Initial BlueGene/L Application Performance Results

Bronis R. de Supinski and Steve Louis
Lawrence Livermore National Laboratory

Conference on High Speed Computing
April 19, 2005

UCRL-PRES-211410
This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.
Wide Variety of Early Applications on BlueGene/L

Blue Matter (IBM) *  SAGE (LANL)
Flash (ANL) *  SPPM (LLNL)
Miranda (LLNL) *  UMT2K (LLNL)
MM5  Sweep3d (LANL)
Amber7, Amber8  MDCASK (LLNL)
GAMESS  GP (LLNL)
QMC (Caltech)  CPMD (IBM/LLNL) *
LJ (Caltech)  TLBE (LBL)
PolyCrystal (Caltech)  HPCMW (RIST)
PMEMD (LBL)  ParaDiS (LLNL)
LSMS (ORNL)  QCD (IBM)*, QCD (BU) *
NIWS (NISSEI)  NAMD
HOMME (NCAR) *  PAM-CRASH (ESI)
Qbox (LLNL)  Raptor (LLNL) *
ddcMD (LLNL)  Enzo (SDSC)
Successful scaling tests and science runs completed for key ASC codes

- Using IBM Rochester BG/L hardware through SC2004 (many, many thanks to Jim Sexton of IBM for his help)

- Several codes running at scale since January 3, 2005 on Livermore’s first 32 racks (16 before this month)

- ASC has concentrated on scaling up the following codes:
  - ddcMD
  - FEQMD
  - GRASP (SNL)
  - hypre/SMG2K
  - LAMMPS (SNL)
  - MDCASK
  - Miranda
  - ParaDiS
  - Qbox
  - Raptor
  - SPaSM (LANL)
  - sPPM (Benchmark)
  - UMT2K (Benchmark)

Excellent scaling seen on runs up to 32K nodes and 64K processors
Current Time Line

1. 16 racks delivered in November 2004
   – System tested in December
   – Runs started in January

2. 16 additional rack delivery early February
   – 32-rack runs started in April

3. 32 additional rack delivery in mid 2005
   – integration to 64 rack system
   – Machine shake out and test
   – 64-rack scaling/science follows

4. Late 2005 expanded early user access and more science runs
BlueGene/L scales to 360 TF with modified COTS and custom parts

**Compute Chip**
- 2 processors
- 2.8/5.6 GF/s
- 4 MiB* eDRAM

**I/O Card**
- FRU (field replaceable unit)
- 25mmx32mm
- 2 nodes (4 CPUs)
- (2x1x1)
- 2.8/5.6 GF/s
- 256/512 MiB* DDR

(compare this with a 1988 Cray YMP/8 at 2.7 GF/s)

**Node Card**
- 16 compute cards
- 0-2 I/O cards
- 32 nodes
- (64 CPUs)
- (4x4x2)
- 90/180 GF/s
- 8 GiB* DDR

**Compute Card**
- 2.8/5.6 GF/s
- 256/512 MiB* DDR

**Cabinet**
- 2 midplanes
- 1024 nodes
- (2,048 CPUs)
- (8x8x16)
- 2.9/5.7 TF/s
- 256 GiB* DDR
- 15-20 kW

**System**
- 64 cabinets
- 65,536 nodes
- (131,072 CPUs)
- (32x32x64)
- 180/360 TF/s
- 16 TiB*
- 1.2 MW
- 2,500 sq.ft.
- MTBF 6.16 Days

Architectural features promote efficiency and scaling for important applications

- Multiple complementary interconnects support diverse application scaling requirements
  - 3D torus with bi-directional nearest-neighbor links
  - 2.1 GB/s combining tree for fast global reductions
  - Low-latency global barrier network

- High reliability expected from high level of integration using system-on-a-chip technology

- Architectural enhancements improve single node performance

- Software architected with very powerful “divide and conquer” technique for software scale-up

The BG/L project is a focused effort to enable important science and to lead the way to cost-effective petaFLOP/s computing
BlueGene/L ~15W compute node has a single ASIC and 9 SDRAM-DDR chips.
PPC440 processor core

- 700 MHz
- 2-way superscalar (2 instructions per cycle)
  - 1.4 Ginstrs per processor
  - 2.8 Ginstrs per node
  - 184 Tinstrs overall
  - \( \leq 6 \) ops per proc per cycle (e.g., 2 FP mac, 1 int mac)

- 3 execution pipelines:
  - Load/store
    - Up to 3 loads pending
  - Simple integer
  - Complex integer, branch
- 32 32-bit integer registers

- “Double hummer” FPU added
- Dynamic/static branch prediction
  - 2-bit branch history
  - 1-cycle branch latency
- Dedicated HW loop counter and special loop branch instruction
Double floating point unit

- 2 64-bit FPUs per core (so 4 per node)
- 32 64-bit register pairs service the 2 FPUs
- An instruction can drive either FPU or both (SIMD)
  - A SIMD Multiply-Accumulate does 4 64-bit Flops
  - Also has complex, other intra-pair instructions

- 2 FMAs × 1 core × 64k nodes @ 700 MHz = 184 Tflops peak
  - Using 2nd core’s DH-FPU gives 367 Tflops peak
  - Using single-op instrs (non-MAC) reduces by 1/2
  - To approach peak, avoid reading off-chip memory
    - Load BW from L3: 64 bits every .25 cycles
    - Load BW from memory: 64 bits every 1.4 cycles
- Quadword load fills a register pair (also useful for comm)
Two ways for apps to use hardware

Mode 1 (Co-processor mode - CPM):  
- CPU0 does all the computations  
- CPU1 does the communications  
- Communication overlap with computation  
- Peak comp perf is $5.6/2 = 2.8$ GFlops

Mode 2 (Virtual node mode - VNM):  
- CPU0, CPU1 independent “virtual tasks”  
- Each does own computation and communication  
- The two CPU’s talk via memory buffers  
- Computation and communication cannot overlap  
- Peak compute performance is 5.6 GFlops
Physics and Materials is initial ASC focus for the early BlueGene/L apps.
BlueGene/L will allow overlapping evaluation of models for first time.
Predicting the mechanical strength of material from first principles is difficult.

- In fact, it has remained a grand challenge for computational materials science for several decades…

- At a fundamental scale, we need to consider the atomistic structure of a “dislocation” and the atomistic mechanisms of dislocation motion.

- At higher scale, we need to consider interaction of many dislocations that form complex patterns on a micron scale to understand the plastic strength of a single crystal.

- Single-crystal strength is used to model a poly-crystal of many crystal grains, which in turn supports yet higher-scale finite-element models of complex-shape object deformation.
Criteria for “First-Wave” Applications

• First-wave target ASC applications
  – Efforts identified to be ready for programmatic science runs with early machine availability, and ongoing assessment of code suitability

• Assessment criteria for early apps
  – Importance to the ASC program
  – Enthusiasm within the code group
  – Potential for good code scaling (i.e., simpler architectural needs)

• Second-wave target ASC apps are also being identified using similar criteria to those above
Also examining many other science applications to run on BlueGene/L

- **ALPS** – Predictive modeling of laser plasma interaction
- **AMRh** – High Reynolds-number fluid flows
- **BOUT** – MFE boundary-layer turbulence
- **DJEHUTY** – 3D modeling of stars
- **DYNA3D** – Structural mechanics
- **EMSOLVE** – Electromagnetic coupling with structures
- **FMC** – Solves Schroedinger Equation for a many-fermion system
- **GFMD** – Greens-function molecular dynamics
- **HYDRA** – Implicit radiation diffusion solver, multi-mode instabilities
- **IRS** – Implicit radiation solver
- **ParaDyn** - LLNL parallel engineering code based on DYNA3D
- **PF3D/Z3** – Predictive modeling of laser plasma interactions
- **ROLEX** – Detailed-accounting opacity
- **SAGE** – LANL widely used adaptive-grid Eulerian hydrodynamics
Classical MD – ddcMD:
Rapid solidification in tantalum

- Scalable, general purpose code for performing classical molecular dynamics (MD) simulations using highly accurate MGPT potentials

- MGPT semi-empirical potentials, based on a rigorous expansion of many body terms in the total energy, are needed to quantitatively investigate dynamic behavior of transition metals and actinides under extreme conditions

64K and 256K atom simulations on 2K nodes are order of magnitude larger than previously attempted; based on 2M atom simulation on 16K nodes, close to perfect scaling expected for full machine (“very impressive machine” says PI...)

Visualization of important new scientific findings already achieved on BG/L: Molten Ta at 5000K demonstrates solidification during isothermal compression to 250 GPa
Excellent scaling of ddcMD on BG/L supports greater understanding of solidification process

- Nucleation of solid is initiated at multiple independent sites throughout each sample cell
- Growth of solid grains initiates independently, but soon leads to grain boundaries which span the simulation cell: size of cell is now influencing continued growth
- 2,048,000 simulation recently performed indicates formation of many more grains

ddcMD is already using 32K* CPUs of BG/L for unprecedented multi-million atom MGPT simulations

* Virtual node mode gives 1.9x performance out to 1K processors
Classical MD - SPaSM

- A high performance (1993 and 1998 Gordon Bell prizes) code for Scalable Parallel Short-range Molecular dynamics simulations

- A variety of finite-range empirical potentials are implemented, including EAM and MEAM for metals, Stillinger-Weber Si/Ge, and a reactive empirical bond-order (REBO) potential for detonation studies.

SPaSM has exhibited excellent scaling for up to 100 billion atoms on 16,384 nodes, and an initial production run on 8k nodes simulated the shock loading of a 2.1 billion atom EAM copper crystal with 0.41% (by volume) voids. BG/L will enable the exploration of an entirely new class of (previously intractable) problems such as this.
SPaSM performance on BG/L with 100,000 — 6,000,000 atoms per node

- Excellent scaling is seen for >1 million atoms per node
- This will extend to smaller sizes for more complex potentials, which typically have much higher CPU/communication ratios
GRASP: Scalable Molecular Dynamics Code for Reactive Force Fields (SNL)

- A scalable, general purpose code for performing classical molecular dynamics (MD) simulations
- Supports a wide range of different force fields: twobody, threebody, Tersoff, EAM, ReaxFF, electrostatics, charge equilibration
- Applications include: Radiation Damage, Materials Interfaces, Explosives, MEMS
- Standard version ported to BGL without difficulties
- Development version combining C++ and Fortran implements ReaxFF force field. Used BGL to test the code. Several software bugs detected and fixed.
- Absolute speed for ReaxFF close to 3 GHz Pentium cluster.
- Stillinger-Weber silicon benchmark scales well
  - 70% efficiency at 122 atoms/CPU on 16k processors.
  - 35% efficiency at only 4 atoms/CPU on 4k processors.
GRASP performance on BG/L and comparison to HP cluster (3 GHz Pentium+Myrinet)

α-Silicon crystal
Rhombohedral periodic cell
Stillinger-Weber force field

RDX Explosive with Oxygen
ReaxFF force field with charge equilibration

---

**CPU Time (µs/timestep-atom)**

- **Pentium**
- **BGL**

**Number of CPUs**

---

**CPU Time (secs/timestep)**

- **Pentium**
- **BGL**

**Number of Processors**
LAMMPS Classical MD (SNL)

- LAMMPS = Large-scale Atomic/Molecular Massively Parallel Simulator

- LAMMPS is a classical molecular dynamics code that models an ensemble of particles in a liquid, solid, or gaseous state. It can model atomic, polymeric, biological, metallic, or granular systems using a variety of force fields and boundary conditions.

- On parallel machines, LAMMPS uses spatial-decomposition techniques to partition the simulation domain into small 3d sub-domains, one of which is assigned to each processor.

LAMMPS has been tested on up to 512 BG/L processors so far, and shown good scaling on a fixed-size (32,000 atoms) problem (strong scaling).
LAMMPS strong scaling on BG/L with 32,000 atom fixed size problem

PI: “With 2048 procs, the benchmark took 1.3 seconds, which is a substantial improvement over our record on that benchmark (5.9 seconds on 64 CPUs of Sandia's Liberty cluster)”
Classical MD - MDCASK

- MDCASK simulates the motion of large collections of individual atoms using the classical laws of Newtonian mechanics and electrostatics.
- Capable of using a wide variety of inter-atomic potentials allowing simulation of metals, semiconductors, insulators, and glasses.

- Weak scaling (with a constant 250,000 atoms per processor) was tested up to 16,384 processors with excellent results.
- Virtual node mode yields a factor of 1.78 speedup.
- To simulate 1 ns with $10^{10}$ atoms requires ~8 days on the full-sized BG/L.
- Strong scaling tests perform well down to ~2,000 atoms / node.

MDCASK is ready to apply the full power of BGL to multi-billion atom simulations.
Qbox is a C++/MPI implementation of the plane-wave, pseudopotential, ab initio molecular dynamics method within Density Functional Theory (DFT). It is developed at LLNL.

- Massively parallel C++ / MPI implementation with specialized 3D FFTs
- Routinely used at LLNL for simulations of condensed matter subjected to extreme such as high pressure and high temperature, as well as in nanotechnology and biochemistry applications.
- 686-atom Mo solid and other heavy metal simulations are under way
- Scalability tests on BG/L show that Qbox can achieve a 3x speedup when solving a given problem on 16384 nodes instead of 4096 nodes. This represents a 75% parallel efficiency. Further optimizations will provide even greater efficiency.
Qbox: (strong) scaling on BG/L
Solid Molybdenenum simulation

- Some lessons learned:
  - Node mapping is critical, can result in a 2x speedup
  - Mixed “AIX/Linux” development environment, with evolving compilers and nascent MPICH-2 BG/L device, has proved challenging
  - 16k task algorithm scaling frequently requires modifications: Rewrote some ScaLAPACK functions to improve scaling above 4k nodes

- Current efforts target generating efficient node mappings, optimization of linear algebra operations and parallel I/O
First-Principles MD - FEQMD

- FEQMD has required a complete “inversion” of its parallelization strategy.
- Data arrangement: \( \{X, Y, Z\} \) in real space, \( \{Z, Y, X\} \) in Fourier space reduces the number of transpose operations.
- Further scaling and optimization work is underway.

\[
H_e \Psi_s = E_s \Psi_s
\]

\( \Psi_{\{k,n,\sigma\}}(r) \) → FFT 3D complex (up to 256 nodes)

\( \Psi_{\{k,n,\sigma\}}(r) \) → FFT 3D complex (up to 256 nodes)

BGL is \(~25\times\) the power of ASCI Q, where we currently simulate 128 atoms.

Runs on BG/L for 64 atoms, one self-consistent calculation without I/O.
Dislocation Dynamics - ParaDiS (Parallel Dislocation Simulator)

- New LLNL code for direct computation of plastic strength of materials
- Tracks simultaneous motion of millions of dislocation lines
- Promises to close the computational performance gap that prevents scientists from understanding the fundamental nature of material strengthening (or hardening)

ParaDiS has run on 16,384 nodes of BG/L, and is currently investigating scaling and dynamic load balancing issues to achieve higher efficiencies.

Killer applications:
- full simulation of poly-crystal solidification from melt
- alloy microstructure evolution during plastic deformation
- science of ultra-fast polymer crystallization
Dynamic load balancing key to scaling ParaDis to large node counts

- **Time per Step**
  - 4096 CPUs
  - 8192 CPUs
  - 16384 CPUs

- **Force vs. Communication Time**
  - Total
  - Force
  - Communication

- **Speedup vs. CPUs**
  - Perfection
  - BGL

- **Percent of Total Time**
  - Force
  - Communication
Instability and Turbulence - Miranda

- High order hydrodynamics code for computing fluid instabilities and turbulent mix
- Employs FFTs and band-diagonal matrix solvers to compute spectrally-accurate derivatives, combined with high-order integration methods for time advancement
- Contains solvers for both compressible and incompressible flows
- Has been used primarily for studying Rayleigh-Taylor (R-T) and Richtmyer-Meshkov (R-M) instabilities, which occur in supernovae and Inertial Confinement Fusion (ICF)

Miranda has successfully run on 16,384 nodes on BG/L and also on 32,768 processors in “virtual node” mode. BG/L enables wide range of scales in space and time necessary to represent turbulent flows of interest. Good time-to-solution improvement from MCR to BG/L.
Miranda Scaling on BG/L

Miranda weak scaling runs - BG/L LLNL
16x16x2048 coprocessor mode - pentadiagonal
8 April, 2005 (driver-100) - custom 8K & 32K maps

Miranda strong scaling runs - BG/L LLNL
512x512x512 coprocessor mode
8 April, 2005 (driver-100) - custom 8K & 32K maps
Instability and Turbulence - Raptor

- Multi-physics Eulerian Adaptive Mesh Refinement (AMR) code used for applications at LLNL including astrophysics, Inertial Confinement Fusion (ICF) and shock-driven instabilities and turbulence

- Can be used to simulate purely fluid dynamics systems and more complex physical systems where the fluids are coupled to the radiation field, such as in ICF or astrophysics

Simulations at full scale on BG/L will offer the computational power to gain an order of magnitude more resolution in simulations of three-dimensional shock-driven systems.
Raptor exhibits good scaling but shows system software immaturity

- Different configurations collapse beautifully
- Super-linear scaling
- Effect of co-processor mode
- Others (larger ones) scale linearly, but linear scaling based on a smaller N=1 time/step
- Per node performance comparable to Intel Linux platforms with higher clock speeds on same problem size per node; however, BG/L exhibits significantly better scaling
BlueGene/L promises to revolutionize DOE mission and high-end computing

BG/L is already the fastest computer in the world, at only 1/4 the size of its eventual 64-rack configuration at LLNL this summer...

Linpack numbers and the Top 500 are certainly exciting news events, with IBM, BG/L and DOE at the top once again...

*BUT, the application results such as those just presented are what all the excitement should really be about:*

- Enabling better science
- Impact on national mission
- Cost-effective path to petaFLOP/s
- Validating BG/L HW & SW design and capabilities
Raptor Linux Performance (ALC)