Introduction to the Blue Waters Project

Dr. William Kramer
NCSA/University of Illinois
Deputy Director of Blue Waters
wkramer@ncsa.uiuc.edu/ - http://www.ncsa.uiuc.edu/BlueWaters
(217) 333-6260/(217) 979-7577
Petascale computing will enable advances in a broad range of science and engineering disciplines:

Molecular Science

Weather & Climate Forecasting

Astronomy

Earth Science

Health
What Users Want From Petascale Systems

• Performance
  • How fast will the system process work if everything is working well

• Effectiveness
  • The likelihood clients can get the system to do their work when they need it

• Reliability
  • The likelihood the system is available to do the work

• Consistency
  • How often will the system process the same or
Goals of Blue Waters Project

- **Science and Engineering**
  - Provide knowledge/expertise/services to help researchers develop applications that take full advantage of Blue Waters

- **Computing System Hardware and Software**
  - Sustain $\geq 1$ petaflops on range of science and engineering applications
  - Enhance petascale applications development environment and systems software

- **Education**
  - Prepare next generation of scientists and engineers for research at the frontiers of petascale computing and computation

- **Industrial Engagement**
  - Enable industry to utilize petascale computing to address their most challenging problems and enhance their competitive position
Watch the Word **Sustained**

- The supercomputing community unfortunately often uses *peak* performance to measure a system’s processing power.
- Peak is like buying a car based solely on the speedometer’s top speed—the car can’t reach it and you can’t use it.
  - Linpack is like measuring a car based on its NASCAR results – highly unrealistic for most except maybe a few moments during a vacation in Las Vegas
- Blue Water’s and NSF focus on *sustained* performance in a way few have been before.
- *Sustained* is the computer’s performance on a broad range of applications that scientists and engineers use every day.
- The Blue Waters concept of *sustained* performance is similar to the Sustained System Performance (SSP) used at NERSC
Selection Criteria for Blue Waters

- Maximum Core Performance
  … to minimize number of cores needed for a given level of performance as well as lessen the impact of sections of code with limited scalability
- Large, Low latency, High-bandwidth Memory Subsystem
  … to enable the solution of memory-intensive problems
- Low latency, High-bandwidth Communications Subsystem
  … to facilitate scaling to the large numbers of processors required for sustained petascale performance
- High-bandwidth I/O Subsystem
  … to enable solution of data-intensive problems
- Maximum System Integration; Mainframe Reliability, Availability, Serviceability (RAS) Technologies
  … to assure reliable operation for long-running, large-scale simulations
The Blue Waters Project is embedded in a cluster of related activities in computer science, engineering and technology at Illinois:

**CCI:** Cloud Computing Initiative  
**N CofE:** NVIDIA Center of Excellence  
**UPCRC:** Universal Parallel Computing Research Center
Great Lakes Consortium for Petascale Computation (www.greatlakesconsortium.org)

**Goal:** Facilitate the widespread and effective use of petascale computing to address frontier research questions in science, technology and engineering at research, educational and industrial organizations across the region and nation.
# Blue Waters Project Components

<table>
<thead>
<tr>
<th>Petascale Education, Outreach and Industry</th>
<th>Petascale Computing Resource Allocations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Petascale Application Collaboration Team Support</td>
</tr>
<tr>
<td></td>
<td>Outstanding User and Production Support</td>
</tr>
<tr>
<td></td>
<td>WAN connections, Consulting, System Management, Security, Operations, ...</td>
</tr>
<tr>
<td></td>
<td>Value added Software – Collaborations</td>
</tr>
<tr>
<td></td>
<td>Value added hardware and software</td>
</tr>
<tr>
<td></td>
<td>Blue Waters Base System – Processors, Memory, Interconnect, On-line Storage, System Software, Programming Environment</td>
</tr>
<tr>
<td></td>
<td>Petascale Computing Facility</td>
</tr>
</tbody>
</table>

Great Lakes Consortium
Users of Blue Waters

- Blue Waters is an open science platform
  - Blue Waters must support many different types of usage
  - Benchmarks are a very limited approximation of some of the usage
  - Blue Waters is the only Leadership class resource NSF has so all types of projects will be selected
- Blue Waters users are selected by the best science projects across all disciplines
  - Projects will change throughout time
- Blue Waters Users are not yet known
  - Will vary (different application types, groups, requirements),
  - Geographically distributed
- Almost guaranteed that the selected users will be
  - Very experienced users
  - Will run applications at other large facilities
  - Could very well use community codes
  - Will be pressed to produce
Petascale Computer Resource Allocation Process (PRAC)

- First round winners to be announced March – June 2009
- Request an allocation on the Blue Waters system.
- Proposers must demonstrate proposed science or engineering research problem requires and can effectively exploit the petascale computing capabilities offered by Blue Waters.
- Receive support from Blue Waters team
- Provides travel funds
Existing Optimization Efforts

Specified Test Problems

- Three petascale applications/problem sizes
  - Lattice-Gauge QCD (MILC)
  - Molecular Dynamics (NAMD)
  - Turbulence (DNS3D)
- Three non-petascale applications/problem sizes
  - Lattice-Gauge QCD (MILC)
  - Materials science (PARATEC)
  - Weather prediction (WRF)
- Ultimate Milestones
  - Time-to-solution target (or 1 PFLOP sustained) for specified problem (size, time, physics, method)
  - Verification of numerical solution
- Future Foci determined as PRAC awards are made
Molecular Dynamics Petascale Application

• Problem Description
  • A periodic system of 100,000 lipids and 1000 curvature-inducing protein BAR domains; total system size of 100 million atoms
    • CHARMM27 all-atom empirical force field
    • Velocity Verlet time-stepping algorithm,
    • Langevin dynamics temperature coupling,
    • Nose-Hoover Langevin piston pressure control,
    • Particle Mesh Ewald (PME) algorithm for electrostatics
  • Time step = 0.002 ps, 64-bit floating point arithmetic
  • Target run time for 10 ns simulation time = 25 hours.
  • Dump positions, velocities, and forces for all atoms to disk every 500 steps.
• Solver
  • NAMD (Schulten, http://www.ks.uiuc.edu/Research/namd/)
Making Progress without target hardware

Key Elements

Applications

Single Core simulator, Network Simulators - BigSim (Kale)/Mercury (IBM)

Application Modeling and Analysis

Execution on Hardware (Power5+, Power6, BlueGene/P, Cray XT4/5, Ranger)

Execution on Power 7 (Early and Final)
Measuring and Improving Performance

- Single chip performance estimation
  - SystemSim performance simulator (with stats)
  - Power5+, Power6, BlueGene/P, Cray XT4/5, Ranger, etc.
- Scalability estimation
  - Power5+, Power6, BlueGene/P, Cray XT4/5, Ranger, etc.
  - BigSim (Kale)/Mercury (IBM) network simulators (June ‘09)
  - LANL performance models + other modeling
- Optimization
  - FP MPI, Tau, HPCS Toolkit, etc. to identify bottlenecks
  - Highly tuned libraries, loop transformations, prefetching, vector intrinsics, reorganized data structures, algorithmic improvements
  - Overlap communication with computation, one-sided communication, improved mapping of tasks to cores, MPI+OpenMP
Blue Waters—Interim systems

An interesting challenge: The hardware on which Blue Waters is based does not exist yet. NCSA has installed four systems to prepare for Blue Waters.

- “BluePrint,” an IBM POWER575+ cluster for studying the software environment.
- IBM POWER6 systems for developing the archival storage environment and scientific applications.
- An x86 system used to simulate the Blue Waters system. This allows application researchers to study the performance of scientific codes on Blue Waters’ hardware implementations.
- A shared testbed for accelerator/GPU investigations
IBM Contract for the Base System - A Superset of DARPA PERCS Demonstration

- Base System contract – sufficient to meet the minimum requirements of Track 1 RFP
- Contract specifies – in detail
  - Base system hardware
  - Base system software
  - Base programming environment
    - Languages
    - Tools
  - Advanced features
- Contract has firm performance requirements
  - Benchmarks, System services, Reliability, On-going Support
- Contract specifies schedule and performance milestones
- Specifies a number of collaborations and a process to create new collaborations
System Components

- Computational Nodes
  - Vast majority of the nodes
- I/O Nodes
  - Supports GPFS and other I/O functions
- Service Nodes
  - Programming environment
  - Interactive login
  - System management functions
- These nodes plus system management assembled from liquid cooled building blocks
# Blue Waters Petascale Computing System

## Power Architecture Trends

<table>
<thead>
<tr>
<th>System Attribute</th>
<th>Power 5</th>
<th>Power5+</th>
<th>Power6</th>
<th>Blue Waters*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock Rate (GHz)</td>
<td>1.9</td>
<td>2.3</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Peak Ops per Clock</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Peak Performance GF</td>
<td>7.6</td>
<td>9.2</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>Threads per core</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>≥ 2</td>
</tr>
<tr>
<td>Number of Sockets/SMP</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Number of Cores per Socket</td>
<td>1/2</td>
<td>2</td>
<td>4</td>
<td>≥ 4</td>
</tr>
<tr>
<td>Number of Cores per SMP</td>
<td>8/16</td>
<td>16</td>
<td>32</td>
<td>≥ 16</td>
</tr>
<tr>
<td>Cache</td>
<td>L1, Shared {L2, L3} L1, Shared {L2, L3} L1, Shared {L2, L3} {L1, L2} Shared L3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory per Core (IH) (GB)</td>
<td>2-4</td>
<td>2-4</td>
<td>2-4</td>
<td>≥ 2</td>
</tr>
<tr>
<td>VMX</td>
<td>HPS</td>
<td>HPS/IB</td>
<td>HPS/IB</td>
<td>New</td>
</tr>
<tr>
<td>MPI Latency</td>
<td>~7</td>
<td>~3.4-5.2</td>
<td>~3.4-5.2</td>
<td>&lt; 5</td>
</tr>
</tbody>
</table>

* Reference petascale computing system (no accelerators).
<table>
<thead>
<tr>
<th>System Attribute</th>
<th>Typical Cluster (NCSA Abe)</th>
<th>Track 2 (TACC)</th>
<th>Blue Waters*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor</td>
<td>Dell</td>
<td>Sun</td>
<td>IBM</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel Xeon 5300</td>
<td>AMD</td>
<td>Power 7</td>
</tr>
<tr>
<td>Peak Performance (PF)</td>
<td>0.090</td>
<td>0.58</td>
<td>&gt;1.0</td>
</tr>
<tr>
<td>Sustained Performance (PF)</td>
<td>~0.005</td>
<td>~0.06</td>
<td>&gt;0.8</td>
</tr>
<tr>
<td>Number of Processor Cores</td>
<td>9,600</td>
<td>62,976</td>
<td>&gt;200,000</td>
</tr>
<tr>
<td>Amount of Memory (PB)</td>
<td>0.0144</td>
<td>0.12</td>
<td>&gt;10</td>
</tr>
<tr>
<td>Amount of Disk Storage (PB)</td>
<td>0.1</td>
<td>1.73</td>
<td>&gt;500</td>
</tr>
<tr>
<td>Amount of Archival Storage (PB)</td>
<td>5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>External Bandwidth (Gbps)</td>
<td>40</td>
<td>10</td>
<td>100-400</td>
</tr>
</tbody>
</table>

* Reference petascale computing system (no accelerators).
Collaborations Basic Principles

- Ensure that lower risk IBM hardware and software is sufficient to support petascale applications
- Invest in higher risk IBM and non-IBM hardware and software that will:
  - Improve user, programmer & operator productivity
  - Broaden range of applications that can be well supported
  - Improve performance
  - Provide risk mitigation strategies
- Ensure that BW software is compatible with software on other high end NSF platforms
  - Contribute to Open Source and help port of open source SW to Blue Waters
Value Added Collaboration Areas

- Application Programming Environments – Eclipse based
- Advanced Programming Models – GSM, UPC, CAF, others
- Common Tools Infrastructure – A
- Petascale-Exascale Hierarchical Storage
- System Monitoring and Response
- Workflow and Data Management
- Compilers
- Interconnect Topology
- Petascale Application Collaboration Teams (PACTS)
  - 5 currently defined based on NSF benchmarks – Molecular Dynamics, Turbulence, QCD, Materials, Weather/Climate
  - ~2-5 others will be based on proposed projects
- Other Collaborations defined as the need arises (e.g. Cybersecurity)
Overall Software Architecture

- Running the System and Basic Services
  - Operating System
  - File Systems
  - Mass storage-HPSS
  - Checkpoint/Restart
  - Integrated, automated console
  - Cybersecurity
- Debugging and Tuning Applications
  - HPCS Tools
  - Performance tuning
  - Debugging
  - IDE

- Running Applications
  - Managing workflow
  - Managing data
  - Investigating results

- Developing Applications
  - Compilers
  - Parallel Programming Models
  - Frameworks
  - Computational Libraries
  - Simulators
  - Performance of programming models
### Full-featured OS Sockets, threads, shared memory, checkpoint/restore

**Languages:** C/C++, Fortran (77-2008 including CAF), UPC

**IO Model:** Global, Parallel shared file system (>10 PB) and archival storage (GPFS/HPSS) MPI I/O

**Environment:** Traditional (command line), Eclipse IDE (application development, debugging, performance tuning, job and workflow management)

**Libraries:** MASS, ESSL, PESSL, PETSc, visualization…

**Programming Models:** MPI/MP2, OpenMP, PGAS, Charm++, Cactus

**Low-level communications API supporting active messages (LAPI)**

### Hardware

- Multicore POWER7 processor with Simultaneous MultiThreading (SMT) and Vector MultiMedia Extensions
- Private L1, L2 cache per core, shared L3 cache per chip
- High-Performance, low-latency interconnect supporting RDMA

### Resource Manager: Batch and interactive access

- Performance tuning: HPC and HPCS toolkits, open source tools

### Parallel Debugging at Full Scale

- Low-level communications API supporting active messages (LAPI)

### Environment: Traditional (command line), Eclipse IDE (application development, debugging, performance tuning, job and workflow management)
File System is GPFS

- IBM is implementing scaling changes in GPFS for the HPCS/DARPA project.
- Blue Waters will implement those changes in a persistent manner.
- GPFS configured to accommodate other local systems in a single namespace.
- Performance requirements are appropriately scaled to BW characteristics.
HPSS Hardware consists of three tape robots and appropriate numbers of tape drives
  • Expect to expand this thru the lifetime of BW
HPSS integrated with BW
  • GPFS-HPSS Interface
  • Import-Export Portal
    • Traditional HPSS commands
NCSA is contribution RAIT implementation to the HPSS community as part of BW
File System is Integrated with HPSS

- Using the GPFS-HPSS Interface
  - GHI Demonstrated at SC 06-07-08
  - Design co-developed by NERSC and IBM
- Single Name Space
- Information Lifecycle management and DMAPI functions automatically move data from rotating disk to near line storage and back
- System Management Control for policy (size of files, ages of file, etc).
- A core part of a new operational concept
  - Transparent data management
  - “Virtual filesystem”
  - Lighter-weight backup
Blue Waters Can Use Either Linux Or AIX

- Same functionality specified under either OS
  - With some caveats on Linux side due to lack of control
- All performance commitments from IBM apply equally to either OS
  - It is possible one or the other will perform better
- IBM and NCSA (as organizations) do not favor one or the other
  - There is lot of internal opinion at the group level
  - What NCSA chooses will likely also be what DARPA uses
- Will be a qualitative decision based on quantitative information
Main Decision Factors

- Impact on application performance
  - Jitter (OS but also overall system)
  - Memory management [large pages]
  - Scheduling [core affinity]
  - Resource consumption [memory]
- Impact on reliability
  - Unknown at this time
- OS services performance (I/O, IP…)
  - Expected to be neutral
- Other functionality
  - No major differences are expected
- User preferences (application programmers, 3rd party tools providers, industrial partners…)
  - Linux preference expected
- Support of user tools/environments
  - Linux advantage expected
Additional Decision Factors

• Risk for long-term lack of support
• Points of responsibility within IBM for SW
  • Development/support
• IBM's testing/debugging resources for each OS
• Feature introduction schedules and timing
• Application specific requirements
• Accelerator drivers and support
• Fallback for entire OS or for components
• Detailed functionality
  • Cyber security (IPSec, IPTables)
  • Resource management
• Advisory Committees
Petascale Computing Facility

- Future home of Blue Waters and other NCSA hardware
- 88,000 square feet total – expandable
- 6’ Raised Floor
- 20,000+ square foot machine room
- Onsite cooling towers augment central chilled water system to save even more energy
Accelerator Task Objective

- Goal: Evaluate the potential of incorporating accelerators for the Blue Waters
  - Small additional acquisition cost
  - Small additional power consumption
  - Small additional application development efforts
  - Potential of Large performance increase for select applications
- Main tasks
  - Select appropriate accelerator technology and schedule
  - Identify the appropriate application development tools/framework
  - Educate application developers about accelerators
  - Provide intermediate application development systems
- Progress
  - Several encouraging studies – Chemistry and Materials
Education - Undergraduate Petascale Education Program

- Undergraduate efforts promote participation of faculty and students from diverse institutions and fields
  - 2- and 4-year primarily undergraduate colleges and universities
  - Minority Serving Institutions
  - EPSCoR institutions
  - Across all fields of study

- Three areas of emphasis
  - Professional development workshops for undergraduate faculty
  - Undergraduate materials development assistance
  - Undergraduate student internships
Education - Faculty Workshops

• **Goal:** prepare faculty for petascale computing education and research
  • Incorporate computational thinking and petascale resources in undergraduate classroom
    • Quantitative Reasoning, Computational Thinking, Multiscale Modeling for all courses
    • Use of PetaApps in upper division courses
    • Emphasize modeling over programming
  • Develop competence and confidence to mentor undergraduate research and collaborate with research universities

• **Leveraging SC08-SC11 Education Programs**
• **Collaborating with TeraGrid EOT, Krell Institute, Sigma Xi, etc.**
Faculty Workshops Announced

• Faculty professional development and curriculum development for petascale computing education and research
• Leveraging SC09 Education Program week-long summer workshops:
  • May 17-23: Oklahoma State U (EPSCoR) - chemistry
  • June 7-13: U Calif Merced - biology
  • June 7-13: Kean University (MSI) - parallel and cluster computing
  • June 14-20: Widener U - physics
  • July 5-11: Atlanta University Center (MSI) - computational thinking
  • July 5-11: Louisiana State U (EPSCoR) - parallel and cluster computing
  • July 12-18: Florida State U - pre-college
  • July 12-18: Ohio Supercomputer Center - engineering
  • Aug 2- 8: Arkansas U (EPSCoR) - computational thinking
  • Aug 9-15: Oklahoma U (EPSCoR) - parallel and cluster computing
• Meeting with instructors in April to incorporate petascale content into the summer workshops
Education - Materials Development

- Goal: prepare students for petascale computing
- Compelling examples in contemporary science, from desktop to grid to petascale incorporating:
  - Quantitative reasoning
  - Computational thinking
  - Multiscale modeling
- Thirty (30) modules targeted for development in all fields of science and engineering - ~10 per year
- Development cycle over 18 months per module
  - Materials development; classroom field test; VVA review; publish
  - Stipend paid to authors after acceptance in CSERD
Education - Undergraduate Internships

- **Goal**: prepare selected students for petascale computing research and education
  - Embed students in Blue Waters research groups
  - Full-year internship starting in summer
  - Participation in SC08-11 Education Programs
- **Interns interact across projects**
  - Wiki, video conferencing
  - Threaded discussions
  - Annual meeting
- **Initial funding for up to 5 per year, starting this summer**
Education - Virtual School of Computational Science and Engineering (VSCSE)

- Goal: Prepare the current & next generation of scientists and engineers to utilize leading edge computer systems
- A multi-state, multi-institutional virtual organization
  - Bring together faculty at CIC, research universities and institutions in the GLCPC to leverage geographically dispersed expertise
  - Build on CIC infrastructure and course-sharing agreements
- Increase and enhance HPC and petascale curricula available to graduate students
  - Define core competencies in HPC & petascale, identify gaps, establish best practices across the CIC
  - Develop on-line learning materials, training modules, courses, summer schools, workshops, and seminars
- Four Phase Development Plan
Education - Virtual School Summer Schools

• Accelerators for Science and Engineering Applications
  2008 Summer School
  • 42 students attended from 179 applications
  • Participants reported summer school contributed “a lot” to their understanding and abilities.
  • Materials from the Summer School are available on-line

• Challenge is to scale-up to reach many more students

• Two 2009 Summer Schools announced
  • Scaling to Petascale - August 3-7
  • Many-core Processors - August 10-14
  • Synchronous HD access at 4 sites plus webcast of sessions
  • All materials to be available on-line
Summary

• The scientific need for Blue Waters exists
• Blue Waters is a well balanced system designed to run a wide range of applications at the sustained petaflop/s
• Blue Waters is much more than just a large computing system
• Blue Waters has the potential of setting the direction towards Exascale computing
• NSF, NCSA, U of I and the GLCPC are the right partners to make Blue Waters sustained Petascale Computing a Reality
Questions

- Blue Waters Contributions
  - PIs
    - Thom Dunning, Marc Snir, Bill Gropp, Wen-mei Hwu, Bill Kramer
    - Past PIs – Rob Pennington, Ed Seidel
  - Partners (more in the future)
    - Shodor, RENCI, LSU, LANL
  - Funding Agents
    - NSF, State of Illinois, University of Illinois
- Blue Waters Staff
- Vendors
  - IBM, others TBD