

Unsurpassed performance, ultrascale computing

IBM @server[®] Blue Gene Solution

The Application Perspective



Introduction

The IBM[®] @server[®] Blue Gene[®] Solution is the result of an IBM supercomputing project begun five years ago dedicated to building a new family of supercomputers optimized for bandwidth, scalability and the ability to handle large amounts of data while consuming a fraction of the power and floor space required by today's fastest systems.

The goal of the Blue Gene project has been to develop a massively parallel computer applied to the study of biomolecular phenomena such as protein folding. The effort would advance the understanding of the mechanisms behind protein folding via large-scale simulation, and explore novel ideas in massively parallel machine architecture and software. The level of performance provided by Blue Gene can enable a tremendous increase in the scale of simulations beyond what is possible with existing supercomputers. Successful simulation studies of protein folding on this scale are expected to advance the techniques, models, and algorithms used in biomolecular simulation.

Starting earlier this year with the availability of operational hardware, hands-on experiences with many differing applications have shown that the Blue Gene architecture is applicable to a broad set of parallel workloads found across a variety of disciplines. Today, IBM and its partners are exploring a growing list of high performance computing (HPC) applications including life sciences, financial modeling, hydrodynamics, quantum chemistry, molecular dynamics, astronomy, space research and climate modeling. Other promising areas of interest include Grid Computing, business intelligence, financial risk and compliance, aerodynamics study and testing, and manufacturing processes.

As greater numbers of scientists and researchers apply large-scale cluster computing to a diverse set of complex problems and build a collective expertise in parallel program development, the relevance of the Blue Gene architecture becomes clearer. The design innovations of Blue Gene offer the promise of advancing vital science across numerous disciplines.

Innovative design benefits parallel applications

The Blue Gene system is built out of a very large number of compute nodes, each of which has a relatively modest clock rate contributing to both low power consumption and low cost. Blue Gene utilizes embedded processors, embedded DRAM and system-on-a-chip techniques that allow for integration of all system functions including compute processor, communications processor, 3 cache levels, and multiple high speed interconnection networks with sophisticated routing onto a single ASIC.

Leveraging PowerPC Technology

The Blue Gene chip contains two standard 32-bit embedded PowerPC 440 cores, each with private L1 32KB instruction and 32KB data caches. The cores also have a 2KB L2 cache each and share a 4MB L3 EDRAM cache. While the L1 caches are not coherent, the L2 caches are coherent and act as a prefetch buffer for the L3 cache.

Each core drives a custom 128-bit “Double” FPU that can perform four double precision floating-point operations per cycle. This custom FPU consists of two conventional FPUs joined together, each having a 64-bit register file with 32 registers. The PPC instruction set has been extended to perform SIMD-style floating point operations on the two FPUs.

Because of a relatively modest processor cycle time, the memory is close, in terms of cycles, to the processor. This is also advantageous for power consumption, and enables construction of dense packages in which 1024 dual-processor compute nodes can be placed within a single rack.

The nodes in the Blue Gene system are diskless.

Two Modes of Operation Offer Flexibility

Depending on the nature of the application to be run on Blue Gene, the programmer may choose to employ both processors in a node for computation, or have one processor dedicated to handling message passing operations.

In *co-processor* mode, the application runs in a single, non-preemptable thread of execution on the main processor (cpu 0). The coprocessor (cpu 1) is used as an off-load engine that runs as part of a user-level application library, communicating with the main processor through a non-cached region of shared memory. Use of the communication co-processor relieves the main processor of handling the network hardware, and allows the overlap of computation and communication; hence it allows optimal use of the available bandwidth.

In *virtual node* mode we provide support for two independent application processes in a node, thus allowing both processors of the compute node to be used for computation. The two processes share the L3 cache, memory, and the networks on the node. Communication between those processes is done through a special region of the non-cached shared memory.

Special –Purpose Networks Enhance Blue Gene Value

The nodes are interconnected through five networks: a 3-dimensional torus network for point-to-point messaging between compute nodes, a global collective network for collective operations over the entire application, a global barrier and interrupt network, a gigabit Ethernet for machine control, and another gigabit Ethernet network for connection to other systems.

The networks of interest to the application programmer are the torus and the global collective network.

The 3D torus allows for each node to have low-latency, high-bandwidth interconnect with its six nearest neighbors and is useful on applications where locality of computation is prevalent. The global collective network is useful for speeding up commonly used MPI collective communications constructs.

The main communication network for point-to-point messages is the torus. Each node contains six bi-directional links for direct connection with nearest neighbors. The network hardware in the ASICs guarantees reliable, unordered, deadlock-free delivery of variable length (up to 256 bytes) packets, using a minimal adaptive routing algorithm. It also provides simple broadcast functionality by depositing packets along a route.

The global collective network is useful for speeding up commonly used MPI collective communications constructs. The collective network supports fast configurable point-to-point, broadcast and reductions of packets, with a hardware latency of 1.5 microseconds for a 64K node system. An ALU in the network can combine incoming packets using bitwise and integer operations, forwarding a resulting packet along the network.

Familiar software environment tuned for Blue Gene

Three fundamental principles were followed when the system software was designed for Blue Gene: simplicity, performance and familiarity. Driving toward simplicity in the software design has allowed development of software that takes advantage of hardware features to deliver high performance without compromising stability and security. And by creating a programming and administration environment based on familiar programming languages, libraries, job management tools, and parallel file systems, clients benefit from the innovative design elements of Blue Gene without facing a steep learning curve.

The compute nodes of a Blue Gene machine run a special Compute Node Kernel (CNK) which provides a simple, flat, fixed-size 512MB address space, with no paging. CNK presents a familiar POSIX interface where the GNU Glibc runtime library has been ported and basic file I/O operations are supported through system calls. Multi-processing services (such as fork and exec) are meaningless in a single process kernel and not implemented. Program launch, termination, and file I/O is accomplished via messages passed between the compute node and its control node over the tree network, using a point-to-point packet addressing mode.

The front-end nodes of a Blue Gene complex are the portals through which programmers access the computational core of the system. They run a standard Linux® distribution which provides a familiar platform from which users compile and debug programs and submit jobs.

Blue Gene systems are supported by standard IBM XL Fortran, C and C++ compilers for PowerPC that have been augmented with a backend that takes advantage of the dual floating-point unit that is unique to Blue Gene.

Programmers will be able to employ the popular IBM Engineering and Scientific Subroutine Library (ESSL), a state-of-the-art collection of over 400 mathematical subroutines that provide optimum performance for floating-point engineering and scientific applications written in FORTRAN, C or C++. Many of these routines will be tuned for the Blue Gene architecture.

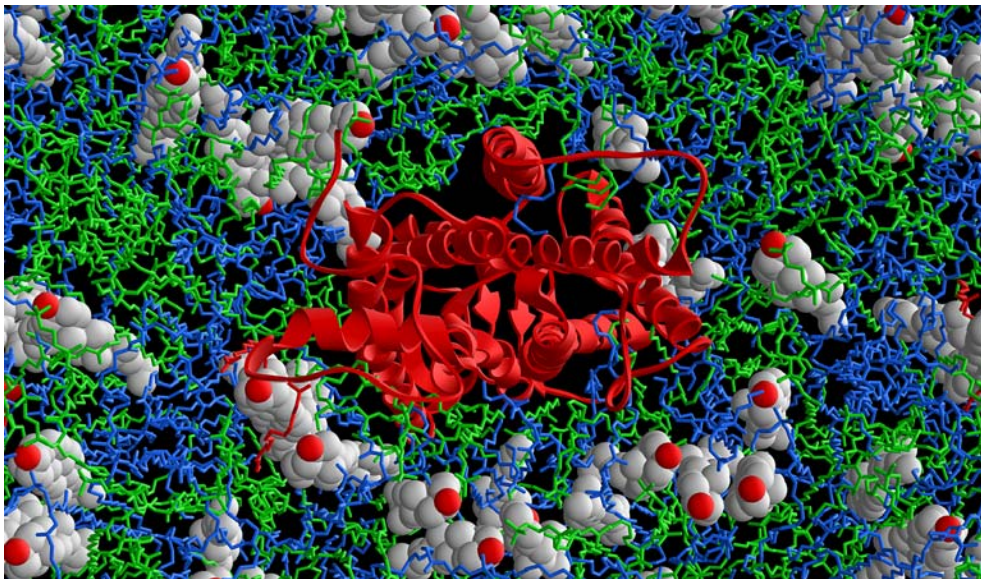
As Blue Gene is designed primarily to run MPI workloads, the degree to which the MPI implementation can be made fast and efficient will to a large extent drive

user satisfaction with the machine. Blue Gene/L MPI is based on MPICH2, a public domain implementation of the MPI 2 protocol. MPICH2 provides better scalability, lower overhead, and a modular software approach. The default implementation of MPI collective operations in MPICH2 relies on sequences of point-to-point messages, building a tree of connections to perform broadcasts and reductions.

Actual experiences with performance results on Blue Gene

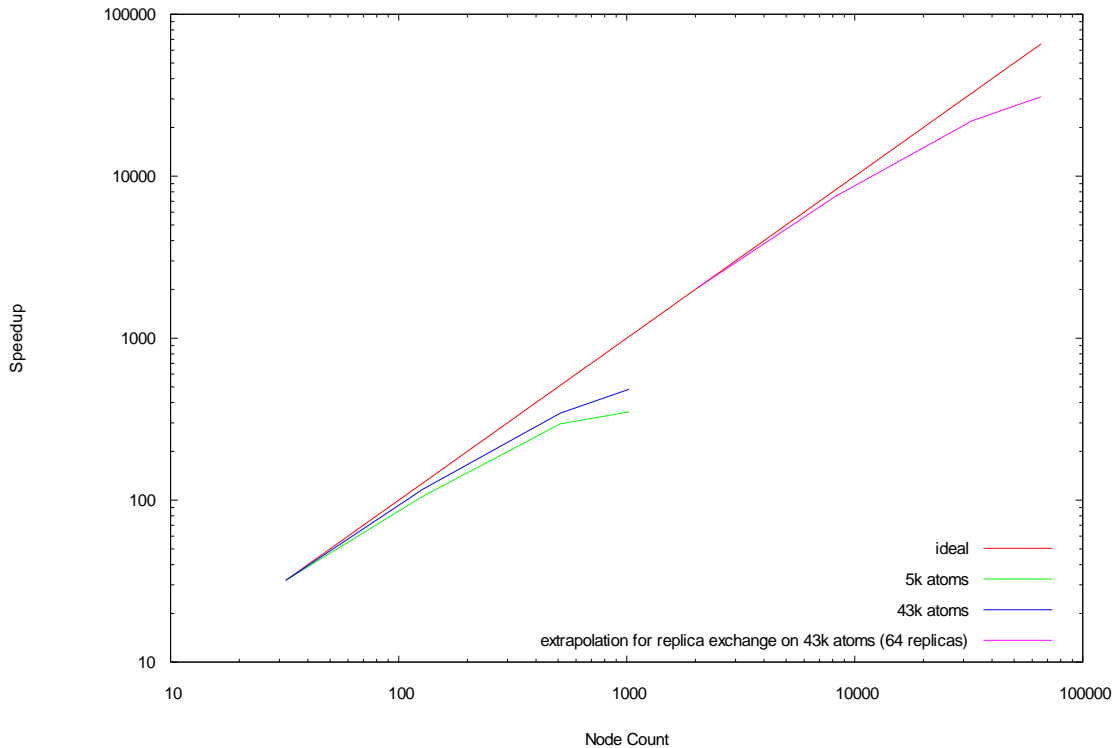
In the short time that Blue Gene has been accessible for running applications, numerous applications have been run on the machine. These include protein folding, quantum chromo-dynamics, weather simulations and astrophysics.

The Computational Biology Center (CBC) at IBM Research in Yorktown Heights has worked on a molecular dynamics framework called Blue Matter to do classical molecular simulation. This group with collaborators has applied Blue Matter to the molecular dynamics simulation of a protein, rhodopsin, in a membrane comprising lipids and cholesterol all immersed in water. The rhodopsin protein is involved in the perception of light by the eye. Rhodopsin is a member of a family of proteins, known as G-Protein Coupled Receptors (GPCRs) that are important in a number of biological processes including cell signaling and are a large fraction of the drug targets of interest to pharmaceutical companies. The details of the organization of the lipid bi-layer in the presence of a GPCR may be crucial to its functioning and are impossible to study via experiment. There are macroscopic observables in this system that can be computed in the simulation and compared with experiment to validate the models used.



Rhodopsin in 2:2:1 SDPE/SDPC/Cholesterol after 120ns

On Blue Gene, the Blue Matter team has conducted some performance studies showing speed up for two different problem sizes as a function of node count. The chart below illustrates this along with a line representing the ideal or perfect speed up and a third line that extrapolates the performance of the replica exchange technique, which uses multiple trajectories, for a 64 replica simulation of the 43,000 atom system. The speedups for the strong scaling cases are normalized to the 32-node performance and similarly, the replica-exchange extrapolated speedup is normalized with respect to the 2048 (64*32) node value.

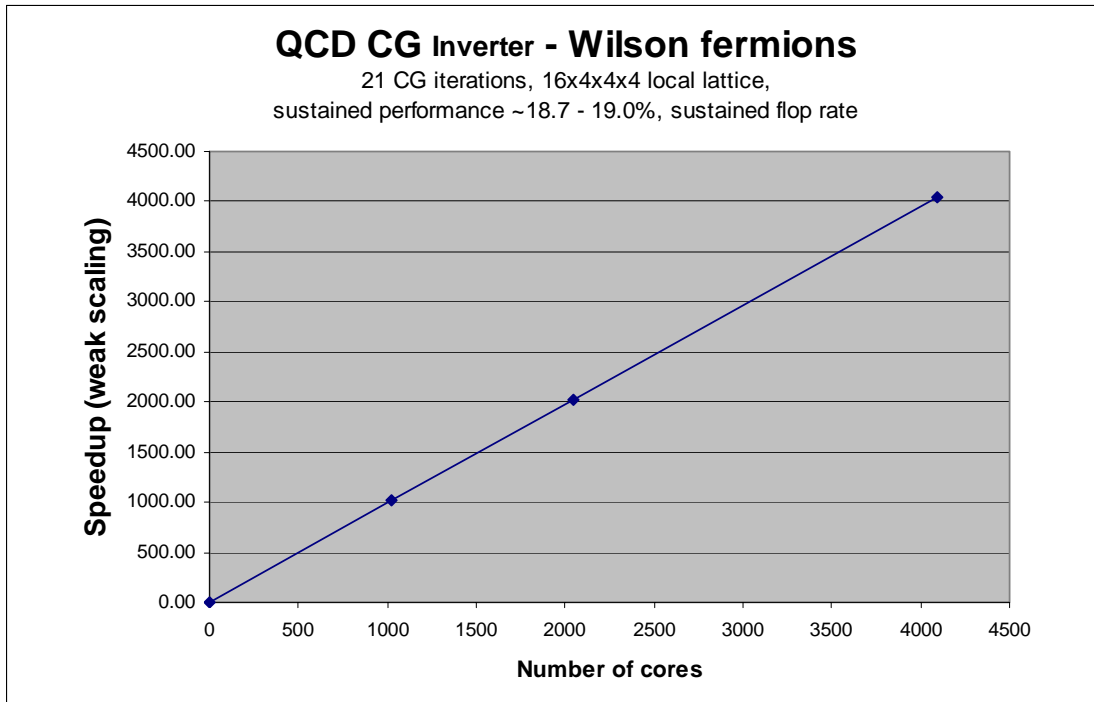


The Blue Gene machine is especially well suited for those calculations associated with Quantum Chromo-Dynamics. A number of quantum chromo-dynamics experts from IBM Research have developed a code called QCD. The QCD code does a first-principles numerical simulation of Quantum Chromo-Dynamics, the theory of the strong nuclear interactions, using Lattice Gauge Theory. This force acts on elementary particles called quarks and binds them tightly into stable nuclear particles such as protons and neutrons. The force is mediated by particles called gluons and it is so strong that the quarks can never escape outside the nuclear particle. This property is responsible for the stability of the matter that most of our universe is made of. However, when nuclear matter is heated up to tremendous temperatures (about 160-170 MeV) it undergoes a phase transition (it melts) into a quark-gluon plasma. Currently an attempt to recreate this condition is underway in big accelerator laboratories. The reverse process is thought to have occurred in the early universe when the fireball from

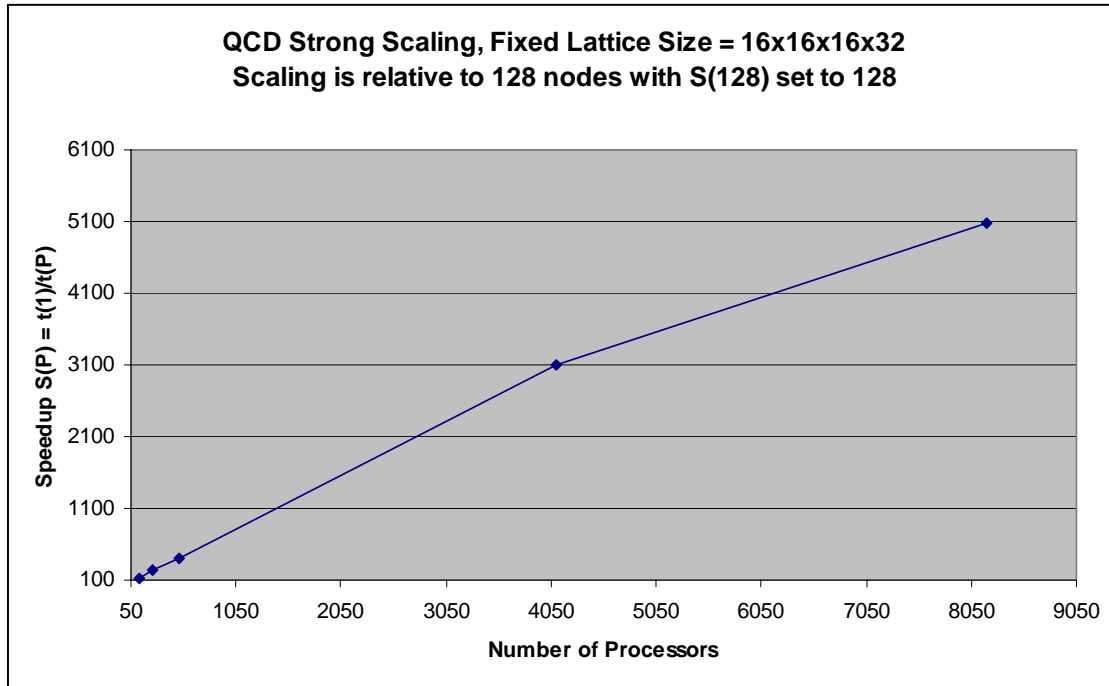
the big bang, containing a quark-gluon plasma, cooled-off and formed the stable nuclear matter and the cosmos as we know it. The QCD simulation changes the temperature and it drives the system through a phase transition. At low temperature, the system is ordinary nuclear matter. At high temperatures the system is a quark-gluon plasma.

The Quantum Chromo-Dynamic calculation requires a tremendous amount of computation. Without performance of ultrascale computing, one is forced to make approximations that give unreliable results. For the first time, the computational capacity of Blue Gene can produce reliable results to this problem. In most QCD calculations, more than 90% of the compute time is spent in a small kernel that is called the Wilson D-slash operator. It is advantageous to optimize the performance of this kernel to take full advantage of Blue Gene hardware features. The optimization consisted of grouping and arranging floating point computations to avoid pipeline conflicts and overlap with load/stores pipeline, selecting a memory storage ordering that minimized pointer arithmetic, arranging floating point load/stores to optimize cache hierarchy, carrying out data transfers through an effective nearest-neighbor communications layer that interacted with the torus network hardware, and using a fast custom global sum routine over the torus network. Further improvement is expected by replacing this global sum routine with one that will use the tree network allowing for additional scaling in multi-rack Blue Gene systems. The results of optimizing this kernel are seen in the following scaling charts.

QCD Weak Scaling on up to 4096 cores (1 core = 1 processor)

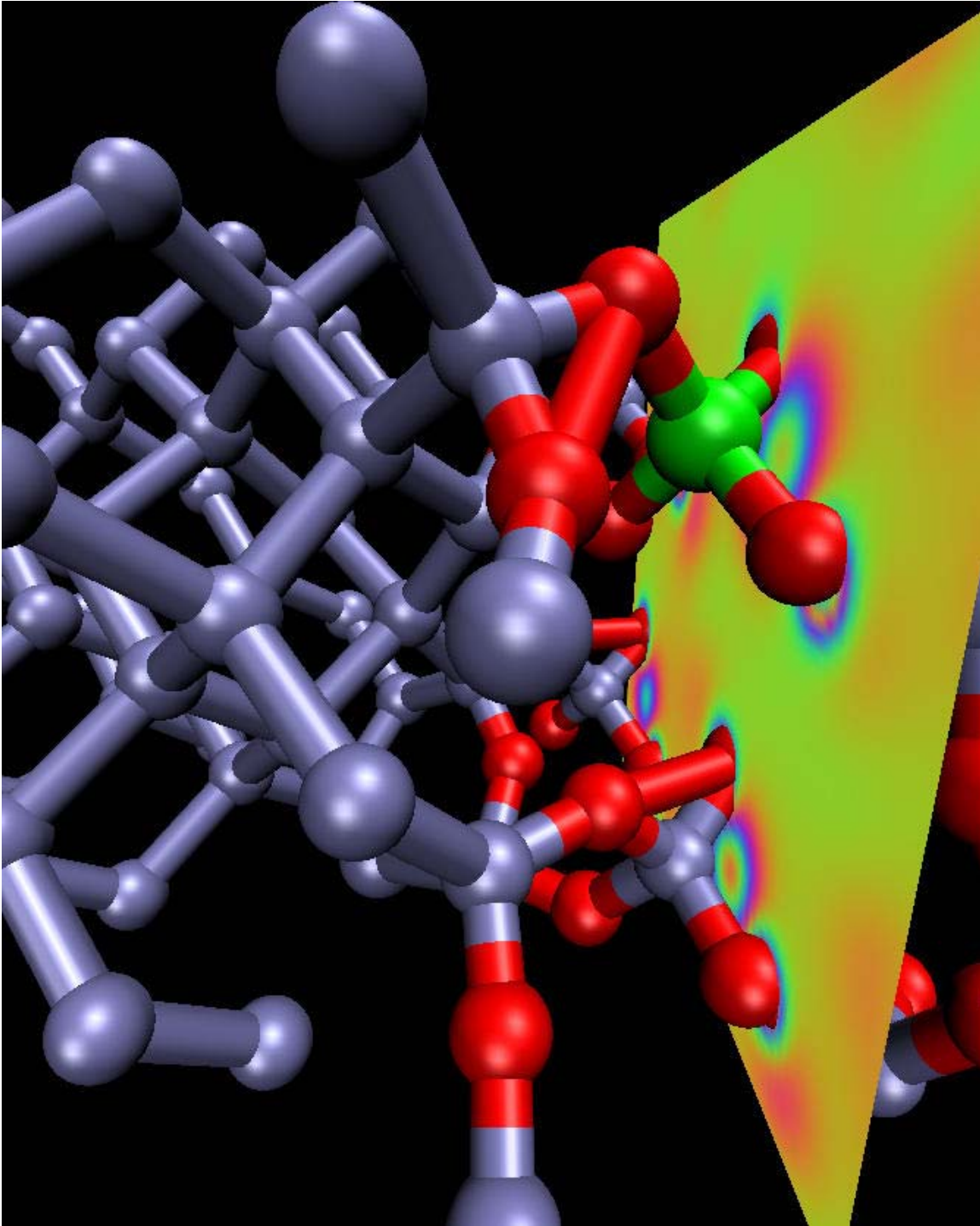


QCD Strong Scaling on up to 8192 processors



The Car-Parinello Molecular Dynamics (CPMD) is a computation and communication intensive application developed from the original Car-Parinello code by Alessandro Curioni and colleagues at IBM Research in Zurich. The CPMD code uses plane wave basis functions, 3D-FFTs and parallel linear algebra routines to study the electronics and structural properties of complex materials from first principles. This code is used worldwide with more than 5000 licenses issued. CPMD was implemented on Blue Gene to do a calculation of the Si/SiO₂ interface. This model is important because it allows us to improve the properties of gate oxides used in CMOS technology.

The electronic structure of a Si/SiO₂ interface is shown in the figure below which is a single snapshot of the simulation. The different atoms are Steel Blue for Silicon atoms, Red for Oxygen atoms, Green for Hafnium atoms - the slice is a contour map of the electrostatic potential.

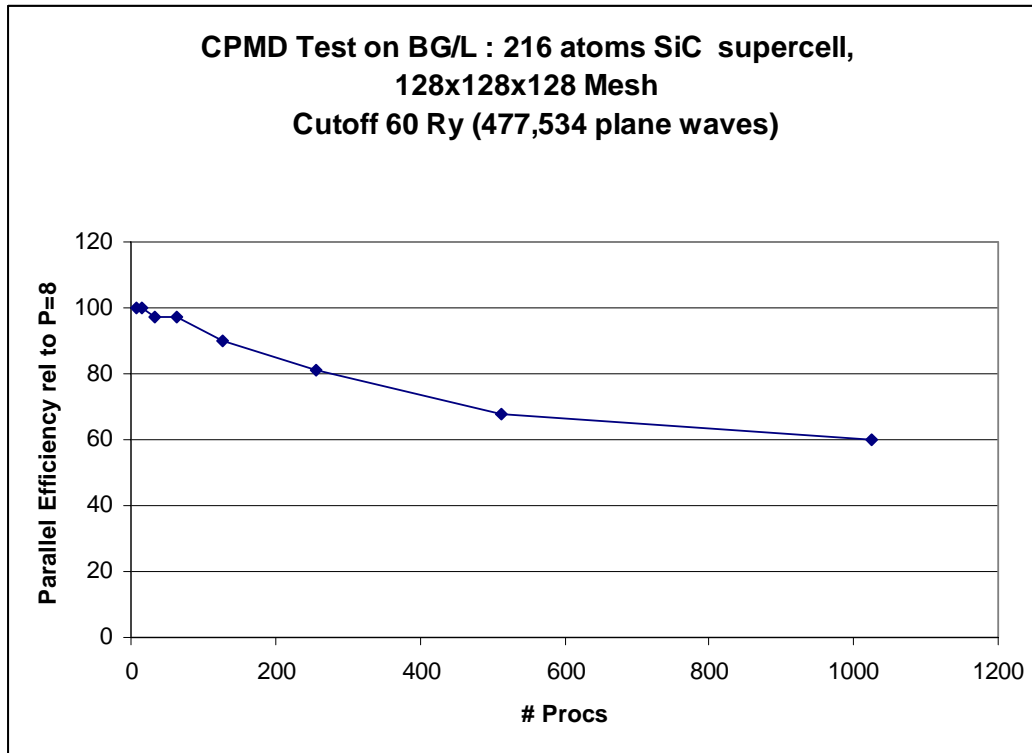


Si/SiO₂ interface

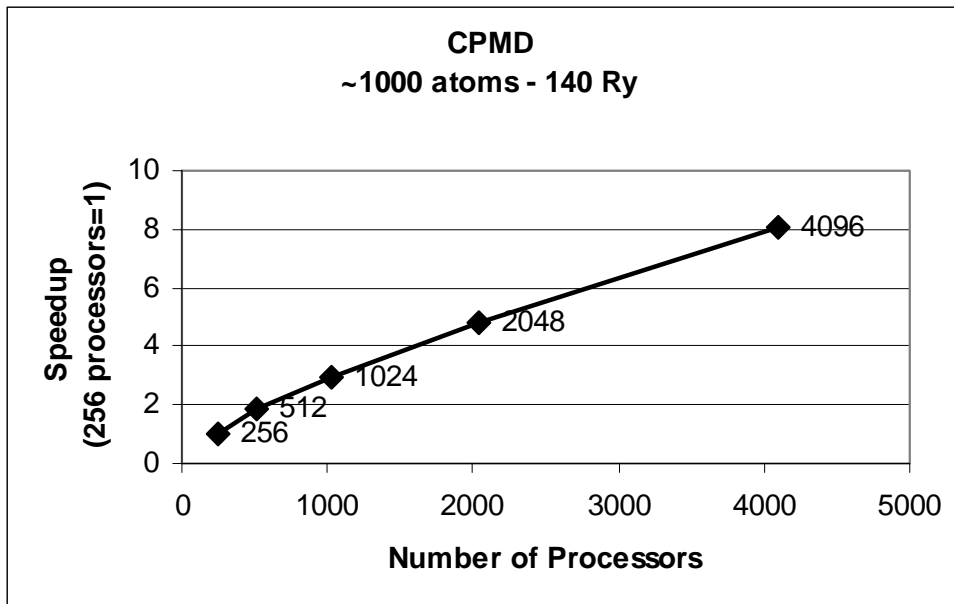
The performance of CPMD on Blue Gene is obtained through the use of optimized Double Hummer routines for most common kernels such as DGEMM for matrix multiplies (DGEMM), DCOPY, AZZERO and FFT on the processor. In addition, the scalability has been improved using a taskgroup implementation of

the FFT with a special mapping to the Blue Gene torus network. Results of this effort are shown in the following chart.

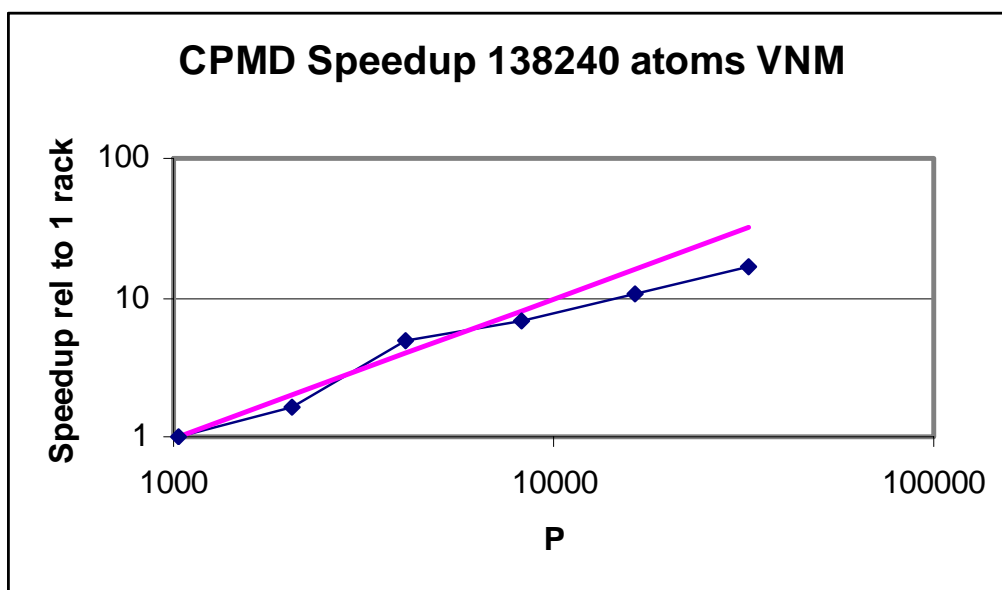
A relatively small system (216 atoms SiC supercell) was used in scaling tests up to 1024 processors, which is at the limit of the granularity of the system. This resulted in 60% parallel efficiency in moving from 1-1024 processors. The execution time on 1024 Blue Gene/L processors was less than 1 sec per step compared to 3.8 sec per step on a 1024 processor Regatta Cluster. This allows to simulate 10 ps/day fully ab-initio on a system of this size.



The CPMD was applied to problem that was published in Science “An ab Initio Molecular Dynamics Study of the Aqueous Liquid-Vapor Interface”, Kuo, I-F and Mundy, C, Science 303 (2004), 658-660. The Methanol-Water interface was studied with an 800 atom system and the CPMD run took 10 sec per step on 2048 Blue Gene/L processors. For this problem thanks to a better granularity of the data good scaling up to 4096 processors was obtained. The results are shown below:



Using a new classical Molecular Dynamics driver for modified Tersoff potentials (developed for QM/MM calculations), tests were done with 138K atoms on up to 32,000 processors using Virtual Node mode, where both CPU processors on the Blue Gene/L node are used for separate MPI tasks. On 32,000 processors, one can simulate 5ns per hour on this system. The speed up chart below is normalized to one Blue Gene rack. The magenta line shows perfect scaling.



Like so many applications, the possibility to scale CPMD on Blue Gene to large numbers of processors will open new frontiers. In the case of CPMD this will be in molecular sciences. It will increase, by an order of magnitude, the time window

that can be explored using ab-initio Molecular Dynamics. This is especially important for applications in Biology. For example, we expect that the use of Blue Gene will enable an accurate study of complex enzymatic reactions and molecule/protein interactions, which are essential for improved drug design. It will also allow for the modeling of bigger systems, with the potential to increase the size of the systems (number of atoms) by an order of magnitude. This is the key ingredient that will allow us to simulate complex systems such as the silicon/high-k oxide interface to improve device design.

The additional modeling time and larger system taken together will enhance the accuracy of molecular simulations in general, giving the predictive power necessary for use in "in-silico" material design and testing.

For more information

To learn more about the IBM @server Blue Gene Solution, please contact your IBM marketing representative or visit the following Web sites:
ibm.com/research/bluegene/index.html

Copyright material

© Copyright IBM Corporation 2004

IBM Corporation
Integrated Marketing Communications
Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States
November 2004
All Rights Reserved

This publication was developed for products and/or services offered in the United States. IBM may not offer the products, features or services discussed in this publication in other countries.

The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

IBM, the IBM logo, the e-business logo, @server, Blue Gene, DB2, LoadLeveler and PowerPC are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both. A full list of U.S. trademarks owned by IBM may be found at: ibm.com/legal/copytrade.shtml.

UNIX is a registered trademark of The Open Group in the United States, other countries, or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product and service names may be trademarks or service marks of others.

IBM hardware products are manufactured from new parts, or new and used parts. In some cases, the hardware product may not be new and may have been previously installed. Regardless, IBM warranty terms apply.

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.

Photographs show engineering and design models. Changes may be incorporated in production models.

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.