


Blue Matter: Application Framework for molecular simulation on Blue Gene

Sarat Sreepathi.

IBM Blue Gene/L

- System: Blue Gene/L DD2 beta-System (0.7 GHz PowerPC 440)
 - Each node has two IBM Power PC processors (700 MHz)
- Rank 1 in Top 500 list.
- **Linpack Benchmark values:**
- Processors:32768
- Rpeak (GFlops):**91750**
- Rmax (GFlops):**70720**
- Nmax:933887

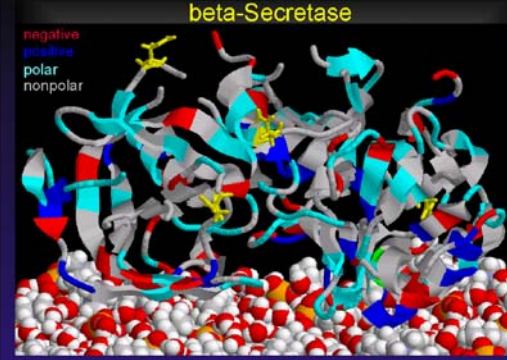
Blue Gene: A spectrum of possible projects



- cover a range of system sizes, topological complexity
- cover a broad range of scientific questions and impact areas:
 - thermodynamics
 - folding kinetics
 - membranes and membrane-bound systems
 - folding-related disease (CF, Alzheimer's, BSE)
- improve our understanding not just of protein folding but protein function

© 2000-2003 IBM Corporation

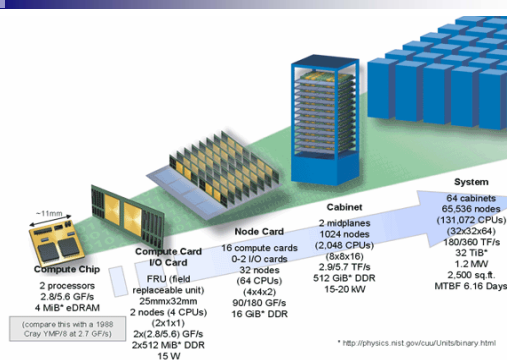
beta-Secretase



© 2000-2003 IBM Corporation

Blue Gene Architecture

- Blue Gene/L is a scalable ultra-computer targeted for 65,536 compute nodes.
- BlueGene/L is a **cellular architecture** in that the basic building block of the system can be replicated in a regular pattern, with no introduction of bottlenecks as the system is scaled up.
- Each BlueGene/L node consists of a single compute ASIC (an application-specific integrated circuit) and SDRAM-DDR memory chips.
- The compute ASIC is a complete system-on-a-chip, including all network interfaces and a modest amount of fast on-chip memory. An on-chip memory controller provides for access to larger external memory chips.



Compute Chip	Compute Card	Node Card	Cabinet	System
2 processors 2.85 G GF/s 4 MB* eDRAM	FRU (field replaceable unit) 2 nodes (4 CPUs) (2x1x1) 2x(2.85 G) GF/s 2x512 MB* DDR 15 W	16 compute cards 0-2 I/O cards (8x8x10) 32 nodes (64 CPUs) (4x4x2) 90/180 GF/s 16 GiB* DDR	2 midplanes 1024 nodes (2,048 CPUs) (32x32x64) 2.9/5.7 TF/s 512 GiB* DDR 15-20 MW	64 cabinets 65,536 nodes (131,072 CPUs) (32x32x64) 180/360 TF/s 32 TiB* 1.2 MW 2,500 sq.ft. MTBF 6.16 Days

* http://physics.mst.gov/oua/UnitsBinary.html

Blue Gene Architecture

- The nodes are interconnected through multiple complementary high-speed low-latency networks, including a 3D torus network and a combining tree network.
- The physical machine architecture is targeted to be most closely tied to the 3D torus, a simple 3-dimensional nearest neighbor interconnect which is "wrapped" at the edges.
- An independent combining tree network provides for fast global operations, such as global max or global sum.
- The ASIC that comprises the nodes is based on IBM's system-on-a-chip technology giving a very compact, low-power building block. This is used to create an extremely high compute-density system with very attractive cost performance and relatively modest power and cooling requirements.

Blue Matter, an application framework for molecular simulation on Blue Gene

B.G. Fitch,^a R.S. Germain,^a M. Mendell,^b J. Pitera,^c M. Pitman,^a A. Rayshubskiy,^a
Y. Sham,^a F. Suits,^a W. Swope,^c T.J.C. Ward,^a Y. Zhestkov,^a and R. Zhou^a

^a IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA
^b IBM Canada, 8200 Warden Avenue, Markham, Ont., Canada L6G 1C7
^c IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099, USA

Journal of Parallel and Distributed Computing
Received 4 December 2002; revised 4 April 2003; accepted 17 May 2003

Presented at JPDC 2003

What is Blue Matter ?

- Application Platform for the Blue Gene Science program.
- Prototyping platform for exploration of application frameworks suitable for cellular architecture machines.
- Blue Matter comprises all of the necessary application components—those that run on the computational core and those that run on the host.

Blue Matter: Design Goals

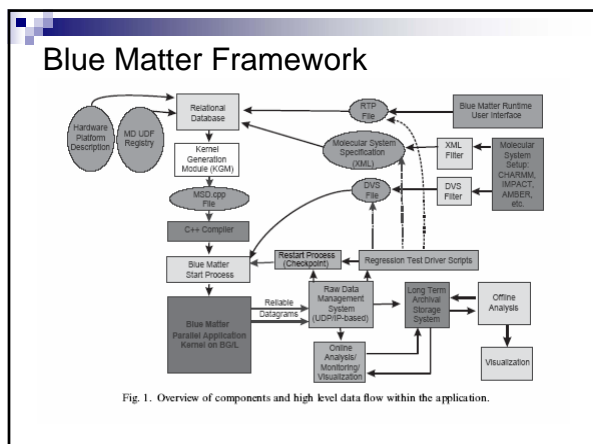
- Separating complexities of Molecular dynamics (MD) Simulation from those of parallel implementation on a particular machine.
- Minimizing system environmental dependencies
 - To enable use of non-pre-emptive low overhead kernels(for Scalability)
- Extensive Infrastructure for validation and binary regression.
- Modular architecture.

Molecular Simulation: Design Issues

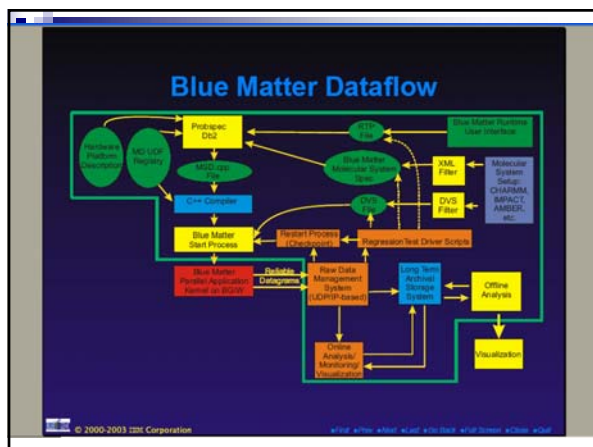
- Validity depends on
 - Representation of water.
 - Treatment of long-range intermolecular interactions.
 - Simulated environment of the protein.

Application Architecture

- Isolation of complexities of MD simulation from those of parallel programming achieved through
 - Modular decomposition and architected interfaces.
 - Use of generic programming concepts.
- Infrastructure for regression and validation.
- Functional correctness
 - By validating results of test simulations
 - Not by examining code



- ## Major Modules
- **Relational Database:**
 - Representation of the molecular system and corresponding simulation parameters is inserted into the database during setup.
 - Useful for analyses of simulation data.
 - **Kernel Generation Module (KGM)**
 - Retrieves information about a particular molecular system from the database and creates a Molecular system definition (MSD) file.
 - The MSD file contains C++ code which is then compiled and linked with framework source code libraries to produce an executable



- ## Setup Phase:
- Incorporate as much domain-specific knowledge as possible into setup phase itself.
 - All non-dynamic information is collected to generate source code specific to that run.
 - First, a commercial simulation package (like CHARMm , Impact, AMBER) is run to assign force field parameters for the molecular system.
 - Then a filter transforms this output into an XML file (a package-neutral form) which is loaded into the database.
 - Ability to execute ad-hoc queries to find molecular systems and parameter sets.

- ## KGM
- **KGM (Kernel generation Module)** generates a customized molecular system definition (MSD) file from the information in the database.
 - The MSD file contains parameters and C++ code containing only desired paths in the parallel kernel.
 - The runtime parameters are provided through the dynamic variable set (DVS)

- **BREAK:**
 - The image shows a portion of the surface of the reverse transcriptase enzyme of the HIV-1 virus with a molecular model of an inhibitor drug compound bound to the enzyme's receptor pocket.
-
- The image shows a complex 3D molecular model. It features a large, multi-colored protein structure (the reverse transcriptase enzyme) with a smaller, more detailed molecular model (the inhibitor drug compound) bound to a specific site (the receptor pocket) on its surface. The colors represent different atoms and their interactions within the protein structure.

Protein folding, illustrated using Chymotrypsin Inhibitor 2 (CI-2).



Generic MD framework

- Framework designed to manipulate domain function in fine grained functional units.(call back functions called UDFs)
- UDF (User defined functions) : Encapsulate localized domain specific data transformations.
- These are resolved at compile time.
- Implementing a new parallel decomposition does not need
 - Changing the UDFs
 - Restating the simulation semantics.

UDF

- This UDF interface abstracts the calling context, providing flexibility. E.g.. Tiling Invocations and Inline functions.
- Reuse of code by multiple UDFs. through
- UDF Adapters:
 - Module that implements framework interface and wraps another UDF.
 - Implemented as C++ template classes.
- UDF Helper:
 - A subroutine called by a UDF through a C++ method invocation interface.
 - Values are passed from the immediate UDF context.

Online Monitoring/Analysis :

- Blue matter avoids direct file I/O and instead communicates results through a raw datagram stream (RDG) to analysis tools listening on remote computers.
- Data is stored for two purposes
 - Trajectory analysis.
 - State information for simulation restart.

Interconnect Topologies

- Three dimensional mesh interconnect topology
- A second high-performance interconnect network to provide low latency and high bandwidth global broadcast and reduce.
- Three dimensional Torus network
- A torus was chosen because it provides high bandwidth nearest neighbor connectivity and due to scalability, cost and packaging considerations.
- A torus requires no long cables and, because the network is integrated onto the same chip that does computing, no separate switch is required.

Torus

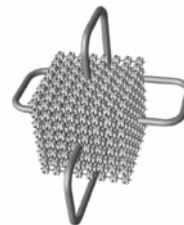
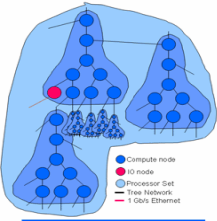


Fig. 2. View of a BC18 "torus" containing 512 nodes, showing the 3D mesh network.

- Each node connected to 6 neighbours.
- Link BW= 175 MB/s bidirectional (2 bits/cycle)
- Global combining/broadcast tree, BW=350 MB/s (4 bits/cycle)
- Latency=1.5 ms one-way latency.

Tree Topology



- ◆ Tree link is $2b+2b @ 1.4\text{GHz} = 350+350\text{MB/s}$ nodes
- ◆ 350 MB/s broadcast from IO node to 64 compute nodes
- ◆ 5.6+5.6 MB/s effective simultaneous point to point bandwidth/compute node
- ◆ I/O link to file system is 1 Gb/s Ethernet
- ◆ Achievable bandwidth below about 90+90 MB/s
- ◆ Aggregate tree BW through a node is 2.1 GB/s
- ◆ Achievable with multiple global combine operations
- ◆ Arithmetic operations implemented in tree
 - ◆ Integer/FP max/min
 - ◆ Integer add/subtract, bitwise logical ops
- ◆ Latency of tree less than 2.5 μs to top, additional 2.5 μs to broadcast to all
- ◆ Global sum over 64K in less than 2.5 μs (to top of tree)
- ◆ Partitioned with Torus boundaries
- ◆ Flexible local routing table

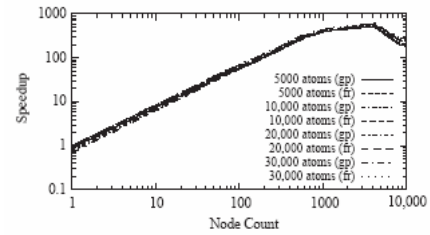
Parallel Decompositions

- Global force reduction
- Pair-wise interactions to be arbitrarily partitioned.
- Once per time step, all nodes – Summation all-reduction.
- Advantages:
 - Load balancing is straight forward for non-bonded forces.
 - Double computation of non-bonded forces is avoided.
- Disadvantages:
 - Replication of dynamics propagation- non parallelizable.
 - Force reduction is non-scalable on the global tree.

Parallel Decompositions

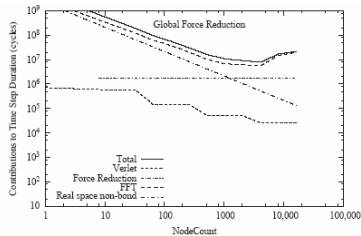
- 3D FFT
 - 3 all-to-all communications within a processor plane or row for each FFT.
 - Volume decomposition is used to map 3D mesh domain onto the machine.
- Global Position broadcast
 - Once per time-step the atom's position OR-reduced then broadcast.

- Speedup:
- 250 at 512 nodes and efficiency of ~50 % - Both decompositions

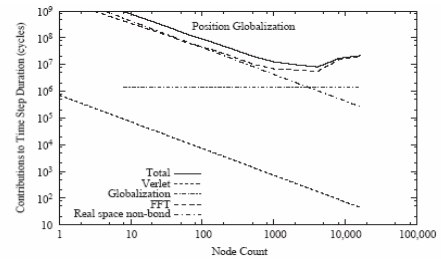


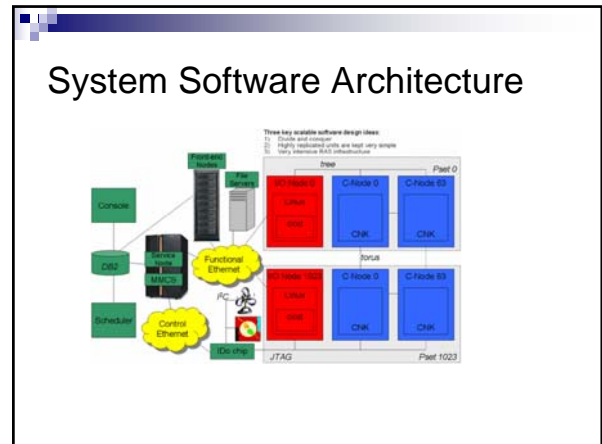
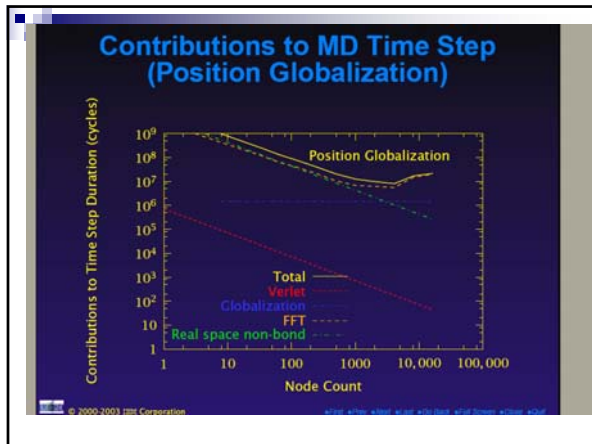
Global Force Reduction

- At very large node counts, performance of both schemes- limited by communications time required for the 3D FFT.



Position Globalization





Conclusion

- Well Organized
- Scalable solution to a grand challenge problem.