Computer Physics Communications 180 (2009) 1495–1502

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

MUPHY: A parallel MUlti PHYsics/scale code for high performance bio-fluidic simulations

M. Bernaschi [a,*], S. Melchionna [b,c], S. Succi [a,d], M. Fyta [c], E. Kaxiras [c], J.K. Sircar [c]

[a] Istituto Applicazioni Calcolo, CNR, Viale Manzoni, 30 – 00185 Rome, Italy
[b] SOFT, Istituto Nazionale Fisica della Materia, CNR, P.le A. Moro, 2 – 00185 Rome, Italy
[c] Department of Physics and School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[d] Initiative in Innovative Computing, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 22 January 2009
Received in revised form 7 March 2009
Accepted 2 April 2009
Available online 5 April 2009

Keywords:
Lattice Boltzmann
Molecular dynamics
Multi-physics
Biopolymer translocation
Parallel processing

ABSTRACT

We present a parallel version of MUPHY, a multi-physics/scale code based upon the combination of microscopic Molecular Dynamics (MD) with a hydro-kinetic Lattice Boltzmann (LB) method. The features of the parallel version of MUPHY are hereby demonstrated for the case of translocation of biopolymers through nanometer-sized, multi-pore configurations, taking into explicit account the hydrodynamic interactions of the translocating molecules with the surrounding fluid. The parallel implementation exhibits excellent scalability on the IBM BlueGene platform and includes techniques which may improve the flexibility and efficiency of other complex multi-physics parallel applications, such as hemodynamics, targeted-drug delivery and others.

1. Introduction

A wide variety of physical systems exhibit complex behavior which involves several spatial and temporal scales. Modeling this behavior requires the combination of several types of physical theories, that can be implemented in multi-scale/multi-physics computational approaches. The need for such sophisticated computational models is driven both from an interest in fundamental understanding of complex systems behavior as well as from a practical perspective to improve technological and industrial applications. Indeed, many advanced-technology applications, from high efficiency turbines to transistors with nano-scale functional elements, rely heavily on multi-scale/multi-physics approaches for their design [1]. Much progress has been made during the last decade in formulating multi-scale approaches based on composite computational schemes in which cross-scale information is exchanged in either a sequential or, less frequently, a concurrent manner [2].

A particularly challenging class of complex phenomena that involve multiple scales and physics are biological processes. Biological systems often exhibit a complexity and diversity that straddles across many decades in space–time resolution, from the quantum mechanical level of interactions responsible for bonding at the atomic scale to the continuum level of fluid motion at macroscopic scales. We focus here on a representative example of such a system, the translocation of DNA through a nanopore with the ultimate goal of ultrafast electronic sequencing. In this example, as in many similar biological systems, the construction of an *efficient and accurate* multi-scale/multi-physics computational method is particularly challenging due to the diverse and competing requirements of the individual scales involved. We report the construction of such a computational method (MUPHY) and demonstrate its efficient implementation on a highly parallel architecture, the IBM BlueGene Supercomputer.

The approach we follow is genuinely multi-physics, as it combines different levels of the statistical description of matter, continuum fluids and individual molecules. It is also genuinely multi-scale, since fluid and molecular degrees of freedom are advanced concurrently, while invoking a sub-cycle time-stepping for these degrees of freedom to take into account the fact that atomistic dynamics is generally faster than the dynamics of the surrounding fluid (solvent). At variance with the vast majority of hybrid fluid/atomistic schemes, the present work is based on a *mesoscopic/hydro-kinetic* representation of the solvent, via the so-called Lattice Boltzmann (LB) method [3–5].

On general grounds, kinetic theory, which lies in-between the continuum and atomistic description of matter, provides a natural candidate for multi-scale fluid applications. In the case of Lattice Boltzmann, this general observation results into a series of specific computational assets, as briefly described in the following. LB is a minimal form of kinetic Boltzmann equation, based on

* Corresponding author.
E-mail address: m.bernaschi@iac.cnr.it (M. Bernaschi).

0010-4655/$ – see front matter
doi:10.1016/j.cpc.2009.04.001

the collective dynamics of fictitious particles, living on the nodes of a regular lattice, and representing a local ensemble of solvent molecules (Boltzmann distribution function). The dynamics of such particles is designed in such a way as to reproduce hydrodynamic behavior in the continuum limit, in which the molecular mean free path is much shorter than typical macroscopic scales.

Full details on our coupled LB/MD schemes are reported in [7]. It is worth noting that LB and MD with Langevin dynamics have been coupled before [6]. However, to the best of our knowledge, this is the first work in which such coupling is carried out using a flexible high performance parallel code able to support long molecules of biological interest, consisting of *tens of thousands* monomers. In addition, the indirect addressing of the main lattice data structure allows our LB code to handle efficiently nearly arbitrary geometries at a minimum extra-memory cost.

The paper is organized as follows: Section 2 reviews the simulation method used to couple the Lattice Boltzmann description for the solvent to the Molecular Dynamics description of the solute. Section 3 describes the issues faced in the development of the parallel code and the solutions we adopted. Section 4 presents the numerical demonstration for the case of biopolymer translocation across a multi-hole membrane. Finally, a brief discussion of the future perspectives of this activity is offered.

2. Coupling between Lattice Boltzmann and Molecular Dynamics

In this section, we review the basic methodology, for which further details can be found in previous works [7,8].

In the LB method the basic quantity is $f_i(\vec{x},t)$, representing the probability of finding a "fluid particle" at the spatial mesh location $\vec{x}$ and at time t with discrete speed $\vec{c}_i$ and mesh spacing Δx. Actually, "fluid particles" do not correspond to individual solvent molecules, but they represent instead the collective motion of a group of physical particles (populations). For the present study, we use the common three-dimensional 19-speed lattice where the discrete velocities $\vec{c}_i$ connect mesh points to first and second topological neighbors [5]. Once the discrete populations f_i are known, the kinetic moments are obtained by a direct summation upon all discrete populations at a given lattice site, with

$$\rho(\vec{x},t) = \sum_i f_i(\vec{x},t) \tag{1}$$

being the local density,

$$\rho\vec{u}(\vec{x},t) = \sum_i f_i(\vec{x},t)\vec{c}_i \tag{2}$$

the flow current and

$$\overleftrightarrow{P}(\vec{x},t) = \sum_i f_i(\vec{x},t)\vec{c}_i\vec{c}_i \tag{3}$$

the momentum-flux tensor. The fluid populations are advanced in time through the following evolution equation:

$$\begin{aligned} &f_i(\vec{x}+\vec{c}_i\Delta t, t+\Delta t) \\ &\quad = f_i(\vec{x},t) + \omega\Delta t\big(f_i - f_i^{eq}\big)(\vec{x},t) + \Delta t F_i(\vec{x},t) + \Delta t S_i(\vec{x},t). \end{aligned} \tag{4}$$

The right-hand side of Eq. (4) represents the effect of fluid–fluid molecular collisions, through a relaxation towards a local equilibrium, f_i^{eq}, typically a second-order expansion in the fluid velocity of a local Maxwellian with speed $\vec{u}$:

$$f_i^{eq} = w_i\rho\left[1 + \frac{\vec{u}\cdot\vec{c}_i}{c^2} + \frac{\vec{u}\vec{u}:(\vec{c}_i\vec{c}_i - c^2\overleftrightarrow{I})}{2c^4}\right], \tag{5}$$

where $c = 1/\sqrt{3}$ is the sound speed, w_i a set of weights normalized to unity, and $\overleftrightarrow{I}$ is the unit tensor in Cartesian space. The relaxation frequency ω controls the kinematic viscosity of the fluid, $\nu = c^2\Delta t(\frac{1}{\omega} - \frac{1}{2})$.

The $F_i(\vec{x},t)$ term represents a stochastic force accounting for fluctuations in the fluid in the framework of fluctuating hydrodynamics. As thoroughly discussed in [9,10], this term is space/time local and acts at the level of the stress tensor and non-hydrodynamic modes. These are constructed via a set of 19 lattice eigenvectors $\{\chi_k\}_{k=0,18}$ orthonormal according to the scalar product $\sum_{i=0}^{18} w_i\chi_{ki}\chi_{li}$. The eigenvectors correspond to the kinetic moments: $k=0$ is relative to mass density, $k=1$–3 to mass current, $k=4$–9 to the (symmetric) momentum flux tensor, the remaining $k=10$–18 eigenvectors to non-hydrodynamic modes. The stochastic forcing reads

$$F_i = \sqrt{\frac{k_B T\rho\omega\Delta t(2-\omega\Delta t)}{c^2}}\sum_{k=4}^{18} w_i\chi_{ki}\mathcal{N}_k, \tag{6}$$

where k_B is the Boltzmann constant, T the temperature and $\mathcal{N}_k$ a set of 15 random numbers with zero mean and unit variance. This stochastic forcing has the merit of producing consistent fluctuations at all spatial scales, in particular at short distances where the effect on embedded molecules is critical.

The source term S_i accounts for the presence of particles embedded in the LB solvent and represents the momentum and momentum-flux input per unit time due to the influence of atomic-scale particles on the fluid. Whenever the LB method is used to simulate a plain fluid, with no thermal fluctuations, the F_i and S_i terms disappear from the fluid evolution equation.

By definition, the particle-fluid back-reaction does not change the fluid density, so that $\sum_i S_i = 0$. Momentum and momentum flux conservation of the solute and solvent systems imply that $\sum_i S_i\vec{c}_i = -(\vec{F}^f + \vec{F}^r)$. The source term then reads

$$S_i = -w_i\frac{(\vec{F}^f + \vec{F}^r)\cdot\vec{c}_i}{c^2}, \tag{7}$$

where $\vec{F}^f$ describes the mechanical friction between a particle and the surrounding fluid and $\vec{F}^r$ is a random force.

Before describing the MD part, we emphasize that a LB solver is particularly well suited to the study of a number of interesting problems for several reasons: first, free-streaming of the fluid proceeds along straight trajectories which greatly facilitates the imposition of geometrically complex boundary conditions, such as those required to describe membranes and nano-pores. Second, fluid diffusivity emerges from the first-order LB relaxation–propagation dynamics, so that the kinetic scheme can march in time-steps which scale only linearly with the mesh resolution. Third, since both fluid–fluid and fluid–particle collisions are completely local, the LB scheme is well suited to parallel computing. These features make the LB an appealing method as compared to other available alternatives, which typically scale superlinearly with the number of particles.

We now describe the MD section of the method, bearing in mind that the embedded solute has a molecular topology, such as DNA. In our case, this is a linear collection of N_0 beads (each bead or solute particle representing a collection of atoms or molecules) composing the polymer. Each solute particle has position $\vec{r}_p$ and velocity $\vec{v}_p$ and experiences the following cumulative force:

$$\vec{F}_p = \vec{F}_p^c + \vec{F}_p^f + \vec{F}_p^r, \quad p = 1, N_0. \tag{8}$$

Here, the conservative force is modeled as

$$\begin{aligned} \vec{F}_p^c = -\partial_{\vec{r}_p}\Bigg[&\sum_q V_{LJ}(|\vec{r}_p - \vec{r}_q|) + \frac{\kappa}{2}(|\vec{r}_{p+1} - \vec{r}_p| - r_0)^2 \\ &+ \frac{\kappa}{2}(|\vec{r}_{p-1} - \vec{r}_p| - r_0)^2\Bigg], \end{aligned} \tag{9}$$

where the bead–bead interaction potential $V_{LJ}(r)$ is a standard Lennard-Jones potential

$$V_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{10}$$

truncated to zero for $r > r_c = 2^{1/6}\sigma$. The remaining interaction arises from a harmonic bonding potential with force constant κ and equilibrium distance r_0, and is required to preserve the polymer connectivity.

The force $\vec{F}_p^f$ represents the mechanical friction acting between a single particle and the surrounding fluid, and is modeled as

$$\vec{F}_p^f = \gamma(\vec{u}_p - \vec{v}_p) \tag{11}$$

with $\vec{v}_p$ being the particle velocity, $\vec{u}_p \equiv \vec{u}(\vec{r}_p)$ the fluid velocity evaluated at the particle position, and γ a friction coefficient. In addition, the particles experience the effects of stochastic fluctuations of the fluid environment through the random term $\vec{F}_p^r$, obeying the fluctuation–dissipation relations.

Clearly, in the LB approach all quantities have to reside on the lattice nodes, which means that the frictional and random forces need to be extrapolated from the particle to the lattice location. This is obtained by extracting the fluid velocity field $\vec{u}_p$ at the nearest lattice point from each particle position and, similarly, by assigning these forces as feed-back on the fluid populations through the same recipe. This simple procedure is as accurate as a more involved bilinear interpolation/extrapolation scheme for the exchange of forces and momentum.

The temporal evolution of particles needs to take into account the stochastic nature of the forces. This is handled through the Stochastic Position Verlet (SPV) algorithm, as introduced in [11], a second-order accurate propagator. A central issue regards the numerical stability of the propagation that depends on the stiffness of the conservative and frictional forces. At low Reynolds regime, frictional forces are well-behaved and do not pose such a problem. Conversely, in order to preserve molecular connectivity at high accuracy, the bonding forces are necessarily stiff, introducing fast oscillations that threaten the numerical stability of the scheme at large time-steps (typically over the LB timestep Δt). A similar, but less severe, problem arises from the Lennard-Jones forces. To take into account such fast movements, a small integration timestep should be used, with a resulting penalty on the computational efficiency. Actually, a multiple timestep algorithm permits to achieve stability without compromising numerical efficiency. For the translocating polymer, the MD solver is marched in time with a fraction of the LB time-step, $dt = \Delta t/M$ (a typical value of the time-step ratio M is 2). A multiple timestep integrator is employed [11] by introducing a nested sub-cycle over a timestep $dt^b = dt/M^b$ as follows:

$$\begin{aligned}
&\vec{r}_p = \vec{r}_p + \frac{dt}{2}\vec{v}_p,\\
&\vec{v}_p = \vec{v}_p + \frac{dt}{2m}\vec{F}_p(\vec{r}),\\
&\left\{\vec{v}_p = e^{-\gamma dt^b}\vec{v}_p + \frac{1-e^{-\gamma dt^b}}{m\gamma}\left(\vec{F}_p^c + \gamma\vec{u}_p\right)\right.\\
&\qquad\left. + \sqrt{\frac{k_B T}{m}\left(1-e^{-2\gamma dt^b}\right)}\mathcal{N}_p\right\}_{M^b \text{ cycle}},\\
&\vec{v}_p = \vec{v}_p + \frac{dt}{2m}\vec{F}_p(\vec{r}),\\
&\vec{r}_p = \vec{r}_p + \frac{dt}{2}\vec{v}_p,
\end{aligned}$$

where, again, $\mathcal{N}_p$ is a random number with zero mean and unit variance. The multiple timestep solver is marched in time with time-step ratios $M = 2$ and $M^b = 5$, providing accurate results in terms of stability and unbiased statistical averages, as verified by monitoring the system average temperature, which remains equal to the preset value. To be noted that the LB/MD coupling in our approach is much tighter than for current Navier–Stokes/MD hybrid schemes, typically featuring an order-of magnitude larger separation, $M \sim 100$ [12,13].

3. Code parallelization

MUPHY (MUlti PHYsics/multi-scale computer code) is written in Fortran 90. We chose MPI as the communication interface for the parallelization since it offers high portability among different platforms and allows good performance due to the availability of highly tuned implementations. The code has been tested on different platforms (including clusters of PC) and in particular on the IBM BlueGene/L system [14], whose main features can be summarized as follows:

- dual processors 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 with superscalar 700 MHz PowerPC 440 cores;
- a large number of nodes (scalable up to at least 65,536);
- three-dimensional torus interconnect with auxiliary networks for global communications, I/O, and management;
- lightweight, Unix-like, OS per node for minimum system overhead.

The parallelization of the Lattice Boltzmann method and Molecular Dynamics algorithms, as two distinct and *separate* problems, has been extensively studied for a number of years [15–22]. However, the coupling of these techniques raises new issues that need to be solved in order to achieve scalability and efficiency for large scale simulations. We addressed these issues starting from a serial version of the combined code, instead of trying to combine two existing parallel versions. The original LB code was primarily updated to take advantage of optimizations like (i) removal of redundant operations; (ii) buffering of multiply-used operations [23] and (iii) "fusion" of the collision and streaming steps in a single loop. This latter technique, already in use in other high-performance LB codes [23], significantly reduces data traffic between main memory and cache/registers of the processor, since there is only one read and one store of all LB populations at each timestep. We store the velocity distribution functions of each lattice site continuously in memory so as to minimize, at least in the loading of the LB populations, the number of cache misses.

With these optimizations in place, we are able to achieve ∼30% of the peak performance of a single BlueGene core for the plain LB method (Eq. (4) with no F_i and S_i terms). This result is in line with other highly tuned LB kernels [23]. Indeed we wish to remind that: (i) the algorithm for the update of the LB populations has an unfavorable ratio between number of floating point operations and number of memory accesses; (ii) no optimized libraries are available like for other computational kernels (*e.g.*, matrix operations or FFTs); (iii) it is not possible to exploit the SIMD-like operations of the PowerPC 440 processor since they require stride one access whereas the LB method has a "scattered" data access pattern. For the generation of the random numbers required by the simulation of the stochastic fluctuations (the F_i term in Eq. (4)) we resorted to the very efficient library functions available in the IBM ESSL.

As to the parallelization, we followed an approach that entails a sort of "run-time" pre-processing. For the LB part of the code, this initial stage can be summarized as follows. Each node of the

LB lattice is labeled with a *tag* that identifies it as belonging to a specific subregion of the computational domain (*i.e.*, fluid, wall, inlet, outlet, solid), as read from an input file. It is possible to define a *default* value for the tag so that nodes that are not explicitly tagged are assumed to have that value for the tag (this assumption reduces significantly the size of the input file). The *tag* file is read by a single task that distributes the input data to all the other tasks by using collective communication primitives. Nodes are stored according to a linearized indirect addressing scheme [24,25]. Obviously this indirection requires, for each node, an additional data structure that contains the list of its neighboring (fluid, wall, inlet, outlet) nodes. At first glance a mechanism like this may appear a waste of time and memory, but, actually, for most (non-trivial) geometries it provides huge savings in memory requirements and simulation times [26]. In the beginning of the simulation, a subset of nodes is assigned to each computational task, attempting to balance the number of nodes *per* task as much as possible (obviously, in some cases, this operation cannot be done exactly). The assignment takes into account the domain decomposition strategy, that can be one-, two- or three-dimensional. All possible combinations (that is, Cartesian decompositions along X, Y, Z, XY, XZ, *etc.*) are supported. Moreover *custom* decompositions, *e.g.*, those produced by tools like METIS [27], which are necessary for irregular domains, are also supported. After the assignment of the nodes to the tasks, the *pre-processing* phase begins. Basically, each task checks which tasks own the nodes to be accessed during the subsequent phases of simulation, in particular for the streaming part of the LB algorithm. Such information is exchanged by using MPI collective communication primitives, so that each task knows the neighboring peers for send/receive operations. Information about the size of data to be sent/received is exchanged as well. In principle, each node could determine by itself all the information required for the communication phase, but the availability of highly tuned collective communication primitives makes more efficient to exchange part of these data among the tasks than computing everything locally. Only at run time it is possible to know exactly the number and the identity of the tasks with which it is necessary to exchange data since this information depends on the geometry of the domain, the total number of tasks, the decomposition method and the boundary conditions. For such reason, all data structures required for the communication are dynamically allocated. There are some obvious "bounds": a 3D Cartesian decomposition requires, at least, 8 tasks and in case of periodic boundary conditions, each task needs to interact with all other tasks.

In a parallel LB scheme there are several sections requiring data exchange: (i) streaming; (ii) handling of periodic boundary conditions; (iii) presence of reflecting or absorbing *walls* within the computational domain. In our code, both boundary conditions and *walls* are managed implicitly during the streaming phase by using indirect addressing. The communication scheme is based on the following approach: the *receive* operations are always posted in advance by using the corresponding non-blocking MPI primitives, then the *send* operations are carried out using either blocking or non-blocking primitives, depending on the parallel platform in use (unfortunately, as it is well know, few platforms allow real overlapping between communication and computation). Then, each task waits for the completion of its receive operations, by using the MPI *wait* primitives. The latter operation, in case of non-blocking *send* operations, is to wait for their completion. Also the choice between blocking and non-blocking *send* can be done at run time. All point-to-point send and receive operations involve only the LB populations strictly required by each task. To this purpose a buffering mechanism is used to *pack* (and unpack) data that need to be exchanged. The evaluation of global quantities (*e.g.*, the momentum along the X-, Y-, Z-directions) is carried out by using MPI collective *reduction* primitives. Table 1 shows the time required by MUPHY to run a small test case ($128 \times 128 \times 128$) using a 3D Cartesian decomposition. It is worth noting that efficiency remains pretty good even though, due to the size of the test case, each task has in charge, for the largest configuration (1024 tasks), only 2048 lattice sites.

Table 1
Times for 100 LB time-steps on a $128 \times 128 \times 128$ lattice using the IBM BlueGene/L system. 32 is the minimum number of tasks with the system configuration used for this test.

Number of tasks	Time (sec)	Number of nodes *per* task	Efficiency (%)
32	12.4	65536	
128	3.36	16384	92.4
512	0.86	4096	90.0
1024	0.44	2048	88.0

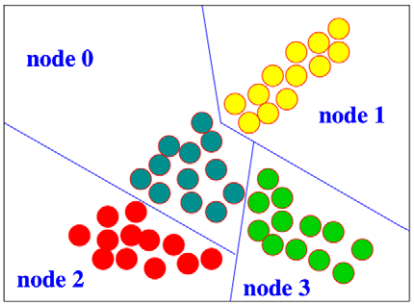

Fig. 1. A possible domain decomposition that achieves a good global MD balancing (each node has in charge exactly the same number of particles) in a simplified 2D case (the color of the beads is used only to identify those belonging to the same subdomain). It is apparent how the size of the spatial subdomains is different, so that the time required for the LB update, that is proportional to the domain size, varies substantially from node to node. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For the Molecular Dynamics section, a parallelization strategy suitable for the multi-scale method described in Section 2 had to be developed. In typical MD applications the spatial distribution of particles may be highly inhomogeneous. A straightforward approach to achieve a good global load balancing is to resort to a domain decomposition such that each task owns (approximately) the same number of particles. However, in this way, the size of the spatial sub-domains assigned to each task may vary significantly as shown in Fig. 1. In a stand-alone Molecular Dynamics simulation, this is acceptable, but in our case the LB component would be hindered, since the computational load is proportional to the size of the spatial domain assigned to each task. One might opt for two separate domain decompositions for the LB and the MD part of the simulation. However, the exchange of momentum among particles and the surrounding fluid would become a non-local operation, with a very high cost due to the long-range point-to-point communications imposed on the underlying hardware/software platform. For the IBM BlueGene such communications are explicitly discouraged.

In the end, we decided to resort to a domain decomposition strategy where the parallel sub-domains coincide with those of the LB scheme. In this way, each computational task performs both the LB and MD calculations and the interactions of the particles with the fluid are completely localized (there is no need to exchange data among tasks during this stage). To compensate the resulting unbalance of the MD computational load, we resort to a *local dynamic* load balancing algorithm, as outlined in the following.

At first, during the *pre-processing* step a subset of particles is assigned to each computational task according to the position in space of the particles. As the simulation proceeds, particles migrate from one domain to another and particle coordinates, momenta and identities are re-allocated among tasks. Non-bonding forces

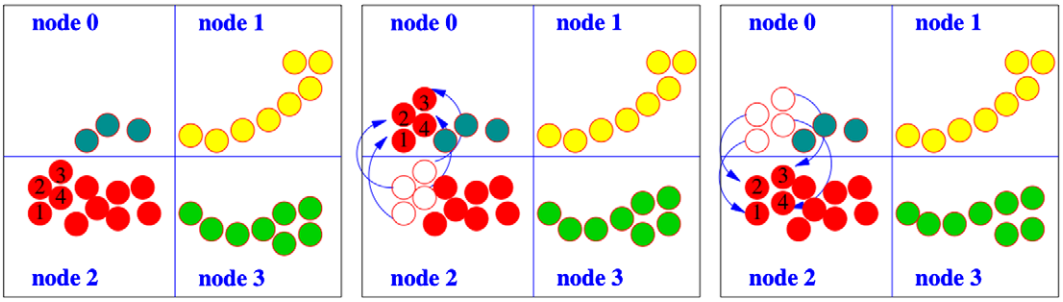

Fig. 2. Local load balancing in a simplified 2D case. Prior to the MD phase (left panel) node 2 has 11 beads whereas node 0 has only three beads. Node 1 and node 3 have the same number of beads. To balance the load during the MD phase, node 2 "virtually" transfers the coordinates of four beads to node 0 (central panel) but it remains the "owner" of those beads. After the calculation of the non-bonding forces, the results (the computed forces) are returned to node 2 (right panel). The color of the beads is used to identify those belonging to the same subdomain. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between intra- and inter-domain pairs of particles, involving the communication of particle positions between neighboring tasks, are computed. The molecular topology is also taken into account by exchanging details on the molecular connectivity, in order to compute bonding forces locally. In this respect, the way parallel MD is implemented is rather standard [22] and will not be described in detail. The main difference with respect to the standard procedure is that each task carries out an additional step, that is the exchange of momentum with the fluid. By design, this operation is carried out by each task without exchanging data with other tasks.

The dynamic load balancing strategy impacts the computation of the non-bonding forces, representing the most CPU demanding part of MD. At first, the strategy requires a communication operation between neighboring tasks that tracks the number of particles assigned to the neighbors. On the IBM BlueGene, it is more efficient to resort to a global collective communication primitive (*i.e.*, the *AllGather*) which involves all processors than to a sequence of 6 point-to-point communications with the nearest neighbor tasks. Our strategy, depicted in Fig. 2, works as follows. Whenever the number of interacting pairs owned by a given task exceeds the number of pairs assigned to a neighbor by a threshold, a fraction of the excess pairs is sent to that neighbor. This way, a *local* load balancing for the computation of non-bonding forces is achieved. A set of *precedence* rules prevents situations in which a task sends pairs to a neighbor and receives pairs from another. The receiving task computes the corresponding non-bonding forces and sends them back to the task that actually owns the particles. For the system under study, the communication/computation ratio is such that this strategy results in a sizeable advantage for the global efficiency. Fig. 3 shows the global structure of the MUPHY code.

```
do Read input data
do Run time preprocessing for the set up of the communication pattern
n = number of particles
ntime = number of time steps
k = frequency of diagnostics
for t=0 to ntime do
  do Lattice Boltzmann step
  if n > 0 then
    do Exchange momentum between fluid and particles
    do Molecular Dynamics
    do Exchange momentum between particles and fluid
  end if
  if (n mod k) = 0 then
    do Diagnostics and periodic output
  end if
end for
do Final output
```

Fig. 3. The structure of the MUPHY code.

4. Test case

Motivated by recent experimental studies, we applied our multi-scale approach to the simulation of the translocation of biopolymers through nanopores. DNA translocation is an important biophysical process occurring in phenomena like viral infection by phages, inter-bacterial DNA transduction or gene therapy [29]. In addition, it is argued that this type of process may open a way to ultrafast DNA-sequencing by sensing the base-sensitive electronic signal as the biopolymer passes through a nanopore with attached electrodes.

The translocation is a complex phenomenon involving the competition between many-body interactions at the atomic or molecular scale, fluid–atom hydrodynamic coupling, as well as the interaction of the biopolymer with the wall in the region of the pores. In our test case, many long polymers translocate through a regular array of pores, a nanodevice that was recently designed as a high-throughput sequencing device [28].

In our simulations, we use a three-dimensional box of size $N_x \times N_y \times N_z$ in units of the lattice spacing Δx. Periodicity is imposed for both the fluid and the polymer in all directions. A separating wall is located in the mid-section of the x-direction, at $x/\Delta x = N_x/2$, with an array of 64 holes of side $h = 3\Delta x$ (distributed in a regular way along 8 rows and 8 columns) through which 64 polymers (each composed by 4000 beads) translocate from one chamber to the other. The total particle and mesh size is 256000 beads and $N_x = N_y = N_z = 512$. At $t = 0$ the polymer resides entirely in the right chamber at $x/\Delta x > N_x/2$. Two snapshots of the polymers configuration in the beginning and in the mid of the simulation are reported in Fig. 4. From this picture, the load unbalancing of the MD part is apparent. Translocation is induced by a constant electric force (F_{drive}) which acts along the x-direction and is confined in a rectangular channel of size $3\Delta x \times \Delta x \times \Delta x$ along the streamwise (x-direction) and cross-flow (y-, z-directions). See Fig. 5. The solvent density and kinematic viscosity are 1 and 0.1, respectively, and the temperature is $k_B T = 10^{-4}$. Additional details can be found in Ref. [30].

In a typical run the Lattice Boltzmann part of the code takes approximately 65% of the computing time. The exchange of momentum between MD and LB takes 3% of the time. The Molecular Dynamics takes approximately 30% of the time. Diagnostics and periodic output take the rest.

We ran our tests for the simulation of 64 polymers traversing 64 holes (for a total of 256000 particles) in a $512 \times 512 \times 512$ box on an IBM BlueGene system with up to 16384 nodes, corresponding to 32768 Virtual Processors. Visual inspection (see Fig. 4)

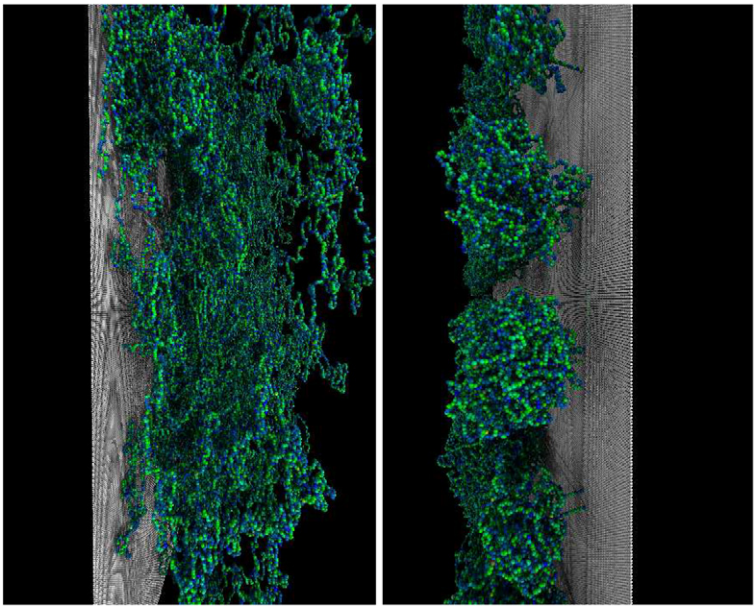

Fig. 4. The left panel shows the configuration right after the beginning of the translocation. The right panel shows the configuration near the end of the translocation. Translocation occurs from right to the left. The beads are colored according to the value of the hydrodynamic work *per* unit time they absorb from the fluid. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

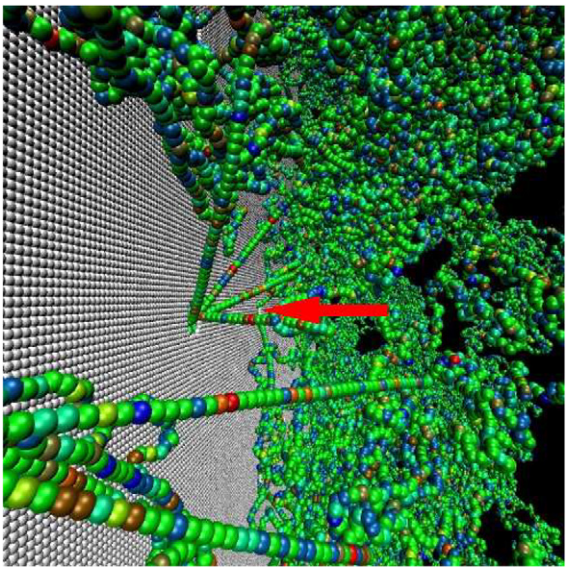

Fig. 5. A detail of the wall with one of the nanopores and a polymer entering the pore driven by the electric field (indicated by the red arrow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reveals that in the initial stage of translocation, the polymers are crowding one chamber in a rather dense and uniform way, and their distribution decays to zero at the far away edge of the initial translocation chamber. The second translocation chamber is initially empty and gradually fills up as translocation proceeds. As shown in Table 2, the time per step of a simulation of this size is < 0.04 sec on 32768 processors in *Virtual Node* (*VN*) mode, with a reasonable scaling for such large number of processors taking into account the special features of the problem under study. By analyzing the different sections of the code, the LB part appears to scale linearly (at times even superlinearly, probably due to a positive side effect of the reduction of lattice nodes *per* task on cache usage). To be noted that also the MD part exhibits a significant speed-up, showing that the saturation regime is still far from being hit. Without the dynamic local load balancing, the time required by the MD part of the simulation increases, on average, by ∼30%. It is also interesting to note that with a $512 \times 512 \times 512$ box and 32768 tasks, each task owns 4096 lattice nodes. The fact that the scalability remains good, especially for the LB part of the simulation, confirms the efficiency of the IBM BlueGene communication network. All floating point computations are performed in double precision.

Table 2
Times (in seconds) for 1000 iterations of a 256000 particles multi-biopolymer translocation in a $512 \times 512 \times 512$ lattice. The simulation requires at least 1024 tasks due to its large memory requirements.

Number of tasks	Total time	Total efficiency	LB time	MD time
1024 (CO)	883.0	N/A	581.4	260.0
2048 (CO)	452.3	98%	290.0	135.2
4096 (CO)	233.3	95%	144.3	70.5
8192 (CO)	118.6	93%	72.1	38.1
16384 (CO)	59.2	93%	36.1	21.1
16384 (VN)	66.2	83%	40.1	23.0
32768 (VN)	36.1	76%	20.1	13.0

For a correct interpretation of the timings reported in Table 2 it is worth noting that in the beginning of the simulation (approximately) half of the processors sit completely idle during the molecular dynamics part of the run. As the simulation proceeds, processors in charge of the region of the lattice where the polymers translocate, receive particles so those processors are no longer idle during that phase. At no stage of the simulation, uniform load balancing is achieved because the translocating molecules never fill up completely the computational domain.

Most results were produced by using the nodes in "CoProcessor" (CO) BlueGene mode. It is worth noticing that switching from CO to the "Virtual Node" (VN) mode, at the same number of computational nodes (i.e. hardware resources) impacts both the LB and the MD section. As a matter of fact, in VN mode, the performance does not double since all resources (cache, memory bandwidth, *etc.*) are shared between the two virtual CPUs of the node.

To measure the number of floating points operations *per* second, we resorted to the IBM HPCT library, which supports multiple instrumentation sections and nested instrumentation. The LB part

Table 3
Communication overhead on one out of 1024 processors for 1000 time-steps of the translocation of 64×4000 particles in a $512 \times 512 \times 512$ lattice.

MPI routine	N. calls	Avg. bytes	Time (sec)
MPI_Send	609080	2260	41.0
MPI_Irecv	605185	393100	0.4
MPI_Waitall	26005	0.0	32.1
MPI_Allgather	2004	8	34.5
MPI_Reduce	712	28	4.8
MPI_Allreduce	14112	128	1.4

of the simulation carries out slightly more than 210 Mflops/sec *per* task. Let us remind that this part includes the simulation of the hydrodynamic fluctuations (as a consequence its performance cannot be measured in the *standard* Lattice-Updates-per-second unit) and the load balancing is almost perfect across all tasks (the tasks whose computational domain includes the separating wall show a limited deviation). As for the MD part, *on average*, each task performs slightly below 160 Mflops/sec, with about 50% of the tasks having zero load. Taken together, *on average*, each task performs at $\simeq 190$ Mflops/sec. On the largest configuration at our disposal (32768 computational tasks) these figures lead to an estimate of a total of 6.2 Teraflops/sec aggregate performance. In addition, we ran some tests on 16 nodes of an IBM/SP system. The nodes are connected by a proprietary IBM switch and each node features 8 Power5 processors at 1.9 GHz with 32 GB of RAM per node for a total of 128 processors and 512 GB of RAM. The execution time for 1000 iterations has been, with this system, 2005 seconds, which leads to estimate that almost 10000 Power5 processors are required to obtain the same performance of the BlueGene system (assuming a similar efficiency).

Table 3 shows the distribution of communication overheads for a biopolymer translocation using 1024 tasks. For one thousand iterations, the total execution time is about 880 seconds with a communication overhead of slightly more than 110 seconds, in line with other molecular dynamics codes. The resulting *sustained* performance of point-to-point communications is $\geqslant 32$ Mbytes/sec. A similar profile when the `MPI_Isend` primitive is used confirms that, at least on the BlueGene platform, there is no real advantage in asynchronous send operations since time apparently saved in the *send* operation is actually spent in the corresponding *wait* operation.

5. Flexibility and future perspectives

To the best of our knowledge, MUPHY is one of the first examples of integrated high performance parallel code for the simulation of multi-physics/scale bio-fluidic phenomena. To place our result in perspective, let us consider that, for the DNA translocation case, one polymer bead corresponds to about 150 base pairs, that is one persistence length of double-stranded DNA (50 nm), whereas a single timestep covers about 170 ps [7]. With our sustained performances, a one-day simulation on the largest BlueGene/L configuration (212992 processors) would cover a physical time span of about 3 ms for a domain of 20 μm in linear size. These space–time scales are clearly beyond reach of fully atomistic simulations and can only be attained by multi-scale/multi-physics approach.

For a number of years, sharp-shooted, highly-tuned, application codes have been used to run large scale simulations in many different fields of computational physics. However the attempts of coupling such application codes to simulate more complex, interdisciplinary, phenomena, have been often based on a simple "serial-pipe" paradigm (*i.e.*, the output of the microscopic code provides the input of the macroscopic code) with very limited (if any) run-time concurrency and system integration.

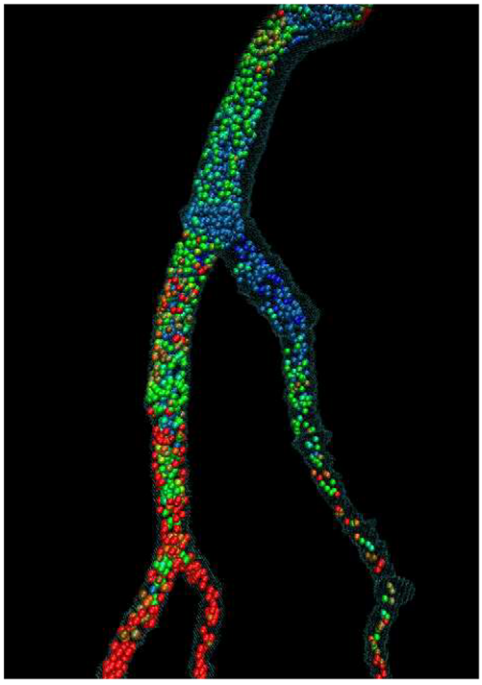

Fig. 6. Blood flow pattern in a human coronary artery. Spheres represent transported white cells, whose size has been magnified as compared to realistic values, to highlight their distribution inside the arterial region. The color of the spheres represent the Shear Stress (SS) experienced locally during blood motion (blue: low SS, green: intermediate SS, red: high SS). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Although in the present paper we focused on just one specific application of MUPHY (the translocation of biopolymers), the code can be used to study a variety of other systems. For instance, thanks to indirect addressing, MUPHY seamlessly handles real-life geometrical set-ups, such as blood flows in human arteries in presence of white cells (see Fig. 6 for an example), lipids, drug-delivery carriers and other suspended bodies of biological interest.

Acknowledgements

M.B., S.S. and S.M. wish to thank the Initiative for Innovative Computing of Harvard University for financial support and the Harvard Physics Department and School of Engineering and Applied Sciences for their hospitality. M.F. acknowledges support by Harvard's Nanoscale Science and Engineering Center, funded by NSF (Award No. PHY-0117795).

We thank Giorgio Amati for useful discussions about the LB method. We thank IBM Research for granting access to their BlueGene installation in Yorktown Heights and CASPUR for access to their IBM/SP system.

References

[1] P. Lethbridge, Multiphysics analysis, The Industrial Physicist, Dec. 2004/Jan. 2005.
[2] G. Lu, E. Kaxiras, Overview of Multiscale Simulations of Materials, in: M. Rieth, W. Schommers (Eds.), Handbook of Theoretical and Computational Nanothechnology, vol. X, American Scientific Publishers, 2005, pp. 1–33.
[3] R. Benzi, S. Succi, M. Vergassola, Phys. Rep. 222 (1992) 145–197.
[4] D.A. Wolf-Gladrow, Lattice Gas Cellular Automata and Lattice Boltzmann Models, Springer-Verlag, New York, 2000.
[5] S. Succi, The Lattice Boltzmann Equation, Oxford University Press, Oxford, 2001.
[6] P. Ahlrichs, B. Duenweg, Lattice-Boltzmann simulation of polymer–solvent systems, Int. J. Mod. Phys. C 9 (1999) 1429–1438;
P. Ahlrichs, B. Duenweg, Simulation of a single polymer chain in solution by combining lattice Boltzmann and molecular dynamics, J. Chem. Phys. 111 (1999) 8225–8239.

[7] M.G. Fyta, S. Melchionna, E. Kaxiras, S. Succi, Multiscale coupling of molecular dynamics and hydrodynamics: application to DNA translocation through a nanopore, Multiscale Model. Simul. 5 (2006) 1156–1173.
[8] M. Fyta, J. Sircar, E. Kaxiras, S. Melchionna, M. Bernaschi, S. Succi, Parallel multiscale modeling of biopolymer dynamics with hydrodynamic correlations, Intl. J. Multiscale Comput. Eng. 6 (2008) 25.
[9] R. Adhikari, K. Stratford, M.E. Cates, A.J. Wagner, Fluctuating lattice-Boltzmann, Europhys. Lett. 71 (2005) 473–477.
[10] B. Dünweg, A.J.C. Ladd, Lattice Boltzmann simulations of soft matter systems, Adv. Polymer Sci. 89 (2009) 221.
[11] S. Melchionna, Design of quasi-symplectic propagators for Langevin dynamics, J. Chem. Phys. 127 (2007) 044108-1/10.
[12] X.B. Nie, S.Y. Chen, W.N. E, M. Robbins, A continuum and molecular dynamics hybrid method for micro and nano-fluid flow, J. Fluid Mech. 500 (2004) 55.
[13] T. Werder, J.H. Walther, P. Koumoutsakos, Hybrid atomistic-continuum method for the simulation of dense fluid flows, J. Comp. Phys. 205 (2005) 373.
[14] http://www-03.ibm.com/systems/deepcomputing/bluegene.
[15] G. Amati, R. Piva, S. Succi, Massively parallel Lattice-Boltzmann simulation of turbulent channel flow, Intl. J. Modern Phys. C 4 (1997) 869–877.
[16] M. Mazzeo, P.V. Coveney, HemeLB: A high performance parallel lattice-Boltzmann code for large scale fluid flow in complex geometries, Comput. Phys. Comm. 178 (2008) 894–914.
[17] L.L. Boyer, G.S. Pawley, Molecular dynamics of clusters of particles interacting with pairwise forces using a massively parallel computer, J. Comp. Phys. 78 (1988) 405–423.
[18] H. Heller, H. Grubmuller, K. Schulten, Molecular dynamics simulation on a parallel computer, Molec. Sim. 5 (1990) 133–165.
[19] D. Rapaport, Multi-million particle molecular dynamics: II. Design considerations for distributed processing, Comput. Phys. Comm. 62 (1991) 198–216.
[20] D. Brown, J.H.R. Clarke, T. Okuda, M. Yamazaki, A domain decomposition parallelization strategy for molecular dynamics simulations on distributed memory machines, Comput. Phys. Comm. 74 (1993) 67–80.
[21] K. Esselink, B. Smit, P.A.J. Hilbers, Efficient parallel implementation of molecular dynamics on a toroidal network: I. Parallelizing strategy, J. Comp. Phys. 106 (1993) 101–107.
[22] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comp. Phys. 117 (1995) 1–19.
[23] G. Wellein, T. Zeiser, S. Donath, G. Hager, On the single processor performance of simple lattice Boltzmann kernels, Computers & Fluids 35 (2006).
[24] A. Dupuis, B. Chopard, Lattice gas: an efficient and reusable parallel library based on a graph partitioning technique, in: P. Sloot, M. Bubak, A. Hoekstra, B. Hertzberger (Eds.), HPCN Europe 1999, Springer, 1999.
[25] M. Schulz, M. Krafczyk, J. Toelke, E. Rank, Parallelization strategies and efficiency of CFD computations in complex geometries using the lattice Boltzmann methods on high-performance computers, in: M. Breuer, F. Durst, C. Zenger (Eds.), High Performance Scientific and Engineering Computing, Proceedings of the 3rd International Fortwihr Conference on HPESC, Erlangen, March 12–14, 2001, in: Lecture Notes in Computational Science and Engineering, vol. 21, Springer, 2002.
[26] L. Axner, J.M. Bernsdorf, T. Zeiser, P. Lammers, J. Linxweiler, A.G. Hoekstra, Performance evaluation of a parallel sparse lattice Boltzmann solver, J. Comp. Phys. 227 (2008) 10.
[27] http://glaros.dtc.umn.edu/gkhome/views/metis/.
[28] M.J. Kim, M. Wanunu, D.C. Bell, A. Meller, Rapid fabrication of uniformly sized nanopores and nanopore arrays for parallel DNA analysis, Adv. Mater. 18 (2006) 3149–3153.
[29] H. Lodish, D. Baltimore, A. Berk, S. Zipursky, P. Matsudaira, J. Darnell, Molecular Cell Biology, W.H. Freeman and Company, New York, 1996.
[30] M. Fyta, E. Kaxiras, S. Melchionna, M. Bernaschi, S. Succi, Quantized current blockade and hydrodynamic correlations in biopolymer translocation through nanopores: evidence from multiscale simulations, Nano Letters 8 (4) (2008) 1115–1119.