Flexible composition and execution of high performance, high fidelity multiscale biomedical simulations

Groen, D.J.; Borgdorff, J.; Bona Casas, C.; Hetherington, J.; Nash, R.W.; Zasada, S.J.; Saverchenko, I.; Mamonski, M.; Kurowski, K.; Bernabeu, M.O.; Hoekstra, A.G.; Coveney, P.V.

Published in:
Interface Focus

DOI:
10.1098/rsfs.2012.0087

Citation for published version (APA):
Flexible composition and execution of high performance, high fidelity multiscale biomedical simulations

D. Groen1, J. Borgdorff2, C. Bona-Casas2, J. Hetherington1, R. W. Nash1, S. J. Zasada1, I. Saverchenko3, M. Mamonski4, K. Kurowski4, M. O. Bernabeu1, A. G. Hoekstra2 and P. V. Coveney1

1Centre for Computational Science, University College London, UK
2Section Computational Science, University of Amsterdam, The Netherlands
3Leibniz-Rechenzentrum, Garching, Germany
4Poznan Supercomputing and Networking Center, Poznan, Poland

Multiscale simulations are essential in the biomedical domain to accurately model human physiology. We present a modular approach for designing, constructing and executing multiscale simulations on a wide range of resources, from laptops to petascale supercomputers, including combinations of these. Our work features two multiscale applications, in-stent restenosis and cerebrovascular bloodflow, which combine multiple existing single-scale applications to create a multiscale simulation. These applications can be efficiently coupled, deployed and executed on computers up to the largest (peta) scale, incurring a coupling overhead of 1–10% of the total execution time.

1. Introduction

Models of biomedical systems are inherently complex; properties on small time and length scales, such as the molecular or genome level, can make a substantial difference to the behaviour observed on much larger scales, such as the organ, full-body and even the population level; and vice versa [1,2]. We therefore need to apply multiscale approaches when modelling many biomedical situations. Examples of multiscale biomedical challenges include predicting the impact of a surgical procedure [3], investigating the effects of pathologies (e.g. arterial malformations and fistulas [4]) or assessing the effects of a targeted drug on a given patient [5]. In all these cases, we need to examine processes that not only occur across several time and/or length scales, but that also rely on different underlying physical and/or biological mechanisms. As a result, modelling these processes may require substantially different algorithms and varying levels of computational effort.

Historically, these problems have often been modelled using single-scale approaches, focusing exclusively on those aspects of the problem which are deemed most relevant. However, applying a single-scale model is frequently insufficient to fully understand the problem at hand, as additional processes occurring on different scales must be incorporated to obtain sufficient accuracy. It is in this need for understanding the composite problem, rather than its individual subcomponents alone, that has driven many research groups to explore multiscale modelling [6–9].

In a multiscale model, the overall system is approximated by coupling two or more single-scale submodels. Establishing and performing the data exchange between these submodels is an important aspect of enabling multiscale modelling. It is often addressed by using coupling tools, such as the Multiscale Library and Coupling Environment (MUSCLE) [10,11], the Multilevel Communicating Interface [12] and GridSpace [13]. Groen et al. [14] provide a review of coupling tools and the computational challenges they address.
2. Multiscale biomedical modelling

Multiscale modelling gives rise to a number of challenges which extend beyond the translation of model and system data. Most importantly, we seek to allow application developers, such as computational biologists and biomedics, to couple multiscale models for large problems, supporting any type of coupling, using any type of submodel they wish to include, and executing this using any type of computational resource, from laptop to petascale. As we cannot expect computational biologists to have expertise in all the technical details of multiscale computing, we also aim to present a uniform and easy-to-use interface which retains the power of the underlying technology. This also enables users with less technical expertise to execute previously constructed multiscale simulations.

Multiscale systems are, in general, characterized by the interaction of phenomena on different scales, but the details vary for different scientific domains. To preserve the generality of our approach, we adopt the multiscale modelling and simulation framework (MMSF) to reason about multiscale models in a domain-independent context and to create recipes for constructing multiscale simulations independent of the underlying implementations or computer architectures. This MMSF is based on earlier work on coupled cellular automata and agent-based systems [17,18] and has been applied to several computational problems in biomedicine [3,19,20]. Within the MMSF, the interactions between single-scale submodels are confined to well-defined couplings. The submodels can therefore be studied as independent models with dependent incoming and outgoing links. The graph of all submodels and couplings, the coupling topology, can either be cyclic or acyclic [21]. In a cyclic coupling topology, the submodels will exchange information in an iterative loop, whereas in an acyclic topology the submodels are activated one after another, resulting in a directional data flow which makes them well-suited for workflow managers. Two parts of the MMSF are particularly useful for our purposes, namely the Scale Separation Map [17] and the multiscale modelling language (MML) [11,22]. The Scale Separation Map is a graphical representation which provides direct insights into the coupling characteristics of the multiscale application. MML provides a formalization of the coupling between the submodels, independent of the underlying implementation. In addition, it simplifies the issues associated with orchestrating submodels by capturing the orchestration mechanisms in a simple model which consists of only four distinct operators. MML definitions can be stored for later use using an XML-based file format (xMML) or represented visually using graphical MML (gMML) [11].

The generic MML definitions allow us to identify commonalities between multiscale applications in different scientific domains. Additionally, the stored xMML can be used as input for a range of supporting tools that facilitate multiscale simulations (e.g. tools that automate the deployment or the submodel coupling of these simulations, as discussed in §3).

3. From multiscale model to production simulation

We have developed a range of tools that allow us to create, deploy and run multiscale simulations based on the xMML specification of a multiscale model. Two of the main challenges in realizing successful multiscale simulations are to establish a fast and flexible coupling between submodels and to deploy and execute the implementations of these submodels efficiently on computer resources of any size. In this section, we review these challenges and present our solutions to them.

3.1. Coupling submodels

Effective and efficient coupling between submodels encompasses three major aspects. First, we must translate our MML definitions to technical recipes for the multiscale simulation execution. Second, we need to initiate and orchestrate the execution of the submodels in an automated way, in accordance with the MML specification. Third, we need to efficiently exchange model and system information between the submodels.

The MUSCLE [10,11] provides a solution to the first two aspects. It uses MML in conjunction with a definition of the requested resources to bootstrap and start the different submodels and to establish a connection for data exchange between the submodels. It also implements the coupling definitions defined in the MMSF to orchestrate the submodels. The submodels can be coupled either locally (on a workstation, for example),
via a local network, or via a wide area network using the MUSCLE Transport Overlay (MTO) [23]. Running submodels in different locations and coupling them over a wide area network is especially important when submodels have very different computational requirements, for example when one submodel requires a local machine with a set of Python libraries, and another submodel requires a large supercomputer. However, messages exchanged across a wide area network do take longer to arrive. Among other things, MTO provides the means to exchange data between large supercomputers while adhering to the local security policies and access restrictions.

The exchanges between submodels overlap with computations in many cases (see §6 for an example), allowing for an efficient execution of the overall simulation. However, when the exchanges are particularly frequent or the exchanged data particularly large, inefficiencies in the data exchange can severely slow down the multiscale simulation as a whole. We use the MPWide communication library [24], previously used to run cosmological simulations distributed across supercomputers [25], to optimize our wide area communications for performance. We already use MPWide directly within the cerebrovascular bloodflow simulation and are currently incorporating it in MUSCLE to optimize the performance of the MTO.

3.2. Deploying and executing submodels on production resources

When large multiscale models are deployed and executed as simulations on production compute resources, a few major practical challenges arise. Production resources are often shared by many users, and are difficult to use even partially at appropriate times. This in turn makes it difficult to ensure that the submodels are invoked at the right times, and in cases where they are distributed on different machines, to ensure that the multiscale simulation retains an acceptable time to completion. We use the QosCosGrid environment (QCG) [26] to run submodels at a predetermined time on large computing resources. QCG enables us, if installed on the target resource, to reserve these resources in advance for a given period of time to allow our submodels to be executed there. This is valuable in applications that require cyclic coupling, as we then require multiple submodels to be executed either concurrently or in alternating fashion. Additionally, with QCG we can explicitly specify the sequence of resources to be reserved, allowing us to repeat multiple simulations in a consistent manner using the same resource sequence each time. A second component that aids in the execution of submodels is the application hosting environment (AHE) [27]. AHE simplifies user access to production resources by streamlining the authentication methods and by centralizing the installation and maintenance tasks of application codes. The end users of AHE only need to work with one uniform client interface to access and run their applications on a wide range of production resources.

4. Using MAPPER for biomedical problems

We apply the MAPPER approach to a number of applications. We present two of these applications here (ISR and cerebrovascular BF), but we have also used MAPPER to define, deploy and execute multiscale simulations of clay–polymer nanocomposite materials [28], river beds and canals [29], and several problems in nuclear fusion.

MAPPER provides formalisms, tools and services which aid in the description of multiscale models, as well as the construction, deployment and execution of multiscale simulations on production infrastructures. It is intended as a general-purpose solution, and as such tackles challenges in multiscale simulation that exist across scientific disciplines. There are a number of challenges which are outside the scope of our approach, because they may require different solutions for different scientific disciplines. These include choosing appropriate submodels for a multiscale simulation and defining, on the application level, what information should be exchanged between submodels at which times to provide a scientifically accurate and stable multiscale simulation. However, MMSF does simplify the latter task by providing a limited number of orchestration mechanisms. The formalisms, tools and services presented here are independent components, allowing users to adopt those parts of our approach which specifically meet their requirements in multiscale modelling. This modular approach makes it easier for users to exploit the functionalities of individual components and helps to retain a lightweight simulation environment.

5. Modelling in-stent restenosis

Coronary heart disease (CHD) is one of the most common causes of death and is responsible for about 7.3 million deaths per year worldwide [30]. CHD is typically expressed as atherosclerosis, which corresponds with a thickening and hardening of blood vessels caused by build-up of atheromatous plaque; when this significantly narrows the vessel, it is called a stenosis. A common intervention for stenosis is stent-assisted balloon angioplasty where a balloon, attached to a stent, is inserted in the blood vessel and inflated at the stenosed location, consequently deploying the stent. The stent acts as a scaffold for the blood vessel, compressing the plaque and holding the lumen open. Occasionally, however, this intervention is followed by ISR, an excessive regrowth of tissue due to the injury caused by the stent deployment [31,32]. Although there are a number of different hypotheses [33], the pathophysiological mechanisms and risk factors of ISR are not yet fully clear.

By modelling ISR with a three-dimensional model (ISR3D), it is possible to test mechanisms and risk factors that are likely to be the main contributors to ISR. After evaluating the processes involved in ISR [19], ISR3D applies the hypothesis that smooth muscle cell (SMC) proliferation drives the restenosis, and that this is affected most heavily by wall shear stress of the BF, which regulates endothelium...
recovery, and by growth inhibiting drugs diffused by a drug-eluting stent. Using the model, we can evaluate the effect of different drug intensities, physical stent designs, vascular geometries and endothelium recovery rates. ISR3D is derived from the two-dimensional ISR2D \[3, 20\] code. Compared with ISR2D, it provides additional accuracy, incorporating a full stent design, realistic cell growth and a three-dimensional BF model, although it requires more computational effort. We present several results from modelling a stented artery using ISR3D in figure 1, which shows the artery both before and after restenosis has occurred.

From a multiscale modelling perspective, ISR3D combines four submodels, each operating on a different time scale: SMC proliferation, which models the cell cycle, cell growth and physical forces between cells; initial thrombus formation (ITF) due to the backflow of blood; drug diffusion (DD) of the drug-eluting stent through the tissue and applied to the smooth muscle cells; and BF and the resulting wall shear stress on SMCs. We show the time scales of each submodel and the relations between them in figure 2. The SMC submodel uses an agent-based algorithm on the cellular scale, which undergoes validation on the tissue level. All other submodels act on a Cartesian grid representation of those cells. For the BF submodel, we use 1,237,040 lattice sites, and for the SMC submodel, we use 196,948 cells. The exchanges between the submodels are in the order of 10–20 MB. The SMC sends a list of cell locations and sizes (stored as 8-byte integers) to the ITF, which sends the geometry (stored as a three-dimensional matrix of 8-byte integers) to BF and DD. In turn, BF and DD, respectively, send a list of wall shear stress and drug concentrations (stored as 8-byte doubles) to SMC. Each coupling communication between SMC and the other submodels takes place once per SMC iteration.

The submodels act independently, apart from exchanging messages, and are heterogeneous. The SMC code is implemented in C++, the DD code in Java and the ITF code in Fortran. The BF code, which unlike the other codes runs in parallel, uses the Palabos lattice–Boltzmann application (http://www.palabos.org/), written in C++. The MML specification of ISR3D contains a few conversion modules not mentioned earlier, which perform basic data transformations necessary to ensure that the various single-scale models do not need to be aware of other submodels and their internal representation or scales. Specifically, the submodels operate within the same domain, requiring the application to keep the grid and cell representation consistent, as well as their dimensions.

5.1. Tests
We have performed a number of tests to measure both the runtime and the efficiency of our multiscale ISR3D simulation. The runs that were performed here were short versions of the actual simulations, with a limited number of iterations. The runs contributed to the integration testing of the code and allowed us to estimate the requirements for future computing resource proposals.

We have run our tests in five different scenarios, using the EGI resource in Krakow (Zeus, one scenario), a Dutch PRACE tier-1 machine in Amsterdam (Huygens, two scenarios) or a combination of both (two scenarios). We provide the technical specifications of the resources in table 1. Since the runtime behaviour of ISR3D is cyclic, determined by the number of SMC iterations, we measured the runtime of a single cycle for each scenario. Because only the BF model is parallelized, the resources we use are partially idle whenever we are not executing the BF model. To reduce this overhead of ‘idle time’, we created two double mapping scenarios, each of which runs two simulations in alternating fashion using the same resource reservation. In these cases, the BF calculations of one simulation takes place while the other simulation executes one or more of the other submodels.

We use a wait/notify signalling system within MUSCLE to enforce that only one of the two simulations indeed executes its parallel BF model. We have run one double mapping scenario locally on Huygens, and one distributed over Huygens and Zeus.

5.1.1. Results
We present our performance measurements for the five scenarios in table 3. The runtime of our simulation is reduced by almost 50 per cent when we use the Huygens machine instead of Zeus. This is because we run the BF code on Huygens using 32 cores, and on Zeus using 4 cores. However, the usage of the reservation was considerably lower.
on Huygens than on Zeus for two reasons: first, the allocation on Huygens was larger, leaving more cores idle when the sequential submodels were computed; second, the ITF solver uses the gfortran compiler, which is not well optimized for Huygens architecture. As a result, it requires only 4 min on Zeus and 12 min on Huygens. The performance of the other sequential submodels is less platform dependent. When we run our simulation distributed over both resources, we achieve the lowest runtime. This is because we combine the fast execution of the ITF solver on Huygens with the fast execution of the BF module on Huygens with the fast execution of the ITF solver on Zeus.

When we use double mapping, the runtime per simulation increases by about 6 per cent (using only Huygens) to 18 per cent (using both resources). However, the double-mapping improves the usage of the reservation by a factor of approximately 1.7. Double-mapping is therefore an effective method to improve the resource usage without incurring major increases to the time to completion of each simulation. The coupling overhead is relatively low throughout our runs, consisting of no more than 11 per cent of the runtime throughout our measurements. Using our approach, we are now modelling different ISR scenarios, exploring a range of modelling parameters.

5.2. Clinical directions

We primarily seek to understand which biological pathways dominate in the process leading to ISR. If successful, this will have two effects on clinical practice: first, it suggests which factors are important for ISR, in turn giving clinicians more accurate estimates of what the progression of the disease can be; second, it may spur further directed clinical research of certain pathways, which will help the next iteration of the model give more accurate results.

The methods for achieving this divide naturally in two directions: general model validation and experiments; and virtual patient cohort studies. For general model validation, we consult the literature and use basic experimental data, such as measurements from animal studies. In addition, we intend to use virtual patient cohort studies to assess the ISR-risk factors of virtual patients with different characteristics. In clinical practice, this will not lead to personalized estimates, but rather to patient classifiers on how ISR3D will progress. Once the primary factors leading to ISR have been assessed, a simplified model could be made based on ISR3D, which takes less computational effort and runs within a hospital.

6. Modelling of cerebrovascular blood flow

Our second hemodynamic example aims to incorporate not only the local arterial structure in our models, but also properties of the circulation in the rest of the human body. The key feature of this application is the multiscale modelling of BF, delivering accuracy in the regions of direct scientific or clinical interest while incorporating global BF properties using more approximate methods. Here we provide an overview of our multiscale model and report on its performance. A considerable amount of previous work has been done where groups combined bloodflow solvers of different types, for example, in the area of cardiovascular [34–36] or cerebrovascular bloodflow [12, 37].

We have constructed a distributed multiscale model in which we combine the open source HemeLB lattice-Boltzmann application for BF modelling in three-dimensions [38, 39] with the open source one-dimensional Python Navier–Stokes (pyNS) BF solver [40]. HemeLB is optimized for sparse geometries, such as vascular networks and has been shown to scale linearly up to at least 32 768 cores [41]. PyNS is a discontinuous Galerkin solver which is geared towards modelling large arterial structures. It uses aortic BF input based on a set of patient-specific parameters, and it combines one-dimensional wave propagation elements to model arterial vasculature with zero-dimensional resistance elements to model veins. The numerical code supports thread-level parallelization and is written in Python in conjunction with the numpy numerical library.

6.1. Simulations

We have run a number of coupled simulations using both HemeLB and pyNS as submodels. Within pyNS, we use a customized version of the ‘Willis’ model, based on [42], with a mean pressure 90 mmHg, a heart rate of 70 beats per minute and a cardiac output of 5.68 l min⁻¹. Our model includes the major arteries in the human torso, head and both arms, as well as a full model for the circle of Willis,
We have modified a section of the right mid-cerebral artery in pyNS to allow it to be coupled to HemeLB in four places, exchanging pressure values in these boundary regions. In HemeLB we simulate a small network of arteries, with a diameter of 0.52 m and consisting of about 4.2 million lattice sites which occupy 2.3 per cent of the simulation box volume. The HemeLB simulation runs with a time step of $2.3766 \times 10^{-4}$ s, a Mach number of 0.1 and a relaxation parameter $\tau$ of 0.52. We run pyNS using a local machine at UCL (Henry), while we run HemeLB for 400 000 time steps on the HECToR supercomputer in Edinburgh. The round-trip time for a network message between these two resources is on average 11 ms. We provide technical details of both machines in table 2. Both codes exchange pressure data at an interval of 100 HemeLB time steps (or 1 pyNS time step). Because HemeLB time steps can take as little as 0.0002 s, we adopted MPWide to connect our submodels, which run concurrently, and minimize the communication response time. The exchanged data are represented using 8-byte doubles and has a small aggregate size (less than 1 kb). As a comparison, we have also run HemeLB as a stand-alone single-scale simulation (labelled ‘ss’), retrieving its boundary values from a local configuration file. The pyNS code requires 116 s to simulate 4000 time steps of our modified circle of Willis problem when run as a stand-alone code.

We present our results in table 4. For 512 cores, the single-scale HemeLB simulation takes 2271 s to perform its 400 000 lattice-Boltzmann time steps. The coupled HemeLB-pyNS simulation is only marginally slower, reaching completion in 2298 s. We measure a coupling overhead of only 24 s, which is the time to do 4002 pressure exchanges between HemeLB and pyNS. This amounts to about 6 ms per exchange, well below even the round-trip-time of the network between UCL and EPCC alone. The communication time on the HemeLB side is so low because pyNS runs faster per coupling iteration than HemeLB, and the incoming pressure values are already waiting at the network interface when HemeLB begins to send out its own. As a result, the coupling overhead is only 24 s, and the multiscale simulation is only 1.2 per cent slower than a single-scale simulation of the same network domain.

When using 2048 cores, the runtime for the single-scale HemeLB simulation is 815 s, which is a speed up of 2.63 compared with the 512 core run. The coupling overhead of the multiscale simulation is relatively higher than that of the 512 core run due to a larger number of processes with which pressures must be exchanged. This results in an overall coupling efficiency of 0.95, which is lower than for the 512 core run. However, the 2048 core run contains approximately 2000 sites per core, which is a regime where we no longer achieve linear scalability [41] to begin with. As a future task, we plan to coalesce these pressure exchanges with the other communications in HemeLB, using the coalesced communication pattern [43].

### Table 3.

<table>
<thead>
<tr>
<th>scenario</th>
<th>BF (time spent)</th>
<th>other (time spent)</th>
<th>coupling (time spent)</th>
<th>total (time spent)</th>
<th>usage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeus</td>
<td>2100</td>
<td>1140</td>
<td>79</td>
<td>3240</td>
<td>80</td>
</tr>
<tr>
<td>Huygens</td>
<td>480</td>
<td>1260</td>
<td>73</td>
<td>1813</td>
<td>27</td>
</tr>
<tr>
<td>Huygens-double</td>
<td>480</td>
<td>1320</td>
<td>131</td>
<td>1931</td>
<td>45</td>
</tr>
<tr>
<td>Zeus – Huygens</td>
<td>480</td>
<td>960</td>
<td>92</td>
<td>1532</td>
<td>32</td>
</tr>
<tr>
<td>Zeus – Huygens-double</td>
<td>480</td>
<td>1080</td>
<td>244</td>
<td>1804</td>
<td>56</td>
</tr>
</tbody>
</table>

### Table 4.

<table>
<thead>
<tr>
<th>ss/ms</th>
<th>cores</th>
<th>init (time spent)</th>
<th>HemeLB* (time spent)</th>
<th>coupling (time spent)</th>
<th>total (time spent)</th>
<th>coupling efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ss</td>
<td>512</td>
<td>47.9</td>
<td>2223</td>
<td>n.a.</td>
<td>2271</td>
<td>n.a.</td>
</tr>
<tr>
<td>ms</td>
<td>512</td>
<td>51.2</td>
<td>2223</td>
<td>24</td>
<td>2298</td>
<td>98.8%</td>
</tr>
<tr>
<td>ss</td>
<td>2048</td>
<td>46.4</td>
<td>815</td>
<td>n.a.</td>
<td>907</td>
<td>95.0%</td>
</tr>
<tr>
<td>ms</td>
<td>2048</td>
<td>50.5</td>
<td>815</td>
<td>37</td>
<td>907</td>
<td>95.0%</td>
</tr>
</tbody>
</table>

6.2. Clinical directions

We aim to understand the flow dynamics in cerebrovascular networks and to predict the flow dynamics in brain aneurysms for individual patients. The ability to predict the flow dynamics...
in cerebrovascular networks is of practical use to clinicians, as it allows them to more accurately determine whether surgery is required for a specific patient suffering from an aneurysm. In addition, our work supports a range of other scenarios, which in turn may drive clinical investigations of other vascular diseases.

Our current efforts focus on enhancing the model by introducing velocity exchange and testing our models for accuracy. As a first accuracy test, we compared different boundary conditions and flow models within HemeLB [39]. Additionally we used our HemeLB-pyNS setup to compare different blood rheology models, which we describe in detail in [44]. Next steps include more patient-specific studies, where we wish to compare the flow behaviour within patient-specific networks of arteries in our multi-scale simulations with the measured from those same patients. We have already established several key functionalities to allow patient-specific modelling. For example, HemeLB is able to convert three-dimensional rotational angiographic data into initial conditions for the simulation, while pyNS provides support for patient-specific global parameters and customized arterial tree definitions.

7. Conclusions and future work

We have shown that MAPPER provides a usable and modular environment which enables us to efficiently map multiscale simulation models to a range of computing infrastructures. Our methods provide computational biologists with the ability to more clearly reason about multiscale simulations, to formally define their multiscale scenarios, and to more quickly simulate problems that involve the use of multiple codes using a range of resources. We have presented two applications, in-stent restenosis and cerebrovascular blood-flow, and conclude that both applications run rapidly and efficiently, even when using multiple compute resources in different geographical locations.

We thank our colleagues in the MAPPER consortium, the HemeLB development team at UCL, as well as Simone Manini and Luca Antiga from the pyNS development team. This work received funding from the MAPPER EU-FP7 project (grant no. RI-261507) and the CRESTA EU-FP7 project (grant no. RI-287703). We made use of computational resources provided by the PL-Grid Infrastructure (Zeus), by PRACE at SARA in Amsterdam, The Netherlands (Huygens) and EPCC in Edinburgh, UK (HECToR), and by University College London (Henry).

References


