Real science at the petascale

BY RADHIKA S. SAKSENA, BRUCE BOGHOSIAN, LUIS FAZENDEIRO, OWAIN A. KENWAY, STEVEN MANOS, MARCO D. MAZZEO, S. KASHIF SADIQ, JAMES L. SUTER, DAVID WRIGHT, AND PETER V. COVENEY

1 Centre for Computational Science, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK
2 Department of Mathematics, Tufts University, Medford, MA 02155, USA

We describe computational science research that uses petascale resources to achieve scientific results at unprecedented scales and resolution. The applications span a wide range of domains, from investigation of fundamental problems in turbulence through computational materials science research to biomedical applications at the forefront of HIV/AIDS research and cerebrovascular haemodynamics. This work was mainly performed on the US TeraGrid ‘petascale’ resource, Ranger, at Texas Advanced Computing Center, in the first half of 2008 when it was the largest computing system in the world available for open scientific research. We have sought to use this petascale supercomputer optimally across application domains and scales, exploiting the excellent parallel scaling performance found on up to at least 32 768 cores for certain of our codes in the so-called ‘capability computing’ category as well as high-throughput intermediate-scale jobs for ensemble simulations in the 32–512 core range. Furthermore, this activity provides evidence that conventional parallel programming with MPI should be successful at the petascale in the short to medium term. We also report on the parallel performance of some of our codes on up to 65 636 cores on the IBM Blue Gene/P system at the Argonne Leadership Computing Facility, which has recently been named the fastest supercomputer in the world for open science.

Keywords: petascale computing; parallel and distributed computing; lattice–Boltzmann methods; molecular dynamics; material simulation; computational biomedicine

1. Introduction

Ranger (http://www.tacc.utexas.edu/resources/hpcsystems/) comprises 15 744 AMD ‘Barcelona’ quad-core processors housed in 3936 nodes, with a total memory of 123 TB. It has a total shared disk space of 1.73 petabytes (PB) and a local disk of 31.4 TB with global, high-speed Lustre file system, running across 72 I/O servers. This massive I/O capacity is critical for our computational

* Author for correspondence (p.v.coveney@ucl.ac.uk).

One contribution of 16 to a Theme Issue ‘Crossing boundaries: computational science, e-Science and global e-Infrastructure I. Selected papers from the UK e-Science All Hands Meeting 2008’.

This journal is © 2009 The Royal Society
endeavours. Ranger is connected to the US TeraGrid (http://www.teragrid.org/userinfo/hardware/resources.php) backbone via four nodes with 10 Gb Ethernet adapters. It has a theoretical peak performance of 0.58 petaflops (Pflops) and, according to recent benchmarks (TOP500, http://www.top500.org/list/2008/06/100), it delivers a Linpack speed of 0.31 Pflops. More recently, we have begun to use the IBM Blue Gene/P system called Intrepid at the Argonne Leadership Computing Facility (ALCF; http://www.alcf.anl.gov/resources/storage.php), Argonne National Laboratory (ANL), and have found excellent scaling on up to 65 536 cores. Intrepid (http://www.mcs.anl.gov/uploads/cels/files/news/2008/ALCF%20Science%20Brochure.pdf) delivers a theoretical peak of 0.56 Pflops and a Linpack speed of 0.45 Pflops, making it the fastest computer available for open science research. Intrepid consists of 163 840 cores and 80 TB of RAM. It has a total storage capacity of over 6 PB with an aggregate transfer speed of $80 \text{ GB s}^{-1}$. The storage is managed by two parallel file systems—PVFS and GPFS. Networking requirements for our work are met by a dedicated 1 Gb JANET (http://www.ja.net/services/lightpath/index.html) lightpath link (Coveney et al. in press) between the Starlight switch/router facility in the USA and our local storage and visualization resources at University College London. In the following sections, we will describe each of our applications that have been ported to these petascale machines and the scientific research that has been made possible. We first describe the lattice–Boltzmann (LB) code HYPO4D that has been ported to both these petascale resources. Next we describe our work on large-scale simulations of liquid crystals performed using an LB code called LB3D. We then discuss our large-scale molecular dynamics (MD) simulations of clay–polymer nanocomposite systems using LAMMPS (http://www.cs.sandia.gov/~sjplimp/lammps.html). Finally, we present two biomedical applications that use the MD code NAMD and the LB code HemeLB, respectively. The applications, which all ran concurrently, have been enabled by the petascale resources mentioned here and involve complex workflows and high-throughput computing.

2. Identification of unstable periodic orbits in the Navier–Stokes equations

The unstable periodic orbits (UPOs) of a dynamical system provide us with a fertile representation of the chaotic dynamics on the attractor of that system and allow for the computation of time averages in terms of an exact expansion, based on the dynamical zeta function (Artuso et al. 1990; Cvitanović et al. 2004). A small number of these UPOs should furnish accurate predictions of turbulent flow. The averages obtained in this way are not stochastic in nature; also there is no need to solve the initial-value problem each time we wish to compute a given observable. One of the more relevant applications of this method lies in turbulent hydrodynamics, described by the Navier–Stokes equations (NSE; Temam 1984), and hailed as one of the great unsolved problems of classical physics.

The approach we adopt in this work (Boghosian et al. 2008) is one that efficiently parallelizes time and space, by locating candidate space–time orbits and then applying a fully four-dimensional numerical relaxation algorithm. The algorithm has been successfully tested on low-dimensional dynamical systems.
such as the Lorenz equations, with favourable results (Boghosian et al. 2008). For the case of the three-dimensional NSE, this approach puts the computational requirements into the petascale as we discuss here.

Within our research, we have developed a scalable LB (Succi 2001) code called HYPO4D, an application to locate and characterize UPOs in the NSE. The code is written in C and parallelized using the Message Passing Interface (MPI), according to a spatial domain decomposition strategy. Asynchronous communications are used to prevent deadlock. We consider periodic boundary conditions so as to exclude wall effects and have an isotropic system. HYPO4D has been tested and deployed on several UK National Grid Service (NGS; http://www.ngs.ac.uk) sites, on HPCx (http://www.hpcx.ac.uk) and on TeraGrid resources with special emphasis on Ranger. Figure 1a shows the linear scaling of HYPO4D on Ranger, with the total system size being simulated kept fixed (‘strong scaling’) as well as with the portion of the system per core kept fixed (‘weak scaling’). Figure 1b shows weak scaling benchmarks, performed on the Blue Gene/P machine at ALCF.

In the simulations performed so far, the fluid starts from a condition of constant density and zero velocity, and an ABC-type force (Dombre et al. 1986) is then applied at each time step to drive the viscous flow. After having identified some relevant and physically valid sets of parameters on the NGS, we performed simulations of millions of LB time steps on Ranger, and then compared many of these in order to identify minima that could indicate the proximity to a periodic space–time trajectory. One of these minima can be seen in figure 2 for a cubic lattice with dimension \( L = 64 \) and a Reynolds number of 500, which corresponds to weakly turbulent behaviour. The quantity \( T \) gives the distance between the two time steps being compared, while \( t \) indicates the value of the time step in LB units. The quantity being plotted is

\[
\Delta(t, T) \equiv \sqrt{\sum_r \sum_{i=1}^l (f_i(r, t + T) - f_i(r, t))^2},
\]  

(2.1)

\[\text{Phil. Trans. R. Soc. A (2009)}\]
where $f_i(r,t)$ are the $l$ distribution functions at each site $r$ of the lattice at time step $t$ and the sum is over the whole lattice. HYPO4D uses the D3Q19 LB model, with $l=19$. Larger lattices were also simulated on Ranger, including cubic lattices of dimensions $L=128$ and 512. For the latter, a total of 3.4 million LB time steps were simulated with checkpoints being kept at every $10^4$ steps, for post-processing. This took close to half a million CPU-hours to run, and produced approximately 10 TB of data.

In this novel approach to turbulence, memory is the main bottleneck. For the minimum shown in figure 2, the relaxation procedure will require approximately 1.2 TB for a dataset consisting of $3 \times 10^4$ time steps. On increasing the system size to simulate flows with higher Reynolds numbers, we easily reach the level of petascale resource requirements, memory-wise. The results obtained so far show that the best candidates for even the smaller UPOs will consist of at least several tens of thousands of time steps. We intend to simulate flows as close to the fully developed turbulent regime as possible. Preliminary calculations show that we can aim to identify UPOs in systems with Reynolds numbers as high as $10^4$, with the memory requirement for a single candidate orbit being approximately 0.5 PB. By identifying and cataloguing these UPOs, we will then be able to systematically compute observables, in a non-stochastic manner, which could have important consequences not only in turbulence studies, but also in other areas involving highly nonlinear dynamical systems.

### 3. Large-scale LB simulations of liquid crystalline rheology

The non-equilibrium behaviour of amphiphilic liquid crystals self-assembled into various periodic phases, such as the lamellar, hexagonal and the cubic diamond, gyroid and primitive phases, is commonly observed in experiments.
These amphiphilic liquid crystalline phases are of great technological and scientific interest in materials science, in the synthesis of mesoporous catalytic templates, in biosensing, ultrafiltration and protein crystallization, and as possible intermediates in various important cellular processes (Schwarz & Gompper 2002; Squires et al. 2002). The non-equilibrium dynamics of these liquid crystalline phases can provide insight into the pathways and kinetics of self-assembly in non-cubic $\leftrightarrow$ cubic as well as cubic $\leftrightarrow$ cubic phase transitions that are also being actively studied experimentally (Seddon et al. 2006). Shear flow fields as well as other fields such as roll casting and extensional flow are used in the processing of functional materials for switchable optical, transport and electronic properties as well as ordered nanostructures (Chen et al. 1997). To design functional materials, one must understand the physical basis of the effects of molecular architecture and processing conditions (especially shear frequency, strain and temperature) on the rheological response of the liquid crystalline mesophases (Berghausen et al. 1998; Rey & Denn 2002).

The distribution of defects within a liquid crystal domain, defined as the ‘texture’ of the liquid crystal, plays a significant role in its rheology (Rey & Denn 2002). Kinetic LB simulations on smaller $64^3$ lattice sites lead to self-assembled monodomains of perfect gyroid, diamond and primitive mesophases (Saksena & Coveney 2008a), while real-world liquid crystalline mesophases usually do not exist as monocrysstals but rather as multiple oriented grains with grain boundary defects as well as dislocations and defects within the grains. Investigation of the rheology of these defect-containing multi-domain systems requires simulation of larger lattices, ideally comprising $256^3$ lattice sites or more. Routine production simulation of such large lattices has recently become possible with the arrival of the Texas Advanced Computing Center (TACC) Ranger machine. The simulation of ternary amphiphilic mixtures composed of oil, water and an amphiphilic species is performed using a kinetic LB implementation in the highly scalable, parallel MPI code LB3D (Love et al. 2003). The code has built-in computational steering capabilities, has been hosted in the Application Hosting Environment (AHE; Coveney et al. 2007) and has been deployed on various grid resources in the USA, UK and EU. Here, we report on the close-to-linear parallel scaling behaviour of LB3D on Ranger and on the new UK Cray XT4 machine, HECToR (http://www.hector.ac.uk/), in figure 3.

The simulations of the rheological yield curve of a $256^3$ defect-containing multi-domain gyroid system on Ranger have given us insight into the defect dynamics under shear flow for the first time (Saksena & Coveney 2008b). We found that regions of high stress magnitude arising from the applied strain are localized in the vicinity of defects (figure 4) and we are currently studying the temporal evolution of the defects and the associated stress field. Additionally, we have observed non-Newtonian shear-thinning behaviour (also seen in smaller $64^3$ and $128^3$ lattice systems) as applied strain rate was increased (Giupponi et al. 2006). The strain velocity profile induced by the Lees–Edwards boundary conditions takes approximately seven times longer to be established, i.e. 70 000 LB time steps from the time shear is first applied, as compared with the smaller $128^3$ system. This increase in simulation time required to achieve steady state in shear flow simulations with increasing system size further contributed to making such large-scale rheological simulations computationally unfeasible before Ranger became available. Future work in this area will focus on following temporal defect dynamics that
involves extremely disk-intensive analysis; scaling of disk I/O will become a major factor in determining code performance as we run simulations on higher and higher core counts. We plan to extend this work to other liquid crystalline phases such as the primitive and diamond phases (Saksena & Coveney 2008a).

4. Material properties of clay–polymer nanocomposites

In the materials science domain, we are using large-scale MD simulation to understand the unique chemical and physical aspects associated with polymer-based nanocomposite materials composed of micrometre-sized clay platelets...
dispersed in a polymer matrix (Pinnavaiam & Beall 2000). These nanocomposites have similar, or improved, mechanical, thermal, barrier and electrical properties to conventional composites, but at considerably lower filler fractions. A fundamental understanding of their properties will lead to widespread usage as with traditional, conventional composite materials (Chen et al. 2008). It is difficult to access the material structures of the clay–polymer nanocomposites experimentally, owing to their small grain size, while at the same time it has proved problematic to predict their materials properties theoretically owing to the great disparity of length scales inherent in these systems. Hence, valuable insights into the structure, behaviour and properties of clay–polymer nanocomposites are provided by atomistic simulations that are large enough to sample all the length scales required to describe the system accurately.

In order to fully exploit petascale resources such as Ranger, we use the open-source, MPI MD code LAMMPS. We find almost linear scaling of LAMMPS on up to 4000 processors with a system approaching 100 million atoms (figure 5). The deviation from linearity depends on the system size, as communication begins to dominate over processing on individual cores.

We have used LAMMPS to investigate the effects of montmorillonite clay filler on poly(ethylene) glycol in clay–polymer nanocomposites (Suter & Coveney in press). The sizes of these simulations were approximately 3 million atoms, two to three orders of magnitude larger than any previous atomistic clay mineral study. This was facilitated through a global infrastructure of federated grid-based computing resources (the US TeraGrid, the UK NGS and the EU DEISA (http://www.deisa.eu) grids) accessed through our lightweight hosting environment, the AHE (Coveney et al. 2007). We find that the conformation of polymer adsorbed on the surface of the clay is distinct from that of polymer far from the surface, owing to complexation with the counterions found near the charged clay surface. The diffusion of the adsorbed polymer is much smaller, which effectively increases the size of the nanofiller and renders gas permeation more tortuous. We also performed non-equilibrium MD simulations by imposing a strain on the model and analysing the stress response. By partitioning the stress response into clay and different polymer layers, each a certain distance from the clay surface, we find that the adsorbed polymer layer is much stiffer than the other layers of polymer molecules in the system. Our next step is the simulation of physically realistic-sized atomistic models, comprising 100 million atoms, which explicitly include the edges of the clay for the first time, effectively reproducing the experimental system. These simulations will run optimally on 4096 cores as per the scaling performance on Ranger (figure 5). In order to simulate polymer penetration into the clay platelet, which occurs on a time scale outside that possible with a single MD simulation, we will use a parallel tempering technique (replica exchange), which involves simulating multiple ensembles of the system, at different temperatures, on multiple partitions of cores. These replicas are then exchanged according to a Monte Carlo procedure. Since the exchanges occur rarely, parallel tempering can be regarded as an embarrassingly parallel problem. For such a scenario, the entire study will comprise at least 16 replicas of the clay–polymer nanocomposite system each running on 4096 cores for an estimated wall-clock time of two weeks. This unprecedented scale is therefore eminently possible with petascale resources.

*Phil. Trans. R. Soc. A* (2009)
An emerging concept in the biomedical domain is that of the use of high-end computational power to confer clinical decision support for the treatment of a range of illnesses (Sadiq et al. 2008b; Sloot et al. 2008). Applied to HIV therapy, a new approach that uses this paradigm aims to compute the binding affinity of any given inhibitor to any sequence of the corresponding viral protein target with sufficient accuracy to discriminate resistant mutations on a patient-specific basis and with enough rapidity to return on clinically relevant time scales. Therefore, provided the method fulfils such requirements and that suitable computational power is available, the binding affinity of a large array of protein–ligand combinations can be calculated for any given patient and in turn used to optimize treatment for that individual.

We have previously investigated the applicability of binding free-energy methods in fully atomistic molecular simulations as a way of ranking drug-resistant mutations of the inhibitor saquinavir (SQV; Stoica et al. 2008). These used the molecular mechanics Poisson–Boltzmann surface area (MMPBSA; Wang & Kollman 2001; Wang et al. 2001) with the inclusion of configurational entropies from normal mode analysis (Brooks et al. 1995) and gave excellent correlation to experimental values (Maschera et al. 1996; Ermolieff et al. 1997). We have also developed a tool, the binding affinity calculator (BAC), for the automated construction of MD simulations of HIV-1 protease–ligand complexes and calculation of binding free energies (Sadiq et al. 2008a), which further generalizes the protocol to any array of mutations and available ligands.

Here, we used BAC to perform an unprecedented number of fully atomistic simulations of an array of HIV-1 protease variants (wild-type and mutants) with six FDA inhibitors as well as seven cleavage-site natural peptide substrates using both single- and three-trajectory MMPBSA approaches up to a production time of 50 ns each. Peptide substrates were the MA-CA, CA-p2, p2-NC, NC-p1, p1-p6, RT-RH and RH-IN with amino acid sequences SQNY-PIVQ, ARVL-AEAM, ATIM-MQRG, QANF-LGKI, PGNF-LQSR, AETF-YVDG and
RKIL-FLDG, respectively (Prabu-Jeyabalan et al. 2002). The G48V, L90M and G48V/L90M mutant proteases bound to the NC-p1 and CA-p2 substrates and the V82A, A431V and V82A/A431V mutations in complex with NC-p1 were also simulated. Inhibitor systems simulated were amprenavir (APV), indinavir (IDV), lopinavir (LPV), ritonavir (RTV) and SQV with the HXB2 wild-type protease sequence (WT) and five MDR mutants, including the hexa-mutant L10I,M46I,I54V,V82A,I84V,L90M (denoted HM) as well as comprising subsets of HM, namely the M46I,I54V,V82A,I84V quatre-mutant (denoted QM) and pairwise mutants L10I,L90M (denoted DM), M46I,I54V (denoted FL) and V82A,I84V (denoted AS). The equilibration period for each run was always 2 ns, followed by a production run for each system up to 10 ns. LPV and RTV sets of systems were extended to 50 ns and the separate three-trajectory MMPBSA method was additionally applied to the LPV system, each up to 50 ns production.

Multiple and/or ensemble MD approaches, employed by projects such as Folding@home (Shirts & Pande 2000; Fujitani et al. 2005; Jayachandran et al. 2006), have been shown to explore conformational space more effectively than single long trajectories. Such approaches, in which multiple replicas of the same system are simulated with slightly varying initial conditions, are particularly suited to petascale resources owing to the overwhelming speed-up in turnaround they can afford over performing single long trajectory simulations of equivalent total simulation time. We simulated all LPV systems using the ensemble MMPBSA method, in which 50 replicas of each LPV-bound system were simulated each to 4 ns production.

The details of the underlying simulation protocol and free-energy calculation method are provided in Sadiq et al. (2008a) and Stoica et al. (2008). Over 3.8 µs of simulation time was achieved within two weeks of elapsed wall-clock time via distribution across a maximum of 150 simultaneously run systems, each system comprising approximately 40 000 atoms and using 32 CPUs. For the first time, we were also able to expand the methodology to the reverse transcriptase system (an approx. 160 000 atom system). The well-established MD code NAMD was used for the production runs, giving a benchmark performance of approximately 8 h ns⁻¹ under the above conditions. Minimum single-trajectory turnaround time (TAT) using 192 cores (at 18 ns d⁻¹) was 416 hours; theoretical minimum ensemble trajectory TAT using 9600 cores (simulating 300 replicas simultaneously at 900 ns d⁻¹) was 48 hours. In practice, we achieved on average 120 ns d⁻¹ with a peak of 300 ns d⁻¹. The theoretical minimum total TAT (simulation + calculation) using this approach was approximately 3 days. In total, over 3.5 TB of trajectory and free-energy analysis data were generated. Importantly, the decoupled nature of each replica means that, using the entire core count of Ranger, approximately 6 µs of MD production simulation data can be generated per day. The protocol used here is thus scalable to any number of cores, being able to turn around approximately 2000 replicas (40 HIV-1 protease–ligand binding affinities) within 3 days on Ranger.

Figure 6 shows the relative ranking of the binding free energy of LPV and RTV studied when bound to wild-type and MDR proteases over a sampling time scale of up to 50 ns. The LPV and RTV systems exhibited excellent peak correlation coefficients (κ) of 0.84 and 0.93, respectively. The results are encouraging and they build on our earlier work (Stoica et al. 2008).
6. Whole-brain patient-specific haemodynamics

Understanding cerebral blood flow behaviour is crucial in the diagnosis and treatment of cardiovascular disease. Blood flow simulation in combination with patient-specific medical imaging data can help not only to understand blood flow patterns in the presence of neurovascular pathologies, but also to provide a tool that surgeons can use in a diagnostic and therapeutic capacity for patients according to their unique vascular structure. HemeLB is an application specifically designed to address fluid flow in sparse geometries, which simulates pulsatile blood flow in the brain (Mazzeo & Coveney 2008). Performance benchmarks show that the core LB code of HemeLB scales superlinearly (figure 7a) on some platforms. The HemeLB algorithm uses a data layout that automatically handles lattices with large unstructured regions covered by non-fluid nodes. A single-pass algorithm is used to spatially decompose the vascular tree into partitions of equal size—resulting in balanced communication between MPI ranks. Partial overlapping of communication between subdomains and on-core computation is also exploited to enhance performance, and the computational kernel is highly optimized. Indeed, the parallel performance of this code outstrips any comparable LB codes that we are aware of (Mazzeo & Coveney 2008). HemeLB has been used in combination with MPIg, the latest

Figure 6. Computed binding free-energy differences compared with experimental results for wild-type and all MDR proteases with inhibitors (a) LPV (black circles, experiment; red squares, 1–10 ns, κ=0.81; green diamonds, 1–20 ns, κ=0.72; blue up triangles, 1–50 ns, κ=0.84; purple left triangles, 31–50 ns, κ=0.79; orange down triangles, 41–50 ns, κ=0.84) and (b) RTV (black circles, experiment; red squares, 1–10 ns, κ=0.93; green diamonds, 1–20 ns, κ=0.86; blue up triangles, 1–50 ns, κ=0.59; purple left triangles, 31–50 ns, κ=0.40; orange down triangles, 41–50 ns, κ=0.46) using 50 ns single-trajectory runs. For comparison, the default experimental correlation is also shown. Correlation coefficients, κ, between ΔG_{theor} and ΔG_{exp} for each set are as listed.
pre-release of the successor to MPICH-G2 (Karonis et al. 2003) on the TeraGrid, Louisiana Optical Network Initiative (LONI; USA) and NGS2 (UK) networks, achieving near-linear scalability across different computational sites (figure 7).

In order to avoid post-processing of data, we employ in situ visualization, an optimal strategy for large-scale visualization where rendering is done within the same cores that perform the simulation in a unified and tightly coupled approach. Within HemeLB, each subdomain is rendered independently, corresponding to the LB domains. Owing to the requirement of global MPI communications to compose the final image, the rendering aspect of HemeLB at each time step is not as efficient as the LB code alone (figure 7). Computational steering is used to control and monitor physical parameters of the simulation.

Figure 7. (a) Scalability of the LB fluid solver (FS), along with the FS in combination with real-time volume rendering (FS+VR) of HemeLB on the HPCx Power5 cluster, ANL IA64 cluster ‘Abe’ and ANL IBM Blue Gene/P cluster. The system tested is a three-dimensional arterial neurovasculature with $7.7 \times 10^6$ fluid sites (filled circles, HPCx and FS; open circles, HPCx and FS+VR; crosses, Abe and FS; filled squares, Abe and FS+VR; asterisks, Blue Gene/P and FS; open squares, Blue Gene/P and FS+VR). (b) Scalability on the LONI network using MPIg with 32 cores per site (dashed curve with circles, FS; dashed curve, ideal and FS; dashed curve with filled squares, FS+VR; dashed curve with open squares, FS+IS). The sites refer to three geographically separate IBM P5 clusters. FS+VR does not scale as well due to the global communications that are needed for rendered frame compositions. However, in reality, VR has little impact on the overall speed since it is only conducted at a rate of 25 times s$^{-1}$.

Figure 8. Real-time interactive rendering at $1024^2$ resolution of patient-specific cerebral haemodynamics. (a) Volume rendering of the blood velocity, (b) volume rendering of the internal blood von Mises stress, (c) pressure against the arterial wall and (d) stress against the arterial wall.
along with visualization parameters (viewpoint, colour maps, etc.), in real time. The code is structured in such a way that the performance of HemeLB is decoupled from the network conditions, such that scenes are rendered only as frames are sent. Thus, the impact of rendering on the overall speed of the simulation is minimal. Depending on the performance of HemeLB on the particular machine as well as the network performance, rendering takes place only every 10–200 LB time steps. As a result, we can volume render in real time, at $1024^2$ pixel resolution and 25 frames s$^{-1}$ while the LB simulation itself runs at near full speed, between our local desktop visualization client (figure 8) and globally distributed remote resources.

The software environment developed for use with HemeLB aims to hide unnecessary grid-computing details from clinicians, while bringing to the forefront the details and processes that they need to be aware of. This is particularly important given the time scales involved in the clinical decision-making process in interventional neuroradiology. During a surgical procedure, three-dimensional rotational angiography is used to rapidly create a volume highlighting the vasculature ($600^3$ voxels, 440 $\mu$m resolution). Flow data are needed within 15–20 min of this acquisition. Within this time, data are segmented into the neurovasculature and boundary conditions are applied, the patient-specific data are staged using the AHE and the job is then executed within a reservation, as a cross-site or single-site run. Finally, as the job commences running, the client connects to the running simulation, which is used for steering and visualization purposes. The availability of an urgent computing system, or pre-reserved slots on remote petascale resources, is imperative. The scalability properties of HemeLB also significantly contribute to the rapid turnaround required. For example, using 1024 cores on the NCSA (http://teragrid.ncsa.uiuc.edu) Abe cluster, a single pulsatile cycle of an entire arterial neurovasculature ($1024^3$ voxels, 0.245 mm resolution, $10^6$ fluid sites) takes approximately 9 min, well within the clinical requirement.

To the best of our knowledge, this is the first method developed with the ability to simulate whole-brain haemodynamics. In collaboration with neuroradiologists at the National Hospital for Neurology and Neurosurgery, London, UK, we are currently carrying out simulations on real medical cases. The longer-term goal is to determine the direct clinical utility of not only cerebral haemodynamics, but also real-time patient-specific medical-simulation methods in general, where validation, verification and clinical studies will be required.

7. Conclusion

In this paper, we describe five novel scientific applications that have been enabled by the availability of petascale resources. We have observed excellent scalability of our codes on up to tens of thousands of cores without any specific optimizations for individual machine architectures. This computational power has allowed us to perform unprecedented simulations in order to investigate complex phenomena over a range of applications, concurrently. One of the major operational issues associated with running such ‘capability’ jobs on large core-count machines is the low mean time to failure compared with the simulation run-time. Future growth in petascale resource usage within our work
will involve integrating fault tolerance (FT) into scientific application codes, for example by using fault-tolerant MPI (FT-MPI; Chen et al. 2005), so that they can survive multiple bursts of process failures with low overhead.

We are grateful for funding from EPSRC grants EP/C536452/1, EP/E045111/1 and EP/F00561X/1, the Virolab project (EU IST STREP Project 027446) and the US NSF MRAC grants TG-DMR070013N and TG-DMR070014N. We acknowledge a recently awarded DOE INCITE grant entitled ‘Large scale condensed matter and fluid dynamics simulations’ at ALCF, which is supported by the US Department of Energy under contract DE-AC02-06CH11357. We are grateful to John Boisseau and the TACC support team for facilitating access to Ranger and to Peter Beckmann and Ramesh Balakrishnan for facilitating access to the Surveyor and Intrepid systems at ALCF.

References


Maschera, B., Darby, G., Paliú, G., Wright, L. L., Tisdale, M., Myers, R., Blair, E. D. & Furline, E. S. 1996 Human immunodeficiency virus. Mutations in the viral protease that confer resistance to saquinavir increase the dissociation rate constant of the protease–saquinavir complex. *J. Biol. Chem.* 271, 33 231–33 235. (doi:10.1074/jbc.271.52.33231)


*Phil. Trans. R. Soc. A* (2009)

