The IBM BlueGene (Supercomputer) Project Capability in Science Application (Molecular Dynamics and Protein Folding)

David Tin Win and Pornsak Budhraja
Faculty of Science and Technology, Assumption University
Bangkok, Thailand

Abstract

A scientific background on proteins and molecular dynamics precedes a technical description of the DOE’s IBM BlueGene/L supercomputer, which is the present and previous No. 1 in the TOP500 supercomputer list (announced November 2005). The system size has doubled, achieving a record Linpack performance of 280.6 TFlop/s - the only system to exceed the 100 TFlop/s mark. Molecular dynamics and protein folding are identified as promising science applications.

Keywords: BlueGene/C, BlueGeneL, BlueGene/P, BlueGene/Q, Cyclops64, molecular dynamics, protein folding, supercomputer, TOP500 supercomputer list.

1. Introduction

Alpha amino acids (α-amino acids) are the basic building blocks of proteins. They may be represented by a general molecular formula as NH₂-CH(R)-COOH, where NH₂- is an amine base functional group, -COOH is an acid functional group, and R is a general representation of alkyl groups such as CH₃-, CH₃.CH₂-, etc.

\[
\text{NH₂-CH(R)-COOH + NH₂-CH(R)-COOH} \rightarrow \text{monomer + monomer}
\]

\[
\text{NH₂-CH(R)-CONH-CH(R)-COOH + H₂O} \rightarrow \text{dimer (dipeptide)}
\]

The initial amino acids are monomers and the result is a dimer, called a dipeptide in protein terminology. The condensation process can carry on continuously resulting in a long chain polymer called a polypeptide. The polypeptides can undergo further condensation reactions to give chains that are even longer than the polypeptide chains. These are the protein chains.

Twenty different choices of amino acids exist for making up the protein chain; and each amino acid residue can occur in the chain many times. Thus protein structures are extremely varied. According to mathematical permutation and combination a 30 residue peptide can have 2030 possible unique sequences. But because of factors such as stability, folding ability and desired functions, only a few actually exist in nature. There are only 100,000 in humans.

Protein functions are essential items for creating and sustaining life. The functions are regulated by protein folding that is made by compacting of the protein chain into various conformations. The compacting processes are influenced by diverse chemical environments. The environments are influenced by intra-molecular or inter-molecular interactions of different neighboring groups. Factors such as pH, temperature and ions also produce different environments.

There are four levels of polypeptide and protein structures. The primary structure of a protein polypeptide determines its secondary, tertiary, and quaternary structures.

“Dysfunctional rendering, hyper or hypo amounts of protein or incorrect timing can
sometimes cause protein misfolding. This can lead to serious diseases such as Alzheimer's, Mad Cow, and cancer. Understanding the folding process allows discovery of treatments for misfolding-related disease, and present protein syntheses techniques. Thus understanding protein folding is of utmost importance (Chemistry.about. 2005).

According to common reaction rate theories such as the collision and activated complex (also called Transition State) Theories reactant molecules must collide with sufficient energy (activation energy) for a reaction to occur. Energy considerations (thermodynamic) determine if a reaction can occur spontaneously or not; but how and how fast it occurs (chemical kinetics) is determined by the movement of molecules or molecular dynamics. In other words what particular reaction takes place, what products are formed and how fast they are produced depend on molecular dynamics. Thus molecular dynamics is crucial to understanding chemical reactions. This in turn will allow design and control of chemical reactions to produce desired products in a fast efficient way. Kinetics of protein folding will also be dictated by molecular dynamics.

Protein folding occurs at a rapidly in the approximate timescale of 10,000 nanoseconds (1 ns = 1 x 10^-9 s). Thus the fundamental challenge to molecular dynamics and analyses that will lead to knowledge of the folding processes is computing power.

Increased computational power allows an increased ability to validate the simulation models used in prolonged probing of biological processes at the microscopic level.

To simulate the folding process on the required nano-timescale, a giant computation infrastructure is required. Two solutions have emerged to address this massive computing requirement (IBM Blue Gene Team 2001).

First Solution - Distributed computing or computing spread over the internet. A global inter-network of volunteers’ computers accepts small protein segments samples to be computed; the results are collected and the overall picture is compiled. University of Stanford and University of Washington protein folding research projects use this method.

Investing in high capacity super computers is unnecessary as workflows are decentralized - a major advantage of distributed computing. The drawback is capability limitation caused by the number of available volunteers.

Second Solution - A TeraFLOP/s (1012 Floating-Point operations per second) super computer provides enough number crunching capability. IBM and Lawrence Livermore National Laboratory (LLNL) have developed a TeraFLOP/s scale system, called BlueGene/L.

The BlueGene project was first announced in December 1999. It was intended for biomolecular phenomena studies such as protein folding. The goal was biomolecular simulations that are orders of magnitude larger than the then current technology permitted. “The mission of the BlueGene scientific program was to use large-scale biomolecular simulation in advancing the understanding of biologically important processes, in particular that of the mechanisms behind protein folding” (IBM Blue Gene Team 2001).

A variety of problems such as protein-drug interactions (docking), enzyme catalysis and protein structures refinement can be influenced by scientific knowledge gained from protein folding studies.

Exploring the supercomputing frontiers, namely computer architecture, and software (program and control massively parallel systems, computation in advancing the understanding of important biological processes such as protein folding) is the aspiration of the BlueGene project. Investigations include searching for the most effective utilization, enhanced use of such massively parallel machines, and using novel machine architectures for achieving performance targets.

Enhanced computational power provides an increased ability for validating models used in simulations. Probing microscopic level biological processes over long time periods is achieved by appropriate validation. According to the TOP500 supercomputer list announced in November 2005, Blue Gene/L (see Fig. 1) is number one. A Blue Gene solution is now offered by IBM. Current applications explored include hydrodynamics, quantum chemistry,
molecular dynamics, climate modeling and financial modeling (IBM Blue Gene Team 2001).

The TOP500 supercomputer list was started in 1993 as a basis for tracking and detecting trends in high-performance computing (IBM Blue Gene Team 2001). A listing of sites operating the 500 most powerful computer systems is released twice a year. The best performance on the Linpack benchmark is used as a measure performance for ranking the computer systems. The list contains a variety of information including the system specifications and its major application areas.

“The present and previous No. 1 is DOE’s IBM BlueGene/L system. It’s size has doubled again, reaching a record Linpack performance of 280.6 TFlop/s. It is the only system ever to exceed the 100 TFlop/s mark” (IBM Blue Gene Team 2001).

In summary, the BlueGene computer architecture project is designed for production of many next-generation supercomputers with operating speeds in the petaflops range. Sustained speeds of over 280 teraflops is now possible. It is a cooperative effort of IBM (particularly the Thomas J. Watson Research Center), the Lawrence Livermore National Laboratory, the United States Department of Energy, and academia.

There are four Blue Gene projects being developed: BlueGene/L, BlueGene/C, BlueGene/P, and BlueGene/Q.

This paper presents the capability of BlueGene/L in dealing with protein folding problems as evidence of the possibility of extension to other scientific problems. A scientific background on proteins and molecular dynamics is followed by a description of known technical specifications on BlueGene/L and selected promising applications.

2. The Scientific Background

The necessary scientific background on proteins and molecular dynamics, including general computing factors concerning the latter that are relevant to understanding the applications of BlueGene/L are presented below.

2.1 Proteins (chemistry.about, 2005; Folding@home, 2005a)

As mentioned above, α-amino acids undergo condensations to form polypeptides (linked with peptide bonds –CONH–), which condense further to form proteins. Protein structure can be extremely varied. Protein functions, regulated via various conformations of protein folding, are essential items for creating and sustaining life. Four levels of polypeptide and protein structures exist. The primary structure influences the other structures. Protein structures dictate their functions. Protein architecture is based on three principles: (chemistry.about, 2005).

1. Polymer chain formation
2. Native structure - folding of chain into a compact function-enabling structure.
3. Modification of the folded structure

2.1 The Primary Structure

The sequence of amino acids in the polypeptide chain with reference to the locations of any disulfide bonds is the primary structure of polypeptides and proteins. It describes all protein covalent bonding.

The standard three-letter abbreviations for representing amino acids are used to denote the amino acid sequence in showing the primary structure. For example: gly-ala-gly-ser denotes a primary structure of a polypeptide, which has the sequence of glycine, alanine,

---

3 These bonds cause cross-linking between polypeptide chains, resulting in a three dimensional network.
glycine and serine in that order, from the N-terminal amino acid (glycine) to the C-terminal amino acid (serine). The protein chain may be written as H₂N-(gly-ala-gly-ser)ₙ-COOH, where n is a very large integer. Note that the sequence gly-ala-gly-ser is repeated many times.

2.2 The Secondary Structure

The secondary structure is the conformation (ordered arrangement) of amino acids in localized regions of a polypeptide chain or protein molecule. Hydrogen bonding is the major factor in stabilizing these folding patterns. Being the two most stable conformers, the alpha helix and the anti-parallel beta-pleated sheet are the two main secondary structures. Polypeptides or proteins may contain multiple secondary structures.

![Fig. 2. Protein Structures. Here the unfolded peptide chain on the left already contains some folded secondary structure, alpha helices (red), and a beta hairpin (blue). It is still a long way from the compact native structure at right. The folding process in different proteins spans an enormous dynamic range from approximately 20 microseconds to approximately 1 second.](image)

Certain conformations, such as the alpha helix (red helices in Fig. 2), are favored by steric hindrance and energy requirements.

The a-helix is a right-handed or clockwise spiral with each peptide bond in a planar trans conformation. The –NH amine groups of the peptide bonds usually point upwards, (parallel to the helix axis); and the carbonyl groups >CO normally point downward.

The beta pleated sheet (blue flattened regions in figure 2) has strands of peptide chains aligned to form a sheet. It is an extensive network of polypeptide chains formed by planar trans peptide bonds. Neighboring chains extend anti-parallel to each other. Hydrogen bonding can occur between adjacent polypeptide chains, as the amine –NH and carbonyl groups >CO of peptide bonds are co-planar and point towards each other.

Stabilization of the helix is by hydrogen bonding between amine and carbonyl groups of the same polypeptide chain. The pleated sheet is stabilized by hydrogen bonds between the amine groups of one chain and the carbonyl groups of an adjacent chain.

The less organized loop or turn regions cement the relatively organized alpha helix and beta sheet sections.

2.3 The Tertiary Structure

The tertiary structure is the three-dimensional atomic arrangement within a single polypeptide chain - the way in which formation of the overall compact protein occurs by combination of localized secondary structure elements (an example is shown on the right in Fig. 2).

The tertiary structure is maintained by disulfide bonds (disulfide bridge -S-S-) like those formed between the side chains of cysteine by oxidation of two thiol groups (SH).

2.4 The Quaternary Structure

Tertiary structures from two or more chains combine to form much larger structures - quaternary structures. Proteins with multiple subunits or those with molecular masses >50,000 usually have such structures, composed of non-covalently-linked monomers. The three-dimensional arrangement of these monomers is the quaternary structure. For example, hemoglobin protein has a quaternary structure formed by package of its monomeric subunits. Four monomers of two different types make hemoglobin. It has a hetero-quaternary structure.

---

4 If the two end bonds of the peptide link are on opposite sides, CONH, it is a trans geometric isomer.

5 Multiple polypeptide molecules, each called a 'monomer'

6 If protein monomers are identical it is a homo-quaternary structure.
The main stabilizing force in quaternary structures is hydrophobic interaction. A three-dimensional shape is formed by folding of single monomers. Its polar side chains are exposed to an aqueous environment. Its non-polar side chains are protected. However some hydrophobic sections are still exposed to water. For protection, two or more monomers assemble so that their exposed hydrophobic sections are in contact and shielded.

2.5 The Protein Folding Problem

“Only knowing amino acid sequence gives us little information on what the protein does and how it does it. In order to perform their function (e.g. as enzymes or antibodies), they must take on a particular shape - a "fold." (Figure 2). Proteins are amazing machines: before they work, they assemble themselves! This self-assembly is called "folding." (Folding@home, 2005b).”

It is believed that protein misfolding results in diseases such as Alzheimer's, Huntington's, cystic fibrosis, BSE (Mad Cow disease), an inherited form of emphysema, and even many cancers. Protein misfolding results in "aggregates". These clumps often gather in the brain causing symptoms of Mad Cow or Alzheimer's disease (Folding@home 2005b).

“The two important items concerning the protein folding problem are prediction of three-dimensional structure from amino acid sequence, and understanding the mechanisms and pathways whereby the three-dimensional structure forms within biologically relevant timescales” (Thomasson, 2006; Bunk 1998).

The interaction of molecular chaperones, a variety of protein cofactors, cause folding of most newly formed cell proteins. These chaperone molecules prevent aggregation and misfolding. They recognize and bind to nascent polypeptide chains and partially folded protein intermediates.

Molecular chaperones include several families of unrelated proteins; many chaperones are also heat shock (stress) proteins. The following chaperone families are those most involved in protein folding: 40-kDa heat shock protein (HSP40; DnaJ), 60-kDa heat shock protein (HSP60; GroEL), and 70-kDa heat shock protein (HSP70; DnaK) families. Detailed understanding of the complex chaperone mechanisms, including the ATP-dependent reaction cycles of the GroEL and HSP70 chaperones (Fink 1999; Ellis 1990; Ruddon and Bedows 1997) were made possible via high-resolution structures.

The scientific knowledge derived from protein folding research can potentially be applied to a variety of related life science problems of scientific and commercial interest, (IBM Blue Gene Team, 2001) including:
- Protein-drug interactions (docking);
- Enzyme catalysis (through use of hybrid quantum and classical methods);
- Refinement of protein structures created through other methods.

3. Background on Molecular Dynamics

(US Government-NIH 2005; Wikipedia 2006a)

Molecular dynamics is a special discipline of molecular modeling and is concerned with molecular motion, which is present in all chemical processes. Simple vibrations, like bond stretching and angle bending, give rise to IR spectra. Many kinds of intramolecular and intermolecular motions are associated with chemical reactions such as hormone-receptor binding, and other complex processes.

Thermodynamics describes the driving force for chemical processes. Kinetics describes the mechanism by which chemical processes occur. Conformational transitions and local vibrations are common topics in molecular dynamics. Energy minimization by a step-wise alteration of intramolecular degrees of freedom is made. Individual energy minimization steps are directed only at establishing a down-hill direction to a minimum; but the molecular dynamic steps represent temporal changes in atomic positions or velocities of atoms.

Bond making and breaking are not involved in many aspects of protein folding.
Hence application of classical techniques, such as molecular dynamics (MD), is sufficient to model proteins in the Blue Gene project.

Inter-atomic interactions are considered in classical atomic approaches. Disulfide bonds influence structures in many proteins, but their formation is ignored.

Numerical solutions of Newton’s equations of motion on an atomic or similar model of a molecular system, is studied by molecular dynamics (MD) simulations. Information on temporal properties of molecular systems is compiled. The system atom forces acting on each other (described by Newton's equation) are dictated by atomic velocities. Initial atomic velocities (total kinetic energy of the system) are influenced by the desired simulation temperature. They are assigned arbitrarily. Then, the system initially at absolute zero was slowly ‘heated’; and the energy is allowed to equilibrate.

The molecular dynamics basic procedures are calculation of force on each atom, and estimation of individual atom positions. The latter is based on the force information obtained, over a specified time, in the order of picoseconds (1 ps = 1 x 10^{-12} s). The MD approach computes all forces on all protein atoms and solvent; then that force is used to compute the new atom positions in a very short subsequent time. “The positions of each atom along a series of infinitesimal small time steps \( dt \), order of femtoseconds (1s =10^{-15} s) are calculated based on atomic forces and masses. The resulting snapshot series of structural changes over time produce atomic coordinates as a function of time, called a trajectory between two states, initial and final states relating to the start and end times of the calculation”

“The atomic positions at time \( t + dt \) are predicted, using initial atomic positions at an initial time \( t \). The positions at \( t + dt \) are then used to predict the positions at \( t + 2dt \). This is reiterated to get positions at time \( t + ndt \), where \( n \) is a positive integer: a common numerical approach” (US Government-NIH, 2005; Wikipedia, 2006a).

To accurately describe the fastest vibrations of the protein and solvent system, small femtosecond time-step size is required. The computational requirements are enormous (Table 1). Some fast-folding systems have times of about \( 10^{-4} \) seconds; and need about \( 10^{11} \) MD time steps.

There is no defined point of termination in molecular dynamics, apart from the time that can be practically covered. “Unfortunately many simulations (large protein conformational transitions) require times longer than the picosecond time limit”.

Table 1. The computational effort required to study protein folding is enormous. Using crude workload estimates for a petaflop/second capacity machine leads to an estimate of three years to simulate 100 microseconds.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical time for simulation</td>
<td>( 10^{-4} ) seconds</td>
</tr>
<tr>
<td>Typical time-step size</td>
<td>( 10^{-15} ) seconds</td>
</tr>
<tr>
<td>Number of MD time steps</td>
<td>( 10^{11} )</td>
</tr>
<tr>
<td>Atoms in a typical protein and water simulation</td>
<td>32000</td>
</tr>
<tr>
<td>Approximate number of interactions in force calculation</td>
<td>( 10^9 )</td>
</tr>
<tr>
<td>Machine instructions per force calculation</td>
<td>1000</td>
</tr>
<tr>
<td>Total number of machine instructions</td>
<td>( 10^{23} )</td>
</tr>
</tbody>
</table>


MD utility include determination of protein structure and refinement; examination of the dynamics of atomic-level phenomena that cannot be observed directly (thin film growth and ion sub-plantation); and prediction of the physical properties of nanotechnology devices.

The interaction between objects is either described by a force field (classical MD), a quantum mechanical model (quantum MD), or a combination of the two. The interactions may also be described by the name of the theory used (chemist’s preference) or approximation used (physicist’s preference).

### 3.1 Design Constraints

Limits of computational power can be encountered in a molecular dynamics simulation design. Simulation sizes and time duration have to be selected for a practically acceptable completion time.
The simulation time duration is affected by the time-step length between which forces are recalculated. For the simulation effect to faithfully represent the real picture, the time-step has to be appropriately small. Also the number of time-steps must be large enough to capture the effect being modeled.

“The simulation size must be large so that the effect of boundary conditions is minimized, so that the behavior expressed is not disrupted. Treatment of boundary constraints involve choosing fixed conditions at the boundary, or choosing periodic boundary conditions in which one side of the simulation loops back to the opposite side. The simulation scalability with respect to the number of molecules is a significant factor in the range of simulation sizes that can be simulated in a reasonable time period”. In Big O notation, common molecular dynamics simulations usually scale by either $O(n \log(n))$, or with good use of neighbor tables, $O(n)$, with $n$ as the number of molecules.” (US Government-NIH 2005; Wikipedia 2006a)

A short digression on Big O notation: it was introduced by P. Brachmann in 1894. Consider real-valued functions $f(x)$ and $g(x)$ that are defined for all $x > x_0$ (fixed positive real). Then it is said

“$f(x)$ is big-O of $g(x)$” and written as $f(x) = O(g(x))$ provided there is some constant $C$ such that $|f(x)| < Cg(x)$. In other words, $f(x)$ is $O(g(x))$ if $f$ has a upper bound given by a constant times $g$. Eg: $53x^2+23x+500 = O(x^2)$, $\sin(x) = O(1)$ and any polynomial in $x$ of degree at most $n$ is $O(x^n)$.

3.2 Physical Principles

All the information needed to calculate system dynamics is contained in the system potential energy function $U$. The force on an atom $i$ in the system is given by $F_i = - \nabla_i U$. (Newtown's laws). Based on these forces, atom trajectories are calculated with an integrator such as the Verlet Integrator.

As temperature is a statistical quantity, system temperature estimation is one difficulty in MD calculations. Provided sufficiently large numbers of atoms were present, statistical temperature can be estimated from the instantaneous temperature, which is found from $KE = \frac{3kT}{2}$.

“Use of small number of atoms in MD simulations, create another temperature-related problem - a variety of thermostat methods are required to remove energy from the boundaries of an MD system in a realistic way.” (US Government-NIH 2005; Wikipedia 2006a)

3.3 Types of MD Systems.

The following system categories have been studied via MD:

3.3.1 Empirical Potentials: Physical properties such as elastic constants and lattice parameters of the atoms being simulated are fitted against potential parameters, which can further be subcategorized into pair potentials, and many-body potentials.

In pair potentials (e.g. Lennard-Jones or 6-12 potential) the total potential energy of a system is calculated from the sum of energy contributions of atom pairs.

An example of a many-body potential is the Tersoff Potential, used to simulate Silicon and Germanium. It involves summation over groups of three atoms. The angles between the atoms are important factors.

The Tight-Binding Second Moment Approximation (TBSMA) is another example. A sum of contributions from surrounding atoms is considered in calculating the electron density of states in the region of an atom. The potential energy contribution is a function of this sum.

3.3.2 Semi-Empirical Potentials: Matrix representation of quantum mechanics is used in semi-empirical potentials. Estimation of matrix elements is made by using empirical formulae, which determines the overlap degree of specific atomic orbitals. The occupancy of the different atomic orbitals is calculated by matrix diagonalization. Then empirical formulae are used to find orbital energy.
contributions. A wide variety of semi-empirical potentials exist. They are known as tight-binding potentials; and vary according to the atoms being modeled.

3.3.3 Ab-initio (First Principles) Methods:
Procedures that use full quantum mechanical formula to calculate the potential energy of a system (atoms or molecules) are called ab-initio (first principles) methods. Calculations are made "locally", for neighboring nuclei close to the reaction coordinate. Various approximations based on theoretical considerations (not on empirical fitting) may be used. Ab-Initio methods yield many information (not available from empirical methods) such as density of states. An example of a well-known ab-initio package is the Car-Parrinello Molecular Dynamics (CPMD) package based on the density functional theory.

3.3.4 Coarse-Graining: “Reduced Representations”:: MD simulations of very large systems require enhanced computer resources, and cannot be studied by traditional all-atom methods. Further more, long timescale (> 1 μs) simulation processes require many time-steps and are therefore prohibitively expensive. Reduced representations, called coarse-grained models, may be used to solve some of this problem. The models involve pseudoatoms representing atom groups instead of explicit representation of every individual system atom

Earliest MD simulations on proteins, lipids and nucleic acids used ‘United atoms’ (represent the lowest level of coarse graining). For example a single pseudoatom taken as a whole, represented all three H-atoms of a methyl group –CH3. One restriction is having the proper distance-dependence with reference to the pseudoatom’s van der Waals interactions with other groups. Bonds, angles, and torsions involving the pseudoatom have also to be considered. All explicit hydrogen atoms are ignored; excepting ‘polar hydrogen’ (those that are capable of forming hydrogen bonds). About 50% of protein or nucleic acid atoms are non-polar hydrogen. This representation is a computer time saver.

At higher levels coarse-graining may not be as accurate. The dynamic description may be less reliable. Fortunately very coarse-grained models have been used successfully. Empirical parameterization is required. It involves matching model behavior with appropriate experimental data.

Examples are:

- **Protein folding studies** - single (or a few) pseudoatoms per amino acid used
- **DNA supercoiling** - 1-3 pseudoatoms per basepair, or lower resolution used
- **Packaging of double-helical DNA into bacteriophage** - one pseudoatom to represent one turn (about 10 basepairs) of the double helix used
- **RNA structure in the ribosome and other large systems** - one pseudoatom per nucleotide used

Some available software: AMBER, CHARMM, CHARMm, NAMD, DL_POLY, GROMACS, LAMMPS, QUANTUM 3.1

4. BlueGene/L

**General Specifications** (TOP500, 2005):
- **Nodes**: 65,536
- **CPUs/node**: 2
- **Total CPUs**: 131,072
- **CPU speed (MHz)**: 667
- **Theoretical system peak performance (GFlop/s)**: 360,000
- **Memory/node**: 512 MiBytes
- **Total memory**: 32 TiBytes
- **Type of memory DDR DRAM**
- **Total disk space**: 800 TiBytes (global)

“Each node in the BlueGene/L system consists of a single ASIC (Application-Specific Integrated Circuit) and SDRAM-DDR memory. A network interface and small amount of on-chip memory can also be found in each ASIC. An on-chip memory controller enabled access to a larger external memory chips. Nodes were connected by low latency network using a combination of Ethernet topology, Tree topology, and 3D Torus network topology. The ASIC was implemented
with IBM's System-On-a-Chip technology, which enables BlueGene/L to be built extremely dense, and requires minimal power and cooling system”.

Blue Gene/L supercomputer is unique in the following aspects (IBM Blue Gene Team 2001; Wikipedia 2006b):

“Trade off between processor speeds and low power consumption. Two processor per node with two working modes: co-processor (1 user process/node: computation and communication work is shared by two processors) and virtual node (2 user processes/node); system-on-a-chip design a large number of nodes (65,536); three-dimensional torus interconnect with auxiliary networks for global communications, I/O, and management; Lightweight OS per node for minimum system overhead (computational noise)”.

A summary (Wikipedia 2006b) of the architecture is as follows: “a single ASIC with associated DRAM memory chips constitute each IO node or Compute. The ASIC combines two 700 MHz PowerPC 440 embedded processors, each with a double-pipeline-double-precision Floating Point Unit (FPU), a cache sub-system with built-in DRAM controller and the logic to support multiple communication sub-systems. Each BlueGene/L node is given a theoretical peak performance of 5.6 GFLOPS by the dual FPUs. Node CPUs are not cache coherent with one another”.

Each Compute or IO node dissipates only about 17 watts, including DRAMs; as all essential sub-systems are integrated on a single chip. For example: a 19” cabinet can hold up to 1024 Compute nodes plus additional IO nodes. Performance metrics in terms of FLOPS per Watt, FLOPS per m² of floorspace and FLOPS per unit, allows scaling to high performances.

Three parallel communications networks: (a 3D toroidal network for peer-to-peer communication between compute nodes, a collective network for collective communication, and a global interrupt network for fast barriers) are attached to each Blue Gene/L node. The I/O nodes, running the Linux operating system, provide world communication via an Ethernet network. A separate and private Ethernet network provides access to any node for configuration, booting and diagnostics.

Gbig/s Ethernet network connecting compute and I/O nodes; and the JTAG network for booting, control and monitoring purposes are also available.

Blue Gene/L Compute nodes use a minimal operating system supporting a single user program. To allow running multiple programs concurrently, Blue Gene/L system is partitioned into electronically isolated sets of nodes. Node number restrictions in each partition are: (a) a positive integer power of 2 (b) contain at least $2^5 = 32$ nodes.

For running a program, prior reservation of a partition of the computer is required. The program is then run on all the nodes within the partition, and no other program may access nodes within the partition while it is in use. Upon completion, the partition nodes are released for future programs to use. Component failures are unavoidable. The system is capable of electrically isolating faulty hardware so as to allow the machine to continue running.

BlueGene/C, renamed as Cyclops64 is a sister-project to BlueGene/L. It is a massively parallel, supercomputer-on-a-chip cellular architecture; and is slated for release in early 2007. BlueGene/P, architecturally similar to BlueGene/L, is expected to operate around one petaflop, and should be complete in 2008. Blue Gene/Q is expected to reach 3-10 petaflop.

5. Some Promising Scientific Applications

Some selected examples of applications to protein folding molecular dynamics studies are presented below.

5.1 Protein Folding Kinetics

Study of protein folding kinetics as a Markov chain consists of a rigorous formalism. The extraction of state-to-state transition functions from a Boltzmann-weighted ensemble of microcanonical molecular dynamics simulations was developed. The long time scale behavior of the folding process was predicted from a pool of short, independent
molecular dynamics simulations (Swope, Pitera and Suits, 2004).

Application to two different systems: a blocked alanine dipeptide in a vacuum and the C-terminal $\beta$-hairpin motif from protein G in water (Swope et al, 2004) was performed. The former displays some of the desired features, but the latter shows challenges in application to more complex biomolecular systems.

Analysis of long-lived states observed during $\beta$ - hairpin simulations suggest that folding processes like formation of non-native hydrogen bonds and salt bridge need to be considered. Figure 2 shows a folding process.

5.2 Molecular Dynamics of Lipid Layers

Investigation of dynamical property of phosphatidylethanolamine lipid bilayers by molecular dynamics was reported (Pitman, 2005a). A 20 ns molecular dynamics simulation described the dynamic behavior of 1-stearoyl-2-oleoyl-phosphatidylethanolamine (SOPE) bilayer. $^2$H spin-lattice relaxation rates, nuclear overhauser enhancement spectroscopy (NOESY) cross-relaxation rates, and lateral diffusion coefficients are used to characterize individual molecular dynamics. Lateral diffusion rates calculated are comparable (4±1×10$^{-8}$ cm$^2$/s); but are lower than pulsed field gradient NMR values.

Study of structural properties of phosphatidylethanolamine lipid bilayers by molecular dynamics was performed (Pitman, 2005b). Examination of a fully hydrated bilayer of 1-stearoyl-2-oleoyl-phosphatidylethanolamine involved using a 14 ns microcanonical (NVE) molecular dynamics simulation. Data were compared with experimental results and simulations of other lipids.

5.3 The free energy landscape

Energy landscape calculations are a theoretical approach to understanding protein folding since the late 1980s (chemistry.about, 2005). The protein energy landscape is the free energy variation as a conformation function, caused by the amino acid residue interactions.

Probable natural protein evolution is via a funnel shaped complicated energy surface that directs towards the native state or the lowest-energy conformation. Because of this "folding funnel" landscape the protein can fold to the native state through a large number of pathways. It is not restricted to a single mechanism. Computational simulations of model proteins support the theory. Improvement of methods for protein structure prediction and design has been made.

An example: the folding free energy landscape of the C - terminal $\beta$ - hairpin of protein G was explored with explicit solvent under periodic boundary conditions and OPLSAA force field. The landscape with respect to various reaction coordinates was found to be rugged at low temperatures and became a smooth funnel-like landscape at about 360 K (Zhou, Berne, and Germain. 2001).

6. Conclusion

The Blue Gene project represents a unique opportunity to explore novel research into a number of areas, including machine architecture, programming models, algorithmic techniques, and biomolecular simulations.

The applications demonstrate that BlueGene/L with its ASIC based on IBM's System-On-a-Chip technology is capable of tackling scientific studies that require large number crunching processes. It opens up a new exciting era for computer applications in Chemistry and Biochemistry.

Research involving competing architectures for future multicore processors is contemplated by competitors of IBM. For example, recently (February 2007) at the International Solid-State Circuits Conference in San Francisco, Intel Corp. demonstrated its Teraflop Research Chip: Polarisat.

"The 80-core chip crunches 1 trillion floating-point operations/second when running at a 3.2-GHz clock speed and consumes 62 watts, to yield a record 16 Gflops/watt. And by cranking the clock up to 5.6 GHz, the chip bested 1.8 teraflops—that's 80 percent faster—albeit by increasing power consumption fourfold to 265 W, or 6.8 Gflops/W." (Johnson 2007).
Acknowledgement

DTW and Pornsak Budhraja, a CT 1101 General Chemistry student in semester 2/2005, are indebted to the Dean of the Faculty of Science and Technology and the Chairperson of the Department of Information Technology (IT) for the opportunity to produce this article.

References


Folding@home. 2005a. On line retrieved from: http://folding.stanford.edu/science.html [Folding@home distributed computing: Scientific Background]

Folding@home, 2005b. On line retrieved from: http://folding.stanford.edu/FAQ-diseases.html [Folding@Home FAQ on diseases studied].


