Towards RealTime 3D Coronary Hemodynamics Simulations during Cardiac Catheterisation

S Moore\textsuperscript{1}, K Halupka\textsuperscript{1}, S Zhuk\textsuperscript{2}

\textsuperscript{1}IBM Research, Melbourne, Australia
\textsuperscript{2}IBM Research, Dublin, Ireland

Abstract

Virtual Fractional Flow Reserve (vFFR) is an emerging technology that assesses the severity of coronary stenosis by means of patient specific of Computational Fluid Dynamics simulations. To be of practical clinical utility within a catheter laboratory, FFR results must be obtainable within minutes to guide intervention. We present the design of a novel Lattice-Boltzmann method code specifically tailored for fully automatic near real-time 3D coronary blood flow simulations. The key contributions of the work include a hybrid multicore-GPU accelerated sparse lattice generation algorithm and specialized 3D-0D coupled hemodynamics solver. We present results on state of the art GPU hardware, simulating hemodynamics within multi segment coronary tree. The results demonstrate that vFFR simulations can be performed in the order of minutes, making the replacement of pressure wire based FFR in a catheter laboratory setting with vFFR simulations feasible, without the need to reduce the fidelity of the hemodynamics modelling.

1. Introduction

Fractional Flow Reserve (FFR) is the gold standard for assessing the hemodynamic significance of a coronary stenosis and is performed during invasive coronary angiography using a diagnostic catheter to measure a patient’s blood pressure both distal and proximal to a lesion [1]. Despite its ability to reduce adverse cardiac events and unnecessary surgical interventions deploying stents [2], it is expensive and often under-utilised. The application of computational fluid dynamics simulations in personalised coronary arterial models (often termed virtual or vFFR) has emerged as a technology to address these shortcomings. The major limitation for use in a catheter laboratory is the compute time required to perform a vFFR simulation, since 3D solutions to the 3D Navier-Stokes can require hours to complete. In this work we present a novel GPU based implementation of the Lattice-Boltzmann Method (LBM), that takes as input a centre-line model of a coronary tree, and specifically designed for coronary blood flow to simulations to address this limitation.

2. Methods

The computational hemodynamics model is comprised of a novel 3D sparse lattice structure, coupled to a 0D model representing the remainder of the cardiovascular system (Figure 1). The LBM implemented in the present study is based upon the standard LBGK approach using the D3Q19 Lattice as:

\begin{equation}
\text{f}_i(x_{\alpha} + e_{i\alpha} \Delta x, t + \Delta t) = \text{f}_i(x_{\alpha}, t) + \frac{\text{f}_i^\text{eq}(x_{\alpha}, t) - f_i(x_{\alpha}, t)}{\tau} \tag{1}
\end{equation}

where $f_i$ denotes the distribution function, $i \in [1-19]$, $\alpha \in [1-3]$, $\tau$ is the relaxation time, and the equilibrium distribution function:

\begin{equation}
\text{f}_i^\text{eq} = w_i \rho \left(1 + \frac{\epsilon_{i\alpha} u_{\alpha}}{c_s^2} + \frac{u_{\alpha} u_{\beta} (\epsilon_{i\alpha} \epsilon_{i\beta} - c_s^2 \delta_{\alpha\beta})}{2 c_s^4} \right) \tag{2}
\end{equation}

where the blood density and moment are defined $\rho = \sum f_i$, $\rho u_{\alpha} = \sum f_i c_{i\alpha}$ respectively, and $c_s = \frac{1}{\sqrt{3}}$ is the speed of sound[3]. In order to capture some of the key phenomena exhibited by coronary hemodynamics, two lumped parameter (LP) models are used for the specification of inlet and outlet pressures respectively (Figure 1). At the inlet, the cardiovascular model of Simaan et al. [4]. At each outlet, the coronary outlet model of Sankaran et al. [5] was used. Parameters were taken from [6] and each LP model is explicitly coupled imposing the systemic and outlets pressures as Dirichlet pressure boundary conditions, solved using the Runge-Kutta method with a time step size of $1.0 \times 10^{-6} \text{s}$. The implementation uses a novel sparse lattice structure and generator, where \textit{only} the fluid sites are stored in memory. Each type of lattice site (interior, wall, inlet, outlet/s) are assigned a unique lattice ID (Figure 1) and sorted into a list such that each type is contiguous in memory. The indices of neighbouring sites hence
explicitly stored in a ‘neighbour list’ table that is used when performing the streaming step, moving distributions within GPU memory. To eliminate errors incurred through obliquely oriented inlet or outlet (iolet) planes with respect to the lattice, the input centreline is first process to extrude and align iolets with the closest lattice plane (Figure 1). The lattice generation process then makes use of threaded CPU and GPU parallelism with a ‘polyballModelling’ filter [7] to locate interior sites in GPU memory and placing these sites into an unordered map structure in CPU memory that can be used to insert wall sites and sorted in order to create the neighbour list. The LBM implementation uses Single precision floating point arithmetic and a standard structure of arrays (SoA) where distributions for a given speed are kept contiguous in memory in order to allow for coalesced memory access. During each time step a collide, boundary, and stream kernel is called. In the boundary kernel, the flowrate is first computed a sum reduction based on warp intrinsics to sum up velocities over an outlet to return to CPU memory and to the lumped parameter model.

3. Validation

Numerical experiments were performed with a coronary centreline model extracted from a Computed Tomography Coronary Angiography (CTCA) dataset. For validation, the LBM simulations were compared against a custom OpenFOAM solver based on the PIMPLE method, using a hex-dominant mesh with uniform resolution of \(7.5 \times 10^{-5} [m]\) (approximately 6 million polyhedral cells) and a temporal resolution of \(1.0 \times 10^{-3} [s]\), which was shown to give a mesh and time step independent solution. Furthermore, to remove the error associated with the explicit coupling between the 0D and 3D domains, three cardiac cycles were simulation, which was shown to produce a periodic solution. Figure 2 presents a visual comparison of the pressure field between the OpenFOAM and LBM simulations, in particular highlighting the spatial discretisation around the stenosis. While the OpenFOAM simulation took approximately 12 hours when run as an MPI parallel job across 56 cores, the Rabid simulation took approximately 12 minutes. Figures 3(a)-3(d) present the OpenFOAM (solid curves) and LBM (dotted curves) iolet pressure and flowrate variation over a cardiac cycle respectively, then the pressure and speed variation along the black line depicted above at four points within the cardiac cycle.

4. Performance

Having shown a faithful solution to the incompressible Navier-Stokes equations commonly used in vFFR simulations, of key importance is the tradeoff between numerical error in terms of the spatial and temporal resolution (which affects the compressibility error) and the required run time. This issue was investigated by performing a number of simulations with varying spatial and temporal resolution and analysing the variation in the computed FFR measurement, versus the performance of the solver. All simulations were performed on a single IBM Cloud\textsuperscript{\textregistered} bare metal server comprised of using an Intel Xeon\textsuperscript{\textregistered} E5-2690 2.60GHz CPU with 125 GB of memory and an NVIDIA Tesla\textsuperscript{\textregistered} V100 PCIE GPU, equipped with 80 streaming multiprocessors and 16GB of memory. Figures 4(a)-4(d) illustrate the average and maximum FFR error as compared to the OpenFOAM simulation, then the simulation time per car-

Figure 1. Overview of the simulation process including centreline preprocessing, sparse lattice generation, and the 0D-3D coupling.
diac cycle, and solver performance in terms of Mega lattice updates per second (MLUPS). The FFR is defined as the ratio of the distal to proximal pressure at the ends of the black line depicted above. The average error is computed by first time averaging these two pressures and the maximum error is computed by defining the ratio after systole where peak flow occurs. Although it is the time averaged values commonly used to define FFR [8], it is during peak flow that the greatest error in the LBM simulation occurs. An interesting result is that accurate computations can be performed at the coarser spatial resolutions, as long as the temporal resolution is kept small enough. The most important result however is that an accurate 3D pulsatile vFFR simulation can be performed in within minutes, using modern algorithms and GPU hardware, in a timeframe to be applicable to a catheter laboratory setting.

5. Conclusions

The results presented demonstrate that high fidelity 3D vFFR simulations can be performed within the time frame necessary to be relevant to a catheter laboratory setting. Future work will seek aim to port the code to the new IBM
Cloud PowerAI servers allowing for execution on multiple GPUs.

References


Address for correspondence:

Stephen Moore
IBM Research Australia
Level 22, 60 City Rd, Southbank, VIC 3006, Australia
tel. +61450009534
stevemoore@au1.ibm.com