Blue Matter: Strong Scaling of Molecular Dynamics on Blue Gene/L

Blake G. Fitch[1], Aleksandr Rayshubskiy[1] Maria Eleftheriou[1], T.J. Christopher Ward[2], Mark Giampapa[1], Yuri Zhestkov[1], Michael C. Pitman[1], Frank Suits[1], Alan Grossfield[1], Jed Pitera[3], William Swope[3], Ruhong Zhou[1], Scott Feller[4], and Robert S. Germain[1]

[1] IBM Thomas J. Watson Research Center, 1101 Kitchawan Road/Route 134, Yorktown Heights, NY 10598, USA
[2] IBM Hursley Park, Hursley, Hursley SO212JN, United Kingdom
[3] IBM Almaden Research Center, 650 Harry Road, San Jose, CA
[4] Department of Chemistry, Wabash College, Crawfordsville, Indiana 47933

Abstract. This paper presents strong scaling performance data for the Blue Matter molecular dynamics framework using a novel n-body spatial decomposition and a collective communications technique implemented on both MPI and low level hardware interfaces. Using Blue Matter on Blue Gene/L, we have measured scalability through 16,384 nodes with measured time per time-step of under 2.3 milliseconds for a 43,222 atom protein/lipid system. This is equivalent to a simulation rate of over 76 nanoseconds per day and represents an unprecedented time-to-solution for biomolecular simulation as well as continued speed-up to fewer than three atoms per node. On a smaller, solvated lipid system with 13,758 atoms, we have achieved continued speedups through fewer than one atom per node and less than 2 milliseconds/time-step. On a 92,224 atom system, we have achieved floating point performance of over 1.8 TeraFlops/second on 16,384 nodes. Strong scaling of fixed-size classical molecular dynamics of biological systems to large numbers of nodes is necessary to extend the simulation time to the scale required to make contact with experimental data and derive biologically relevant insights.

1 Introduction

Blue Matter [1, 2, 3] is a molecular simulation framework and application developed to support the scientific goals of IBM's Blue Gene project [4], to serve as a platform for research into application programming patterns for massively parallel architectures, and to explore ways to exploit hardware features of the Blue Gene/L architecture. A major design goal for Blue Matter has been to achieve strong scaling of molecular dynamics for moderately sized systems (10,000 – 100,000 particles) to node counts corresponding to ratios of atoms per node of order one. This supports one of the aims of the scientific component of the project, to carry out simulations on a scale that allows meaningful comparisons with experimental data. Results on a 43,222 atom protein/lipid system obtained from early production use of prototype Blue Gene/L hardware were recently

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 846–854, 2006.

published in the Journal of the American Chemical Society [5], another paper on the same system in in press and larger scale studies including microsecond-scale simulations of solvated protein and membrane protein systems are currently underway.

The use of Blue Gene hardware to advance our understanding of biologically important processes has been an integral part of the Blue Gene mission from the beginning of the project [4]. As part of that strategy, we started an application effort to support the scientific goals of the project and to also act as a concrete test-bed for research into application development for massively parallel machines. The Blue Matter development effort has tracked the evolution of the machine architecture with the goal of exploiting hardware facilities on the target machine and enabling a systematic exploration of parallel decompositions for molecular dynamics. The most recent results of this exploration are described in this paper. As part of our efforts to exploit the current Blue Gene/L machine architecture [6] we have explored the following: (1) Use of the global collective network (2) Machine topology effects (3d-torus) [7] (3) Low level interfaces vs. MPI (4)Use of both processors on the BG/L compute chip.

The Blue Matter architecture requires infrastructure to support extensive regression and validation because of the aggressive and experimental nature of the computational platform we are targeting and because of its support for multiple force fields (the models and their parameters used for classical molecular simulation). The two main requirements are that the force field parameters be properly implemented, and that the integrator correctly measures the forces on each atom and makes the appropriate update of position and velocity for each time step. The JACS publication [5] was based on a 118ns NVE simulation of a membrane-bound protein, and the total and kinetic energy drift over that long period of simulation was negligible, indicating that there was consistently correct bookkeeping and integration of all the interaction forces.

2 Parallelization Strategies and Challenges

Classical molecular dynamics uses a model of the interactions between particles as the basis for a numerical integration of the equations of motion of the n-body system. In the case of biomolecular simulation, the existence of molecules with well-separated partial charges means that long range electrostatic interactions must be treated properly or unphysical behavior can be observed [8]. This issue is most commonly addressed through the use of periodic boundary conditions and the Ewald [9] or related mesh [10] techniques. Use of these techniques involves partitioning the computation of the long range forces into a real-space component that is short-ranged and a reciprocal space component. In the case of the mesh techniques, the reciprocal space component involves a convolution, implemented using three-dimensional FFTs, of the "meshed" charge distribution.

One of our goals for the Blue Matter framework was to allow a systematic exploration of parallelization strategies, progressing from the relatively straightforward to the more complex. Our starting point was a version of the

"replicated data" [11] approach that leveraged the Blue Gene/L hardware collective network (to globalize positions) as well as the torus (to perform a global force reduction) [2]. This allowed us to get out onto hardware very quickly and allowed us to create the testing and validation infrastructure that has been used from very early on the Blue Matter effort. While the replicated data approach makes load balancing straightforward, its scalability is limited by the performance of the floating point "all reduce" collective used for the forces.

In the current phase of this exploration, we use a minimal communication radius decomposition, a class of spatial decompositions which enables a very fine-grained decomposition of interactions and effective load balancing across a large number of nodes[3]. Our requirements included the ability to load balance based on pair interactions and the maintenance of locality for the real-space portion of the calculation so that a "natural" domain decomposition of the simulation volume onto the 3D torus layout of BG/L would minimize contention on the links. Currently, we assign the interaction between two particles to the node containing the mid-point of the pair of interacting particles although any node in the intersection of the broadcast spheres of the pair could carry out this computation and other algorithmic choices for the assignment are possible.

Another approach that combines a spatial decomposition with some of the advantages of the interaction decomposition invented by Plimpton and Hendrickson [11] is taken by the NAMD code[12] which has enabled this code to scale to over 1500 processors[13] prior to the advent of Blue Gene/L and more recently to 8192 nodes on BG/L[14]. Recently, two additional approaches that combine spatial and interaction decompositions have been proposed [15, 16], but neither outlines a detailed strategy for dealing with load imbalance and to our knowledge, no published performance results on biomolecular systems using either technique are available at this time.

Our decomposition requires many-to-many personalized communication operations [17] that are not efficiently represented by collective operations within the MPI standard. These operations entail each node concurrently originating a broadcast of positions to a local neighborhood and a corresponding concurrent reduction of computed forces back to the originating node. Given a three-dimensional simulation domain and a three-dimensional torus interconnect, communication locality on the machine can be achieved via a "natural" mapping of the simulation domain onto the machine.

Although MPI only allows a task to participate in one collective operation at a time, it is possible to construct equivalent function within the standard in several ways:

1. Sequentially invoking MPI broadcast/reduce collectives with common members (collectives involving disjoint task groups can proceed concurrently).
2. Using `ISEND/IRECV` to implement the same communication function in a non-serialized fashion.
3. Use of `ALLTOALLV` on `MPI_COMM_WORLD` with many node pairs transferring no data to implement the same communication function while avoiding MPI internal message-handling overhead. This can be achieved by using a lower

overhead messaging protocol within the implementation of `ALLTOALLV` rather than just using a set of `ISEND/IRECV` calls.

We have implemented the second and third options and have found that as a result of optimizations of the MPI collectives for BG/L [20], the third option gives superior performance. Even so, the realized performance on MPI does not yet reflect the full capabilities of the hardware. We have implemented the collective operations required by Blue Matter via the low-level System Programming Interface (SPI) of the Blue Gene/L Advanced Diagnostics Environment [21]. This is the environment used by the BG/L hardware team to test and validate hardware performance. Results comparing the performance of one communication-intensive kernel, the 3D FFT, on both the MPI and SPI communication layers have been presented previously[22] and measurements in the context of Blue Matter of the time taken by the FFT as well by the neighborhood broadcast and reduce are provided in Table 2.

Table 1. Details about the systems benchmarked with Blue Matter. Runs on SOPE and Rhodopsin were made with the velocity Verlet integrator [18] while the ApoA1 runs used RESPA[19]. All used the a particle mesh technique (P3ME or PME) to handle long range electrostatic interactions and were constant particle number, volume, and energy (NVE) simulations.

System	Total Atoms	Cutoff/Switch (Å)	P3ME Mesh
SOPE	13,758	9.0/1.0	128^3
Rhodopsin	43,222	9.0/1.0	128^3, 64×128^2
ApoA1	92,224	10.0/2.0	128^3

3 Performance Results

We present detailed performance scaling results for three molecular systems whose sizes span the range from 10,000 to 100,000 atoms in Table 2 for both MPI and BG/L ADE SPI communications protocols. Two of these systems, a small solvated lipid bilayer system, SOPE, and a solvated membrane protein system, Rhodopsin, have been used in published scientific work by the Blue Matter science team[5, 23, 24]. The third, the 92,224 atom ApoA1 solvated lipoprotein system, is used as a benchmark by the NAMD package[13]. Additional details about these systems are provided in Table 2.

A summary plot of the strong scaling behavior of Blue Matter is shown in Figure 1, in which the computational rate in time-steps per hour is plotted as a function of atoms per node. This scaling plot shows that use of the BG/L ADE SPI communications interfaces allows continued performance gains to values of atoms per node well below those achievable using MPI. The MPI implementation on Blue Gene/L[20] is quite good as evidenced by the results achieved on the 3D-FFT[22], but the scalability of Blue Matter using MPI appears to be limited by the performance of the neighborhood broadcast and reduce collectives discussed above as can be seen in Table 2.

Table 2. Tabulated performance data for the various molecular systems described in this paper using Blue Matter on BG/L. The total time per time-step and selected components are provided for both MPI and BG/L ADE SPI implementations. All of these data were taken in a dual core mode in which k-space and real-space operations are carried out on separate cores and used a complex to complex single precision 3D-FFT. The SOPE and Rhodopsin benchmarks used the velocity Verlet integrator and carried out the FFT-based P3ME operations on every time-step while the ApoA1 benchmark used RESPA[19] and only carried out P3ME on every fourth time-step. Performance is reported for two different processor mesh geometries at 4096 nodes because communication-intensive operations can be sensitive to the aspect ratio of the processor mesh.

Nodes				Time/time-step (milliseconds)								Atoms/
				Total		FFT		Bcast		Reduce		
Total	P_x	P_y	P_z	MPI	SPI	MPI	SPI	MPI	SPI	MPI	SPI	Node
512	8	8	8	7.46	6.81	2.17	1.88	0.55	0.35	0.44	0.35	26.8
1024	8	8	16	5.24	4.29	1.33	1.21	0.63	0.30	0.52	0.29	13.4
2048	8	16	16	4.66	2.80	0.93	0.77	0.87	0.24	0.85	0.22	6.7
2048	16	8	16	4.55	2.81	0.91	0.70	0.87	0.24	0.77	0.23	6.7
4096	16	16	16	5.07	1.94	0.71	0.46	1.39	0.22	1.30	0.21	3.3
4096	8	32	16	5.61	2.56	1.00	0.78	1.38	0.24	1.32	0.23	3.3
8192	16	32	16	7.31	1.88	0.84	0.52	2.48	0.22	2.20	0.20	1.6
16384	32	32	16	12.51	1.84	0.95	0.50	4.88	0.22	4.29	0.20	0.8

(a) SOPE (13,758 atoms)

Nodes				Time/time-step (milliseconds)								Atoms/
				Total		FFT		Bcast		Reduce		
Total	P_x	P_y	P_z	MPI	SPI	MPI	SPI	MPI	SPI	MPI	SPI	Node
512	8	8	8	16.77	16.82	1.27	1.97	0.76	0.47	0.53	0.50	84.4
1024	8	8	16	9.42	9.49	0.91	1.28	0.77	0.38	0.58	0.38	42.2
2048	8	16	16	6.45	5.57	0.73	0.81	0.94	0.34	0.77	0.33	21.1
2048	16	8	16	6.40	5.61	0.70	0.76	0.94	0.34	0.75	0.33	21.1
4096	8	32	16	5.82	3.54	0.78	0.81	1.44	0.28	1.23	0.25	10.5
4096	16	16	16	5.56	3.47	0.59	0.50	1.42	0.28	1.15	0.27	10.5
8192	16	32	16	7.17	2.50	0.75	0.54	2.46	0.24	2.04	0.22	5.2
16384	32	32	16	12.88	2.27	0.96	0.52	5.29	0.25	4.52	0.24	2.6

(b) Rhodopsin (43,222 atoms)

Nodes				Time/time-step (milliseconds)								Atoms/
				Total		FFT		Bcast		Reduce		
Total	P_x	P_y	P_z	MPI	SPI	MPI	SPI	MPI	SPI	MPI	SPI	Node
512	8	8	8	35.51	36.30	0.59	0.53	1.04	1.07	0.67	0.97	180
1024	8	8	16	19.28	19.17	0.36	0.33	1.02	0.73	0.71	0.75	90
2048	16	8	16	11.34	10.70	0.25	0.19	1.04	0.59	0.84	0.55	45
4096	16	16	16	7.53	5.96	0.19	0.13	1.37	0.50	1.22	0.47	22.5
4096	8	32	16	8.57	6.31	0.26	0.20	1.67	0.50	1.53	0.46	22.5
8192	16	32	16	7.32	3.67	0.21	0.14	2.21	0.43	2.15	0.38	11.3
16384	32	32	16	11.82	2.50	0.24	0.13	4.83	0.39	4.79	0.34	5.6

(c) ApoA1 (92,422 atoms)

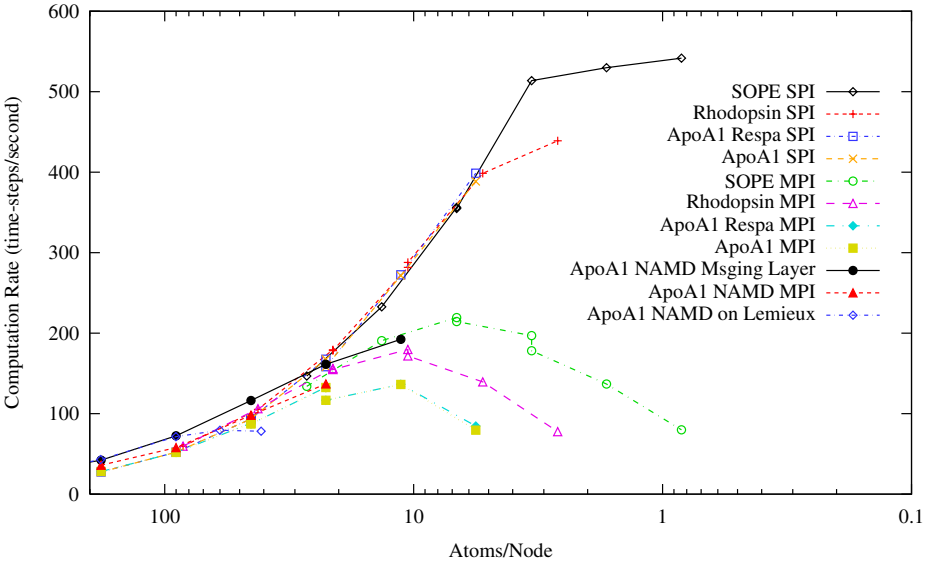

Fig. 1. This figure compares the strong scaling characteristics of the Blue Matter MPI and BG/L ADE SPI-based implementations for the molecular systems described in Table 2. The simulation parameters employed for the SOPE and Rhodopsin systems were those used in production, specifically a Verlet integrator with P3ME being computed on every time-step. For ApoA1, we attempted to match the parameters used by the NAMD benchmark as closely as possible, but we were forced to use a larger, $128 \times 128 \times 128$, mesh for our FFT because our current implementation is restricted to dimensions that are powers of two. This plot also includes data reported for NAMD on Blue Gene/L [14] and for the Lemieux Alpha system at the Pittsburgh Supercomputing Center [13]. Because benchmarking data for NAMD was not available for the SOPE and Rhodopsin system, the only comparison possible was on the 92,224 atom ApoA1 system. The NAMD benchmarks used a smaller FFT mesh size ($108 \times 108 \times 80$) than that used by Blue Matter (128^3) and on the Lemieux system used a specially tuned version of the Charm++ library written to the Elan communication library provided by Quadrics. On Blue Gene/L, NAMD results are shown for both a specially tuned version of the Charm++ library using a custom messaging layer and for an MPI-based implementation of Charm++[14].

Table 3. Floating point performance of Blue Matter on the ApoA1 system using multiple time-stepping (P3ME every four time steps) derived from measurements using floating point performance counters in the BG/L chip

Node Count	512	1024	2048	4096	8192	16384
GFLOP/sec.	127	241	433	774	1258	1846

As a way to place Blue Matter running on Blue Gene/L in context, Figure 1 also shows published results using the NAMD package [12] on the Lemieux system at the Pittsburgh Supercomputing Center [13] and on Blue Gene/L [14]

and Blue Matter on Blue Gene/L. The results on Lemieux were obtained using a version of the Charm++ library written to the Elan communication library provided by Quadrics. The systems benchmarked by Blue Matter and NAMD (ApoA) are identical, and we have made every effort to use either the same (cut-off distances) or higher cost (FFT mesh size) parameters in the Blue Matter runs as were used in the NAMD study to get as close as possible to an "apples-to-apples" comparison. Also, our comparison was made using a multiple time step integration technique that only carried out the P3ME operations once in every four time steps because this was the mode that gave NAMD the best performance on the PSC Lemieux system. Table 3 gives the realized floating point performance for Blue Matter running on the ApoA1 system using the BG/L ADE SPI communication protocols.

4 Summary and Conclusions

We have presented strong scaling data on biomolecular systems spanning a range of sizes using two communications protocols (MPI and the BG/L ADE SPI interface) on Blue Gene/L and we have compared our performance results on the largest system, ApoA1 with published values obtained using the NAMD code on Blue Gene/L and on the PSC Lemieux system. Using Blue Matter on BG/L with communications via the BG/L ADE SPI interface, we have achieved under 2.3 milliseconds per time-step on 16,384 nodes for a 43,222 atom protein/lipid system. The continued speed-up through values of less than three atoms/node on this system and to under one atom/node on the smaller SOPE system is the first time that this level of strong scaling has been obtained with classical biomolecular simulation.

The improvement in performance over the MPI baseline obtained through use of the SPI communications interface shows the advantages that can be realized through use of application-aware communications collectives that fully leverage the available hardware capabilities.

The time-to-solution measured for the 43,222 atom rhodopsin system on 16,384 nodes corresponds to over 76 nanoseconds of simulation time per day or a microsecond of simulation in only two weeks. This capability enables studies of biologically relevant systems on time-scales that were previously impractical. Scientific results using Blue Matter on prototype BG/L hardware have already been published and additional scientific studies are underway.

Work is currently underway to explore further optimizations of the 3D-FFT, such as implementing a real FFT to reduce communication data volume below that of the current single precision complex FFT implementation and to continue to improve the performance of neighborhood broadcast and reduce operations. We are also continuing to refine our load balancing techniques and are working with the compiler team to improve the floating point efficiency of the Blue Matter code.

References

1. Fitch, B., Germain, R., Mendell, M., Pitera, J., Pitman, M., Rayshubskiy, A., Sham, Y., Suits, F., Swope, W., Ward, T., Zhestkov, Y., Zhou, R.: Blue Matter, an application framework for molecular simulation on Blue Gene. Journal of Parallel and Distributed Computing **63** (2003) 759–773
2. Germain, R., Zhestkov, Y., Eleftheriou, M., Rayshubskiy, A., Suits, F., Ward, T., Fitch, B.: Early performance data on the Blue Matter molecular simulation framework. IBM Journal of Research and Development **49**(2/3) (2005) 447–456
3. Germain, R.S., Fitch, B., Rayshubskiy, A., Eleftheriou, M., Pitman, M.C., Suits, F., Giampapa, M., Ward, T.C.: Blue Matter on Blue Gene/L: massively parallel computation for biomolecular simulation. In: CODES+ISSS '05: Proceedings of the 3rd IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis, New York, NY, USA, ACM Press (2005) 207–212
4. Allen, F., et al.: Blue Gene: a vision for protein science using a petaflop supercomputer. IBM Systems Journal **40**(2) (2001) 310–327
5. Pitman, M.C., Grossfield, A., Suits, F., Feller, S.E.: Role of cholesterol and polyunsaturated chains in lipid-protein interactions: Molecular dynamics simulation of rhodopsin in a realistic membrane environment. Journal of the American Chemical Society **127**(13) (2005) 4576–4577
6. Gara, A., et al.: Overview of the Blue Gene/L system architecture. IBM Journal of Research and Development **49**(2/3) (2005) 195–212
7. Adiga, N., et al.: Blue Gene/L torus interconnection network. IBM Journal of Research and Development **49**(2/3) (2005) 265–276
8. Bader, J., Chandler, D.: Computer simulation study of the mean forces between ferrous and ferric ions in water. The Journal of Physical Chemistry **96**(15) (1992)
9. De Leeuw, S., Perram, J., Smith, E.: Simulation of electrostatic systems in periodic boundary conditions I. lattice sums and dielectric constants. Proc. Roy. Soc. Lond. A **373** (1980) 27–56 and references therein.
10. Deserno, M., Holm, C.: How to mesh up ewald sums. i. a theoretical and numerical comparison of various particle mesh routines. J. Chem. Phys. **109**(18) (1998) 7678–7693
11. Plimpton, S., Hendrickson, B.: A new parallel method for molecular dynamics simulation of macromolecular systems. Journal of Computational Chemistry **17**(3) (1996) 326–337
12. Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., Schulten, K.: NAMD2: Greater scalability for parallel molecular dynamics. Journal of Computational Physics **151** (1999) 283–312
13. Phillips, J., Zheng, G., Kumar, S., Kale, L.: NAMD: biomolecular simulation on thousands of processors. In: Supercomputing 2002 Proceedings. (2002) `http://www.sc2002.org/paperpdfs/pap.pap277.pdf`.
14. Kumar, S., Huang, C., Almasi, G., Kale, L.V.: Achieving strong scaling with NAMD on Blue Gene/l. 20th IEEE International Parallel and Distributed Processing Symposium, IEEE (2006) http://charm.cs.uiuc.edu/papers/NAMDIDPDS06.pdf.
15. Snir, M.: A note on n-body computations with cutoffs. Theory of Computing Systems **37** (2004) 295–318 DOI: 10.1007/s00224-003-1071-0.
16. Shaw, D.E.: A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. Journal of Computational Chemistry **26**(13) (2005) 1318–1328

17. Kale, L., Kumar, S., Varadarajan, K.: A framework for collective personalized communication. In: Parallel and Distributed Processing Symposium, 2003. Proceedings. International, IEEE (2003) http://dx.doi.org/10.1109/IPDPS.2003.1213166.
18. Swope, W., Andersen, H., Berens, P., Wilson, K.: A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. Journal of Chemical Physics **76** (1982) 637–649
19. Tuckerman, M., Berne, B., Martyna, G.: Reversible multiple time scale molecular dynamics. J. Chem. Phys. **97**(3) (1992) 1990–2001
20. Almasi, G., et al.: Design and implementation of message-passing services for the Blue Gene/L supercomputer. IBM Journal of Research and Development **49**(2/3) (2005) 393–406
21. Giampapa, M., et al.: Blue Gene/L advanced diagnostics environment. IBM Journal of Research and Development **49**(2/3) (2005) 319–332
22. Eleftheriou, M., Fitch, B., Rayshubskiy, A., Ward, T., Germain, R.: Performance measurements of the 3d FFT on the Blue Gene/L supercomputer. In Cunha, J., Medeiros, P., eds.: Euro-Par 2005 Parallel Processing: 11th International Euro-Par Conference, Lisbon, Portugal, August 30-September2, 2005. Volume 3648 of Lecture Notes in Computer Science., Springer-Verlag (2005) 795–803
23. Suits, F., Pitman, M.C., Feller, S.E.: Molecular dynamics investigation of the structural properties of phosphatidylethanolamine lipid bilayers. Journal of Chemical Physics **122**(24) (2005)
24. Pitman, M.C., Suits, F., Gawrisch, K., Feller, S.E.: Molecular dynamics investigation of dynamical properties of phosphatidylethanolamine lipid bilayers. Journal of Chemical Physics **122**(24) (2005)