PROCEEDINGS OF SCIENCE

Large-scale lattice-Boltzmann simulations over lambda networks

Radhika S. Saksena*
Centre for Computational Science, Department of Chemistry, University College London
20 Gordon Street, London WC1H 0AJ, United Kingdom
E-mail: r.saksena@ucl.ac.uk

Peter V. Coveney
Centre for Computational Science, Department of Chemistry, University College London
20 Gordon Street, London WC1H 0AJ, United Kingdom
E-mail: p.v.coveney@ucl.ac.uk

Robin L. Pinning
Manchester Computing, University of Manchester
Oxford Road, Manchester M13 9PL, United Kingdom
E-mail: pinning@manchester.ac.uk

Stephen P. Booth
Edinburgh Parallel Computing Centre, University of Edinburgh
Edinburgh EH9 3JZ, Scotland, United Kingdom
E-mail: s.booth@epcc.ed.ac.uk

Amphiphilic molecules are of immense industrial importance, mainly due to their tendency to align at interfaces in a solution of immiscible species, e.g., oil and water, thereby reducing surface tension. Depending on the concentration of amphiphiles in the solution, they may assemble into a variety of morphologies, such as lamellae, micelles, sponge and cubic bicontinuous structures exhibiting non-trivial rheological properties. The main objective of this work is to study the rheological properties of very large, defect-containing gyroidal systems (of up to 1024^3 lattice sites) using the lattice-Boltzmann method. Memory requirements for the simulation of such large lattices exceed that available to us on most supercomputers and so we use MPICH-G2/MPIg to investigate geographically distributed domain decomposition simulations across HPCx in the UK and TeraGrid in the US. Use of MPICH-G2/MPIg requires the port-forwarder to work with the grid middleware on HPCx. Data from the simulations is streamed to a high performance visualisation resource at UCL (London) for rendering and visualisation.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)012

1. Introduction

Our objective is to simulate the rheological properties of ternary amphiphilic mixtures which undergo self-assembly into cubic and non-cubic periodic structures. Lattice-Boltzmann simulations of cubic and non-cubic self-assembled mesophases exhibit interesting rheological behaviour. Such discoveries can be exploited to design functional materials with specific rheological properties. Self-assembled mesophases are also finding application in the synthesis of mesoporous nanomaterials which have interesting structural and electronic properties. These simulations involve traversing complex parameter spaces in order to identify regions where self-assembled mesophases are formed. Self-assembly phenomena are known to suffer from hysteresis whereby a meta-stable state can persist for long times and one requires long simulations to identify the final self-assembled phase. Additionally, defects in the periodic mesophases are known to significantly influence rheological properties: in order to simulate defect dynamics correctly, one needs to perform simulations of large system sizes which can be deemed to be free of finite-size effects. Thus, in order to simulate physically realistic system behaviour, large-scale and long simulations need to be performed, making this an extremely computationally demanding endeavour.

2. Simulation Code

We use the lattice-Boltzmann code, LB3D, to perform large-scale lattice-Boltzmann simulations. The implementation of the lattice-Boltzmann model for ternary amphiphilic fluids in LB3D has been discussed previously [1]. LB3D correctly simulates the self-assembly dynamics [2] and rheology [3] of cubic and lamellar mesophases in these systems. The code has been under development for over 7 years and is widely deployed on the UK NGS [4], the UK supercomputing resource HPCx [5] and on various supercomputing resources on the US TeraGrid [6]. LB3D is a scalable parallel MPI code written in Fortran 90. It has been parallelised according to the domain decomposition scheme [7]. LB3D is a memory intensive application code requiring approximately 1 kilobyte of memory per lattice-site to store state data. The compute-intensive part of the algorithm consists of the collision and propagation steps. Because of the non-local interaction forces between different species in the amphiphilic mixture, two communication steps per cycle are required to exchange state data for lattice-sites on the sub-domain boundaries between neighbouring processors. LB3D checkpoints system state and visualisation data-sets at regular intervals. For large lattices, the size of checkpoints becomes non-trivial. For a 1024^3 lattice-sites system, LB3D requires 1.07 TB of total memory to run; writing checkpoint files requires O(TB) of disk space. Each visualization step requires emission of a 4.3 GB visualisation dataset which has to be transferred to and rendered by a high performance visualisation resource.

3. Technical Challenges

In order to simulate physically realistic rheological behaviour of multi-domain, defect containing, ternary amphiphilic mixtures we need to perform long-time simulations of large systems containing at least 1024^3 lattice sites. As mentioned in the previous section, the simulation checkpoint and visualisation data-sets can reach up to terabytes and require significant network bandwidth for

transfer to storage and visualisation resources. Such large-scale data transfers have been performed for these lattice-Boltzmann simulations over UKLight [9] and in other RealityGrid projects, e. g. [8]. The large amount of memory required to carry out these simulations is often not available to us on a single supercomputer. Here we discuss a new network-intensive meta-computing approach called geographically distributed domain decomposition or GD^3 [10] that can overcome this memory bottleneck. In this approach, a single MPI simulation is split across processors on geographically distributed supercomputers. Network provisionability, bandwidth, latency and reliability during the simulation run are all critical in the GD^3 approach. The grid-middleware that we use to launch cross-site GD^3 simulations is called MPICH-G2 [11] and its newer pre-release version called MPIg.

Our initial aim is to split a 1024^3 lattice-sites simulation across supercomputers, the obvious candidates for this being the US TeraGrid resources and HPCx in the UK. HPCx is connected via the UKLight optical network [12] and Starlight network to the TeraGrid optical backbone. From an application scientist's perspective there are many technical challenges that need to be overcome in order to efficiently run cross-site simulations. Firstly, the simulation code needs to achieve maximum overlap between computation and communication by taking advantage of MPI's asynchronous communication calls. Unlike the previous MPICH-G2 version, MPIg implements asynchronous communications and is well-suited to take advantage of latency hiding optimisations in the code. Also UDT communication protocol was proposed in future versions of the grid middleware instead of the currently used TCP protocol. This is estimated to improve cross-site performance by a factor of two [10]. Secondly, in order to be included in MPICH-G2/MPIg cross-site framework, the participating machines need to have externally addressable nodes. This poses a problem for relatively less grid-enabled machines like the Cray XT3 machine (Bigben) at the Pittsburgh Supercomputing Center and the new UK HEC resource, HECToR [13]. Thirdly, we face issues in using MPICH-G2 on the UK's HPCx machine due to the port-forwarding mechanism in place on that machine. However, within the Vortonics project at SuperComputing Conference 2005 [14], a trans-atlantic cross-site run, over UKLight on TeraGrid machines and the now decomissioned Newton machine at CSAR on the UK NGS [4], was performed by Boghosian *et al* using their lattice-Boltzmann simulation code which has a similar communication pattern as LB3D [10]. In situations where the memory requirements of the simulation are too large to fit onto a single supercomputer, their results provide support for the viability of GD^3 as compared to alternatives like swapping portions of the simulation to disk or worse, waiting for a bigger machine to become affordable. From a usability point-of-view, cross-site simulations depend critically on the availability of automatic mechanisms for the advanced reservation and co-scheduling of compute resources and networks. To this end, tools have been developed for automated reservation and co-scheduling of grid resources [15] and of dynamically provisioned networks (within the ESLEA project). These tools used in conjunction with the grid application hosting middleware like the Application Hosting Environment [16] can allow scientists to efficiently and frequently schedule and launch cross-site simulations. Although there is a significant effort on middleware development, the support from grid resource managers on this front is less forthcoming. The UK NGS and EU DEISA [17] grid resource providers have not shown any serious indication of providing such a facility. There is a mechanism in place in the US TeraGrid to request for advanced reservation of resources through a web page, however, this requires manual intervention on the application scientist's part and as far

as we understand also on the resource manager's part. The beta version of the NAREGI [18] grid software stack, however, has a super-scheduler component to support co-scheduling and advanced reservation of grid resources. This is an encouraging development and further efforts like this are required to allow scientists to exploit advances in various areas of computing in a coherent fashion and be able to do science that was not possible before.

4. Summary

In this paper, we have described the scientific motivation for our lattice-Boltzmann simulations and the primary simulation issues that need to be overcome. We describe the main features of the LB3D code which determine network requirements. Finally we discuss a new meta-computing approach called geographically distributed domain decomposition (GD^3) for which high bandwidth, low latency, reliable network connections are critical and the technical challenges in deploying these simulations on transatlantic grids connected via UKLight.

This research has been funded by ESLEA project's EPSRC grant GR/T04465/01 and the EPSRC grants GR/R67699, EP/C536452/1, EP/E045111/1, GR/T27488/01 and through the OMII Managed Programme grant GR/290843/01. Access to the US TeraGrid resources was provided under the NRAC and PACS grants MCA04N014 and ASC030006P. We would also like to acknowledge useful discussions with Giovanni Giupponi, Marco Mazzeo and Steven Manos, and Nicola Pezzi's help with network issues.

References

[1] J. Harting, J. Chin, M. Venturoli, and P. V. Coveney. *Phil. Trans. R. Soc. A.*, 1833:1895–1915, 2005.

[2] J. Chin and P. V. Coveney. *Proc. R. Soc. London Series A,*, 462:3575–3600, 2006.

[3] G. Giupponi, J. Harting, and P. V. Coveney. *Europhys. Lett.*, 73:533–539, 2006.

[4] http://www.ngs.ac.uk.

[5] http://www.hpcx.ac.uk.

[6] http://www.teragrid.org.

[7] W. Gropp, E. Lusk, and A. Skjellum. In *Using MPI*, pages 59–97. MIT Press, 1994.

[8] M-A. Thyveetil, S. Manos, J. L. Suter and P. V. Coveney *in Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project, POS (ESLEA), 013*, 2007.

[9] M. Venturoli, M. J. Harvey, G. Giupponi, P. V. Coveney, R. L. Pinning, A. R. Porter, and S. M. Pickles. *Proceedings of the UK e-Science All Hands Meeting 2005*, 2005.

[10] B. Boghosian, L. I. Finn, and P. V. Coveney. *http://www.realitygrid.org/publications/GD3.pdf*.

[11] N. Karonis, B. Toonen, and I. Foster. *J. Parallel and Distributed Computing*, 63(5):551–563, 2003.

[12] http://www.uklight.ac.uk.

[13] http://www.hector.ac.uk.

[14] http://hilbert.math.tufts.edu/ bruceb/VORTONICS/index.html.

[15] J. Maclaren and M. Mc Keown. HARC: A Highly-Available Robust Co-scheduler, 2006. Proceedings of the 5th UK e-Science All Hands Meeting.

[16] P. V. Coveney, R. S. Saksena, S. J. Zasada, M. McKeown, and S. Pickles. *Comp. Phys. Comm.*, 176:406–418, 2007.

[17] http://www.deisa.org.

[18] http://www.naregi.org.