpenMP parameters, link optimized math libs</td>
</tr>
<tr>
<td>Lamb/UC</td>
<td>FLASH/RTFlame</td>
<td>✔</td>
<td>S</td>
<td>OpenMP threading</td>
</tr>
<tr>
<td>Mackenzie/Fermilab</td>
<td>MILC, Chroma, CPS</td>
<td>✔</td>
<td>L</td>
<td>Full threading, QPX intrinsics/assembler, kernel on SPI comm.</td>
</tr>
<tr>
<td>Moser/UTexas</td>
<td>PSDNS</td>
<td>✔</td>
<td>S</td>
<td>Compile dependency libs, add OpenMP directives for threading</td>
</tr>
<tr>
<td>Pieper/ANL</td>
<td>GFMC</td>
<td>✔</td>
<td>S</td>
<td>Tune no. threads &amp; ranks.</td>
</tr>
<tr>
<td>Roux/UC</td>
<td>NAMD, Charm++</td>
<td>✔</td>
<td>L</td>
<td>Threads, PAMI implementation of Charm++</td>
</tr>
<tr>
<td>Tang/Princeton</td>
<td>GTC</td>
<td>✔</td>
<td>S</td>
<td>Improve OpenMP implementation</td>
</tr>
<tr>
<td>Voth/UC, ANL</td>
<td>NAMD, LAMMPS, RAPTOR</td>
<td>✔</td>
<td>M</td>
<td>OpenMP threads &amp; serial optimizations in RAPTOR/LAMMPS</td>
</tr>
</tbody>
</table>
Status of selected ESP projects
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
HACC on Mira

- **Hardware/Hybrid Accelerated Cosmology Code Framework**
  - Hybrid particle/grid solver for the cosmological Vlasov-Poisson system
    - Designed using practical co-design principles and associated algorithms
    - Targeted for very high performance on the largest science problems
    - Runs at scale on all current hardware (BG, CPU, CPU/GPU, CPU/MIC, --)
    - Novel tunable algorithmic structure; multiple programming paradigms

- **The Cosmological Structure Formation Problem**
  - Very high global dynamic range in space, time, and mass
    - Global spatial dynamic range requirement of 1,000,000:1
    - Mass resolution dynamic range of 100,000:1
    - Future surveys (e.g., LSST) demand multi-trillion particle simulations

- **HACC on the BG/Q**
  - Algorithms, implementations, and optimizations for the BG/Q
    - Split force computations into 3 domains, close-, medium-, and long-range
      - (i) Close -- direct Particle-Particle (PP), (ii) Medium -- Recursive Coordinate Bisection (RCB) Tree, (iii) Long -- Spectral Particle-Mesh (SPM)
    - MPI+OpenMP model, QPX intrinsics, very high-performance force kernel
    - **Gordon Bell Finalist**: Performance at >50% of peak with >90% parallel efficiency up to 786,432 cores, multi-trillion particles at full dynamic range
HACC on Mira: Early Science Test Run

- **First Large-Scale Simulation on Mira**
  - 16 racks, 262,144 cores (one-third of Mira)
  - 14 hours continuous, no checkpoint restarts

- **Simulation Parameters**
  - 9.14 Gpc simulation box
  - 7 kpc force resolution
  - 1.07 trillion particles

- **Science Motivations**
  - Resolve halos that host luminous red galaxies for analysis of SDSS observations
  - Cluster cosmology
  - Baryon acoustic oscillations in galaxy clustering

Zoom-in visualization of the density field illustrating the global spatial dynamic range of the simulation -- approximately a million-to-one
Early Science High Speed Combustion and Detonation Project

- Direct Numerical Simulation of the deflagration-to-detonation transition (DDT) in hydrogen-oxygen gaseous mixtures for hydrogen safety

- PI: Alexei Khokhlov, University of Chicago
- Joanna Austin, University of Illinois
- Charles Bacon, Argonne National Laboratory
- Andrew Knisely, University of Illinois
- Ben Clifford, Argonne National Laboratory
- Joe Bernstein, Argonne National Laboratory
Code features

- 3-d reactive flow Navier-Stokes with 8-species and 19 reaction kinetics H2-O2 burning, multi-species NASA7 equation of state, multi-species temperature dependent viscosity, mass and heat conduction, and radiative cooling
- Adaptive mesh refinement on a regular rectangular grid
Reflected shock tube validation
# Single-node scaling on Q

<table>
<thead>
<tr>
<th>Thread count</th>
<th>Time per step</th>
<th>Efficiency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>349</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>190</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>110</td>
<td>79</td>
</tr>
<tr>
<td>8</td>
<td>70 (68)</td>
<td>62 (64)</td>
</tr>
<tr>
<td>16</td>
<td>64 (48)</td>
<td>34 (45)</td>
</tr>
<tr>
<td>32</td>
<td>45 (40)</td>
<td>24 (27)</td>
</tr>
<tr>
<td>64</td>
<td>50 (41)</td>
<td>11 (13)</td>
</tr>
</tbody>
</table>

Parenthetical numbers come after increasing the size of the array of cells passed to the work functions – high rank counts were getting not enough work per thread from the original setting.
Main loop times

- BG/P -> BG/Q speedup = 2.5x/core, 9.2x/node

<table>
<thead>
<tr>
<th>BG/Q Node count</th>
<th>Time</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>705</td>
<td>100</td>
</tr>
<tr>
<td>256</td>
<td>358</td>
<td>98.5</td>
</tr>
<tr>
<td>512</td>
<td>188</td>
<td>93.8</td>
</tr>
<tr>
<td>1024</td>
<td>101</td>
<td>87.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BG/P Node count</th>
<th>Time</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>1779</td>
<td>100</td>
</tr>
<tr>
<td>1024</td>
<td>887</td>
<td>100</td>
</tr>
<tr>
<td>2048</td>
<td>488</td>
<td>91.1</td>
</tr>
<tr>
<td>4096</td>
<td>250</td>
<td>89</td>
</tr>
</tbody>
</table>
FLASH Overview

*PI: Don Lamb, University of Chicago*

*Postdoc: Christopher Daley*

- FLASH simulates problems from astrophysics, cosmology, HEDP, incompressible fluid dynamics
- Collection of code units that a user assembles into a custom application
- Portable and scalable
- ½ million lines of code
  - Fortran90 and C
  - MPI and (recently) OpenMP
  - AMR
  - Parallel I/O
FLASH Early Science Goals

- Improve understanding of type 1a supernova
- The early science applications include
  - RTFlame (flame in a rectilinear domain with constant gravity)
  - White Dwarf (full supernova problem)
RTFlame: FLASH performance on BG/Q

Best overall compromise:
16 MPI ranks/node
4 threads/MPI rank
White Dwarf: Weak scaling on BG/Q

Source: Chris Daley, postdoc for FLASH ESP
White Dwarf: Weak scaling on BG/Q summary

- Good weak scaling on Intrepid BG/P (up to 131K MPI ranks) and Vesta BG/Q (up to 16K MPI ranks - whole machine)
  - Hard to create perfect weak scaling adaptive grid tests
    - Oscillations in evolution time happen because the work per MPI rank does not stay exactly constant

- BG/Q to BG/P node-to-node ratio between 7.4x and 7.9x
  - Less than RTFlame node-to-node ratio
    - White Dwarf application includes additional physics and so more guard cell fills per time step
    - Does not include multipole solver multithreading - removed to avoid a strange crash during access of threadprivate data

Source: Chris Daley, postdoc for FLASH ESP
GAMESS
(General Atomic and Molecular Electronic Structure System)

- **Ab initio** quantum chemistry package
- Maintained by the research group of Prof. Mark Gordon at Iowa State
  - ESP Postdoc: Maricris Lodriguito Mayes
  - ALCF catalyst: Graham D. Fletcher
- Implements all major quantum mechanical methods
  - Hartree-Fock
  - Møller Plesset perturbation theory
  - Coupled-cluster
  - Multiconfiguration consistent field
  - Configuration interaction
  - Density functional theory
- Ported to all major architectures
- Free and widely used on everything from laptops to supercomputers
- About a million lines of code, with an associated parallelization library comprising 15,000 lines
- Highly scalable, including many distributed data algorithms
The Performance of FMO2-MP2/aug-cc-pvdz force calculations for several water clusters. (All timings are in minutes)

<table>
<thead>
<tr>
<th>Waters</th>
<th>Blue Gene/P</th>
<th>Blue Gene/Q*</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>4.8</td>
<td>12.0</td>
</tr>
<tr>
<td>256</td>
<td>10.5</td>
<td>31.1</td>
</tr>
<tr>
<td>512</td>
<td>28.9</td>
<td>94.2</td>
</tr>
</tbody>
</table>

* Using -O2 IBM XL compiler option and 4 ranks per node

- Even without tuning and optimizing GAMESS, a Q/P speed-up per node is about 4.6 to 6.4 times.
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
  - Threaded & nonthreaded running on BG/Q
  - Underlying Charm++ framework ported
    - MPI implementation
    - PAMI implementation
    - NAMD using PAMI Charm++ 20% faster (than MPI)
Tokamak Plasma Microturbulence

Stephane Ethier (PPPL)
Bei Wang (Princeton U.)

- **Code: GTC-P**
- **Particle-in-cell simulation of plasma**
  - Study energy loss through turbulence
  - Validate key assumption about scaling in ITER
Tokamak Plasma Microturbulence

- 2D data decomposition for MPI tasks
  - Zeta (long way round torus)
    - Equal number of grid points
  - Radially in poloidal plane
    - Equal number of particles
    - Guard layers must overlap by max gyroradius

- OpenMP threads
  - Particle loops

- ~10x better performance per node (BG/Q vs BG/P)

- Tang: “Such reduction in time to solution will enable meaningful scientific progress.”

ITER plasmas
- \( O(10^{10}) \) particles
- \( O(10^8) \) grid cells
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)

QMCPACK on Blue Gene/Q
- Operations depend on type of wave function: LCAO, real-space, PWs.
  - Spline interpolation
  - Small DGEMM and DGEMV
- Reformulating loops to use BLAS2+3
- Adding QPX intrinsics
- Add nested parallelism to MCWalker evaluation
  *(needs OpenMP ≥ 3.0)*
  - Good efficiency going from 1 to 2 OpenMP threads
  - Going from 2 to 4 threads: 1.5x speedup

IBM Zurich optimizing CPMD for Blue Gene/Q
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**
- **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
  - Full hybrid Monte Carlo evolution now running on BG/Q
    - Dirac matrix solver kernel:
      - 60% of theoretical peak on BG/Q chip
      - >80% of theoretical peak communication bandwidth BG/Q chip
        » Using low level SPI communications layer
  - Dirac solver kernel using pthreads
  - OpenMP used in rest of code
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
- 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
### Timeline for Mira availability

<table>
<thead>
<tr>
<th>ESP access to Mira</th>
<th>Plan for 2013 INCITE time on Mira</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESP projects will have access to Mira late summer 2012</td>
<td></td>
</tr>
<tr>
<td>- access to 1-rack BG/Q system since March, some access to racks at IBM</td>
<td></td>
</tr>
<tr>
<td>An ESP project completed full-machine performance run on DATE XX, 16-rack science run on DATE YY</td>
<td></td>
</tr>
<tr>
<td>Currently in preparation for acceptance phase</td>
<td></td>
</tr>
<tr>
<td>Early science runs likely to start second half CY2012</td>
<td></td>
</tr>
<tr>
<td>Mira is committed to go live for INCITE on October 1, 2013 with 768M core-hours for allocation</td>
<td></td>
</tr>
<tr>
<td>Start date of production INCITE time is likely to happen earlier in CY 2013</td>
<td></td>
</tr>
<tr>
<td>Guidance for INCITE 2013</td>
<td></td>
</tr>
<tr>
<td>- Propose science based on a 3B core-hour pool, 100-300M per project</td>
<td></td>
</tr>
<tr>
<td>- 2013 INCITE Allocation scenarios</td>
<td></td>
</tr>
</tbody>
</table>

**2012**
- Q1 2012: VEAS and T&D
- Possible Mira

**2013**
- Q1 2013: ESP Mira Access
- Q4 2013: Possible INCITE

**Possible INCITE**
Summary

- Early experience confirms that Mira will enable advances in a broad spectrum of applications
- The Early Science Program is paying off
  - All the applications are running
  - Many valuable insights on tuning and scaling are being obtained
  - Applications are being enhanced to model more complex phenomena, with higher fidelity
- The BG/Q Tools and Libraries Project has yielded substantial software tools very early in the life of Mira
- We look forward to exciting scientific results in the next few months