



KTH Engineering Sciences

GPU Simulation of Rigid Fibers

ERIC WOLTER

Master's Thesis at School of Engineering Sciences

Supervisor: Katarina Gustavsson

Examiner: Michael Hanke

TRITA xxx yyyy-nn

Abstract

The major objective of this Master's thesis is to accelerate a serial implementation of a numerical algorithm for the simulation of slender fiber dynamics by using Graphical Processing Units (GPU). We focus on rigid fibers sedimenting due to gravity in a free-space Stokes flow. The ability to simulate a large number of fibers in a reasonable computational time on a high-performance parallel platform opens exciting new research opportunities.

The previous serial implementation is rewritten for parallel execution. The algorithm is implemented in single precision using the Compute Unified Device Architecture (CUDA) on NVIDIA GPUs. In addition, we develop an OpenMP version of the parallel implementation to run on multi-core CPUs. Using both implementations, we perform a number of benchmarks to determine the fastest variant of the algorithm. We observe a speedup of 20 \times to 40 \times on the NVIDIA GTX 970 compared to an Intel Core i7 4770. The GPU implementation can simulate up to 2000 fibers on a desktop computer and it takes only in the order of 8 seconds to advance one time step.

Furthermore, we have performed a number of simulations of known experiments for sedimenting fibers to validate the algorithm and to explore the numerical precision of the results. The results show an excellent agreement with results from prior experiments in the field.

Referat

GPU simulering av stela fibrer

Huvudsyftet med denna masteruppsats är att accelerera en seriell implementation av en numerisk algoritm för simulering av tunnfiberdynamik med hjälp av grafikprocessorer (GPU:er). Vårt fokus ligger på sedimentering av stela fibrer som en konsekvens av gravitationens påverkan i ett frirums-Stokes-flöde. Att kunna simulera ett stort antal fibrer med en relativt kort beräkningstid på en högpresterande parallelplattform öppnar för nya spännande forskningsmöjligheter.

Den föregående seriella implementationen var skriven för parallelexekvering. Algoritmen är implementerad i enkelprecision med hjälp av Compute Unified Device Architecture (CUDA) på NVIDIA-GPU:er. Vi utvecklar dessutom en OpenMP-version av parallelimplementationen för flerkärnsprocessorer. Genom att använda båda implementationerna utför vi ett antal prestandatester för att hitta den snabbaste varianten av algoritmen. Vi observerar en ökning från 20× till 40× på NVIDIA GTX 970 jämfört med en Intel Core i7 4770. GPU-implementationen kan simula upp till 2000 fibrer på en skrivbordsdator, och det tar endast i ordningen 8 sekunder att avancera ett tidssteg.

Vi har dessutom utfört ett antal simuleringar av redan kända experiment för sedimentering av fibrer för att bekräfta att algoritmen är korrekt, samt för att utvärdera resultatens numeriska precision. Resultaten överensstämmer utomordentligt väl med tidigare resultat från experiment i området.

Acknowledgements

Contents

List of Figures	viii
List of Tables	x
List of Listings	xi
1 Introduction	1
2 Theoretical Foundation	5
2.1 Stokes flow	5
2.2 Boundary integral formulation	6
2.2.1 Fundamental solutions	6
2.2.2 Formulation for objects immersed in a fluid	7
2.3 Slender fibers	9
3 Numerical algorithm and serial implementation	13
3.1 Assemble system	13
3.1.1 Closed linear system	14
3.1.2 Inner integral	14
3.2 Solve system	15
3.3 Update velocities	16
3.4 Update fibers	16
3.5 Algorithm summary	16
4 GPU Programming	19
4.1 General Purpose Computing on GPUs	19
4.2 CUDA vs. OpenCL	21
4.3 CUDA programming model	22

5 Parallel implementation	29
5.1 Development environment	29
5.2 Linear solvers	30
5.3 Kernels	31
5.4 Optimizations	34
5.4.1 Numeric vs. analytic integration	37
5.4.2 Shared memory	37
5.4.3 Thread block dimension	38
5.5 OpenMP	40
6 Benchmarks	43
6.1 Methodology	43
6.1.1 Hardware	43
6.1.2 Benchmark scheme	44
6.2 Optimizations	46
6.2.1 Numeric vs. Analytic Integration	46
6.2.2 Shared memory	49
6.2.3 Thread block dimension	49
6.3 Linear solvers	51
6.4 Individual steps	54
6.5 GPU vs. CPU	55
7 Experiments	59
7.1 Tumbling orbits	59
7.2 Fiber concentration effect on GMRES iterations	61
7.3 Sedimenting sphere	63
7.3.1 Sphere break-up	63
7.3.2 Fiber concentration effect on break-up time	65
7.3.3 Number of fibers effect on break up time	66
7.4 Mixed density sphere	68
8 Conclusion	71
Bibliography	73
Appendices	76

List of Figures

2.1	Two immersed objects in stokes flow, using a rigid body motion.	8
2.2	Illustration of the slenderness parameters.	9
2.3	Slender fiber approximation.	10
2.4	Two interacting fibers with five quadrature points.	11
4.1	Year over year increase in theoretical floating-point operations per second for CPUs and GPUs.	20
4.2	Overview of the CUDA platform.	23
4.3	Automatic scaling of blocks across an arbitrary number of Streaming Multiprocessors.	24
4.4	CUDA block addressing using 2D grid with 2D thread blocks.	26
4.5	CUDA memory hierarchy.	28
5.1	Illustration of 1D and 2D thread block dimension using simulation system matrix.	39
6.1	Benchmark computing integrals using OpenMP.	47
6.2	Benchmark computing integrals using CUDA.	48
6.3	Benchmark of assemble system step for different thread block dimensions.	50
6.4	Benchmark of solve system step for different Fortran solvers.	52
6.5	Benchmark of solve system step for different GPU linear solvers.	53
6.6	Benchmark for individual steps with OpenMP.	55
6.7	Benchmark for individual steps with CUDA.	56
6.8	Benchmark of overall timestep for both OpenMP and CUDA.	57
7.1	Visualization Tumbling Orbit.	60

7.2	Comparison of sedimentation velocity for single- and double-precision simulation.	61
7.3	Effect of fiber concentration on GMRES iterations.	62
7.4	Visualization Sedimenting Sphere.	64
7.5	Torus composition.	66
7.6	Effect of fiber concentration on torus break up time.	67
7.7	Effect of number of fibers on torus break up time.	67
7.8	Unmixed sphere	69
7.9	Mixed sphere	69

List of Tables

6.1	Benchmark Hardware Specification.	44
6.2	Warp Exection Efficiency of Numerical vs. Analytical Integration.	47
6.3	Atomic transactions of 2D vs. 3D thread block dimensions.	51
6.4	Overall speedup factor for 2000 fibers.	58

List of listings

4.1	Pseudocode for CUDA vector addition.	23
4.2	Pseudocode for CUDA matrix addition, illustrating 2D thread blocks.	25
5.1	Pseudocode for parallel algorithm on the host.	32
5.2	Pseudocode for the assemble system step with a 1D thread block.	33
5.3	Pseudocode for the updating velocities simulation step.	35
5.4	Pseudocode for the updating fibers simulation step.	36
6.1	Pseudocode for benchmark scheme.	45

Chapter 1

Introduction

Predicting the physical behavior of particles suspended in fluids is of great interest in a variety of different fields. The ability to accurately model and simulate different categories of particle suspensions allows their properties to be analyzed and optimized for a large number of applications. Examples include medical applications where the delivery and distribution of the active agent has to be modeled, waste management to efficiently extract waste from water and the paper industry which is trying to improve the characteristics of their material.

Even in the simplest flow cases, the flow of a particle suspension exhibits very complex and complicated dynamical behavior. The rheological properties of the suspension depend strongly on features such as the concentration of particles, particle shapes and particle interactions. In order to accurately capture the complex dynamics of a suspension using numerical simulations, a large number of particles are required in the simulation. Hence, the ability of the numerical algorithm to efficiently handle a large amount of particles is of crucial importance.

The work in this thesis focuses on increasing and optimizing the efficiency of a numerical algorithm of the simulation of sedimenting rigid and slender fibers in a viscous fluid. The fibers are modeled and simulated on a particle-level in a 3D free-space Stokes fluid. The mathematical model is based on a boundary integral formulation and a non-local slender body approximation by Tornberg and Gustavsson [29].

There are many numerical studies of fiber suspension and several different methods have been developed for both rigid and flexible fibers. One approach is the so-called beads-model, where the fibers are modeled as a set of connected spherical beads (e.g. [3][10][27][32]). The immersed boundary method discretizes

the fibers with Lagrangian markers and distributes the force onto a background grid which is then used to modify the fluid flow (e.g. [23][28]). The last approach is based on slender body theory which uses the large aspect ratio of the fibers to simplify the underlying model (e.g. [8][29][30]). A comprehensive review of the numerical studies of fiber suspension can be found in Guazzelli and Hinch [7].

In this thesis we develop a high performance GPU implementation of the numerical algorithm for simulating fiber suspension. By taking advantage of the massively parallel architecture of modern GPUs many more fibers can be simulated compared to the previous CPU based implementation of the algorithm. The major goal is to easily and efficiently perform simulations on a high-end desktop computer or workstation readily accessible by the researcher. This will allow the researcher to explore the huge problem space and simulate a large number of fibers in a short amount of time. Thus, it enhances the capacity to rapidly iterate and discover interesting test cases. These cases can then be used as a starting point for large scale simulations using computing clusters.

The most costly part of the numerical algorithm is to account for the interactions between all fibers in the system. For the calculations we chose the naive algorithm, where the interactions between every pair of fibers is computed. The development of such a naive algorithm is both easier and more cost effective than an alternative fast summation approach. A fast summation approach, like the fast multipole method, is both complex to implement and introduces a potentially large performance overhead. Therefore it might not result in the desired performance increase compared to the naive algorithm. Since the interactions between every pair of fibers is computed, the naive algorithm is more accurate. Our choice of consumer-grade GPU hardware limits the simulation to single precision, thus maintaining the accuracy of is very important. By using the naive algorithm on the GPU the goal is to strike a good balance.

This thesis is divided into the three major parts: *Previous Work*, *GPU Implementation* and *Results*. First we introduce the theoretical foundation of the numerical method in Chapter 2 and refer to the original paper by Tornberg and Gustavsson [29] for in-depth details. The numerical algorithm and the serial implementation on the CPU developed in the original paper [29] is presented in Chapter 3. Together these two chapters form the *Previous Work* part of the thesis. Based on this the *GPU Implementation* part follows. Chapter 4 briefly introduces general purpose computing on the GPU. Combining the previous work and GPU computing, we then describe our new parallel implementation of the rigid fibers

simulation on GPUs in Chapter 5. This represents the base for the efficiency and the performance improvements and therefore is the core of the thesis. Afterwards, the final *Results* are presented. The benchmarks in Chapter 6 illustrate the achieved performance increase of the parallel *GPU Implementation*. Finally, we performed and compared our simulation results to a few known experiments in Chapter 7.

Chapter 2

Theoretical Foundation

The introduction discussed different applications of rigid fiber simulations. It especially stressed the importance of being able to simulate a large number of fibers in order to generate the various patterns found in real world experiments.

In this chapter, we will present the theoretical foundation of the physics and the mathematical model the simulations are based on. This is required to be able to understand the numerical method used throughout the rest of the thesis. However, the model is only presented in a reduced summary, and for more details please refer to Tornberg and Gustavsson [29].

2.1 Stokes flow

We are interested in modeling the behavior of objects immersed in an incompressible fluid. We restrict ourselves to rigid bodies, which are slowly sedimenting in the fluid due to gravity. In the most general case, the flow can be modeled by the Navier-Stokes equations,

$$\begin{aligned} \rho \left(\frac{\delta \mathbf{u}}{\delta t} + (\mathbf{u} \cdot \nabla) \mathbf{u} \right) &= -\nabla p + \mu \nabla^2 \mathbf{u} + \mathbf{f} \\ \nabla \cdot \mathbf{u} &= 0. \end{aligned} \tag{2.1}$$

where $\mathbf{u}(x)$ denotes the velocity field, $p(x)$ the pressure field and $\mathbf{f}(x)$ the force acting on the fluid at the location $x = (x, y, z) \in \mathbb{R}^3$. The constant μ is the viscosity of the fluid. Solving these equations is quite challenging due to their time dependence and non-linearity. However, in this work we are only concerned with slowly moving objects and the Reynolds number $Re = \frac{\rho UL}{\mu}$ is assumed to be

very small. Given the constraint that $Re \ll 1$, the inertial and acceleration terms in the Navier-Stokes Eqns. (2.1) can be neglected and we arrive at the linear and time independent Stokes equations,

$$\begin{aligned}\nabla p - \mu \Delta \mathbf{u} &= \mathbf{f} && \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega.\end{aligned}\tag{2.2}$$

In addition to these equations we need two boundary conditions to solve the system. The first boundary condition accounts for the influence on the fluid from the presence of the objects. Hence, we force the fluid velocity at the boundary of the object to be equal to the velocity of the object itself. These no-slip conditions on the surface of the objects are defined as

$$\mathbf{u} = \mathbf{u}_\Gamma \quad \text{on } \Gamma,\tag{2.3}$$

where Γ denotes the union of all object surfaces and \mathbf{u}_Γ the corresponding surface velocity, thus constraining the fluid to have zero velocity relative to the object surfaces.

The second boundary condition models the fact that the fluid velocity far from the object is not affected by its presence. This is modeled by stating that the velocity field should be equal to a background velocity \mathbf{U}_0 at infinity

$$\mathbf{u} \rightarrow \mathbf{U}_0 \quad \text{as } \|\mathbf{u}\| \rightarrow \infty.\tag{2.4}$$

In our simulations this background velocity is always set to 0. Through the motion of the immersed objects and the no-slip boundary conditions defined in Eqn. (2.3) the dependency on time is reintroduced to the time independent Stokes Eqns. (2.2).

2.2 Boundary integral formulation

The Stokes equations are linear in both velocity and pressure, which allows them to be solved using a number of different methods for linear partial differential equations. The approach used in this thesis is the boundary integral method by Pozrikidis [24].

2.2.1 Fundamental solutions

We start from a boundary integral formulation of the Stokes equations, for which analytical solutions can be derived, the so-called fundamental solutions. One

2.2. BOUNDARY INTEGRAL FORMULATION

such solution is the Stokeslet. If \mathbf{f} in Eqn. (2.2) is given by a point force acting at \mathbf{y} , i.e. $\mathbf{f} = \mathbf{F} \cdot \delta(\mathbf{x} - \mathbf{y})$, where $\delta(\mathbf{x} - \mathbf{y})$ is the Dirac delta function, the velocity field

$$u_i(\mathbf{x}) = \frac{1}{8\pi\mu} S_{ij}(\mathbf{x}, \mathbf{y}) F_j \quad i, j = 1, 2, 3, \quad (2.5)$$

where the tensor product

$$S_{ij} F_j = S_{i1} F_1 + S_{i2} F_2 + S_{i3} F_3, \quad (2.6)$$

is a solution to the Stokes equations. The term S_{ij} is the Stokeslet and is given by

$$S_{ij}(\mathbf{x} - \mathbf{y}) = \frac{\delta_{ij}}{|\mathbf{x} - \mathbf{y}|} + \frac{(x_i - y_i)(x_j - y_j)}{|\mathbf{x} - \mathbf{y}|^3}. \quad (2.7)$$

Later, we will need higher order fundamental solutions to the Stokes Eqns. (2.2). These can be obtained by simply differentiating the Stokeslet. One example is the so-called doublet

$$D_{ij}(\mathbf{x} - \mathbf{y}) = \frac{1}{2} \Delta S_{ij}(\mathbf{x} - \mathbf{y}) = \frac{1}{8\pi\mu} \left(\frac{\mathbf{I}}{|\mathbf{x} - \mathbf{y}|^3} - \frac{3((x_i - y_i)(x_j - y_j))^2}{|\mathbf{x} - \mathbf{y}|^5} \right). \quad (2.8)$$

2.2.2 Formulation for objects immersed in a fluid

Using the fundamental solution from Eqn. (2.5) we can now model the influence of immersed objects on the Stokes flow. Assume that we have a total of M immersed rigid objects in the fluid. Each object m , for $m = 1, 2, \dots, M$ is centered at \mathbf{x}_c^m with an associated orthonormal basis \mathbf{t}^m and surface Γ^m . Given a rigid body motion and the no-slip boundary condition from Eqn. (2.3) we can model the velocity field $\mathbf{u}(\mathbf{x})$ at each surface point $\mathbf{x} \in \Gamma^m$ of the object as

$$\mathbf{u}(\mathbf{x}) = \mathbf{U}^m + \boldsymbol{\omega}^m \times (\mathbf{x} - \mathbf{x}_c^m), \quad (2.9)$$

where \mathbf{U}^m is the translational velocity of the object and $\boldsymbol{\omega}^m$ the rotational velocity. This setup is illustrated for $M = 2$ fibers in Fig. 2.1.

By combining this rigid body motion with the boundary integral formulation (Eqn. (2.5)) of the Stokes equations we can now relate the objects velocities with the force exerted by all other objects. For this we take advantage of the linearity of the Stokes equation and apply the superposition principle. The superposition principle states that the combined effect from all objects at a single point in the flow is simply the sum of all the individual effects caused by each object. For M

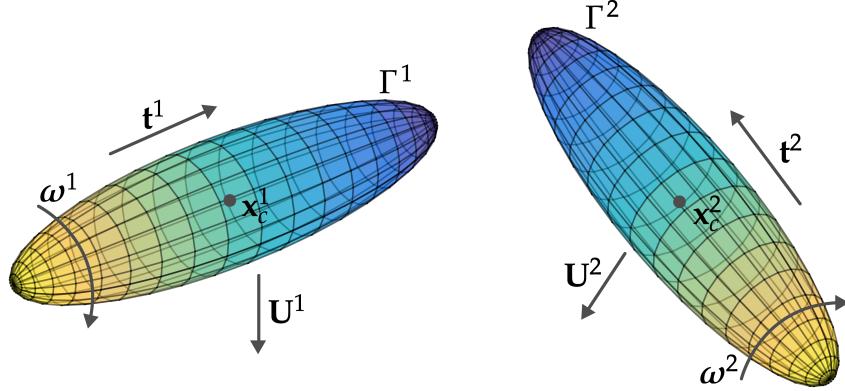


Figure 2.1: Two immersed objects in Stokes flow, using a rigid body motion.

fibers this yields the following relationship between the objects velocities and forces

$$\mathbf{U}_i^m + (\boldsymbol{\omega}^m \times (\mathbf{x} - \mathbf{x}_c^m))_i = \frac{1}{8\pi\mu} \sum_{l=1}^M \int_{\Gamma^l} S_{ij}(\mathbf{x}, \mathbf{y}) f_j^l(\mathbf{y}) dS_y \quad i, j = 1, 2, 3. \quad (2.10)$$

The resulting system has three unknowns, the translational and rotational velocities \mathbf{U}^m and $\boldsymbol{\omega}^m$ as well as the force distribution \mathbf{f}^m . Thus in order to be able to solve the system we add two constraints

$$\mathbf{F}_{\text{object}}^m = \int_{\Gamma^m} \mathbf{f}^m(\mathbf{y}) dS_y, \quad \mathbf{T}_{\text{object}}^m = \int_{\Gamma^m} (\mathbf{x} - \mathbf{x}_c^m) \times \mathbf{f}^m(\mathbf{y}) dS_y, \quad (2.11)$$

stating that the integrated force and torque over each object must be equal to the externally applied force and torque on the object. Given these additional constraints we are now able to solve Eqns. (2.10) and (2.11) for \mathbf{f}^m , \mathbf{U}^m and $\boldsymbol{\omega}^m$ for all fibers $m = 1, 2, \dots, M$. Using the resulting velocities we can then simply update the position and orthonormal basis of each object through time t by

$$\frac{d}{dt} \mathbf{x}_c^m = \mathbf{U}^m, \quad \frac{d}{dt} \mathbf{t}^m = \mathbf{t}^m \times \boldsymbol{\omega}^m. \quad (2.12)$$

Additionally, given the force distribution \mathbf{f}_m of each fiber we can compute the velocity field $\mathbf{u}_i(\mathbf{x})$ at any point \mathbf{x} of the Stokes flow. We again take advantage of the superposition principle and apply it to Eqn. (2.5) to express the velocity field as

$$\mathbf{u}_i(\mathbf{x}) = \frac{1}{8\pi\mu} \sum_{l=1}^M \int_{\Gamma^l} S_{ij}(\mathbf{x}, \mathbf{y}) f_j^l(\mathbf{y}) dS_y. \quad (2.13)$$

2.3. SLENDER FIBERS

2.3 Slender fibers

The formulation developed in the previous section is now specialized for a particular rigid suspension. In our case these are suspended slender fibers. Consider a straight, rigid body of length $2L$ and radius a , and let $\epsilon = a/2L$ denote a slenderness parameter. If $\epsilon \ll 1$ the body is referred to as a slender body (slender fiber) as illustrated in Fig. 2.2.

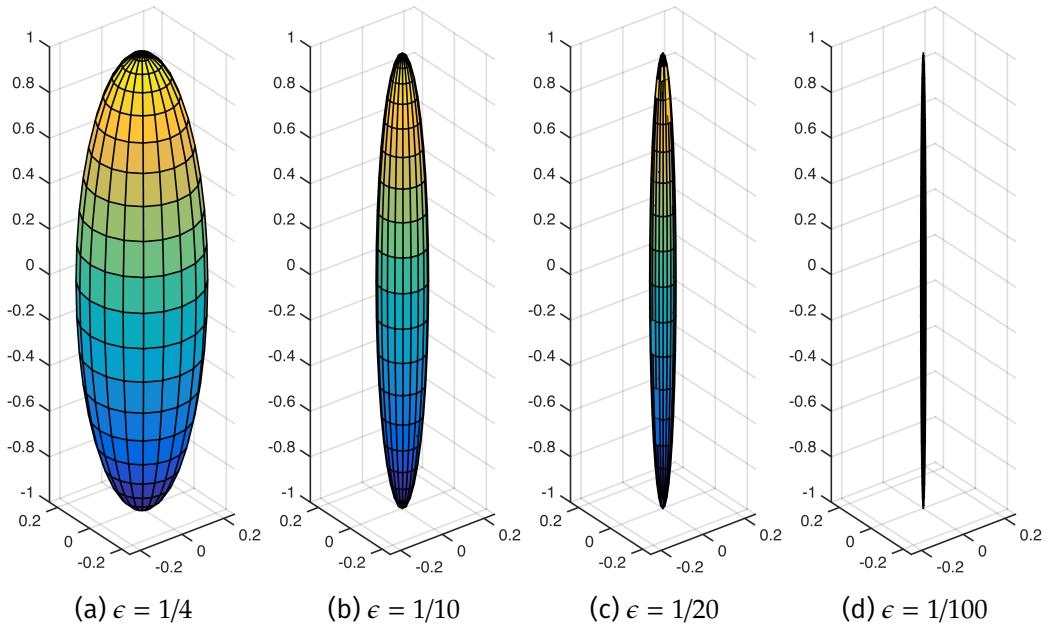


Figure 2.2: Illustration of the slenderness parameters, when ϵ becomes smaller the shape of the body asymptotically approaches that of an infinitesimal slender fiber.

The goal is to simulate many slender fibers, but using the 2D boundary integral formulation would be very expensive to solve numerically. As the aspect ratio $1/\epsilon$ of the fiber increases, the requirement of the number of quadrature points also increases in order to accurately resolve the flow field around the fibers as shown in Fig 2.2. However, for slender bodies a slender body approximation can be used instead.

The slender body approximation is derived from the boundary integral formulation of the Stokes Eqns. (2.13). Using asymptotic analysis the governing 2D equations are reduced to 1D equations along a centerline of the fiber. This could be achieved by matching the fluid velocity at a virtual boundary of a slender ellipsoid to the fiber centerline velocity [5]. The accuracy of this approximation is

in $O(\epsilon)$. This reduction in dimensionality from 2D boundary integral equations to 1D integral equations is crucial for the ability to include a large number of fibers in the simulations.

The slender body approximation yields a coupled system of 1D integral equations over two fundamental solutions of the Stokes equations, the Stokeslet (Eqn. (2.7)) and the Doublet (Eqn. (2.8)). The coupled system relates the forces exerted on the fibers to their velocities and captures the non-local interaction of the fiber with itself (as mediated by the fluid), as well as with any other structures with the fluid, such as other fibers or external boundaries. The system of integral equations is solved using the boundary integral method from Pozrikidis [24]. Details of the model are given by Tornberg and Gustavsson [29].

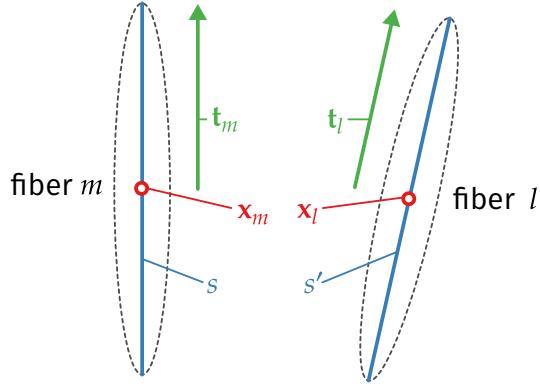


Figure 2.3: Slender fiber approximation.

Assume that we have M fibers immersed in the fluid. As shown in Fig. 2.3 each fiber is now defined by its centerline and parameterized by the arc-length, $s \in [-L, L]$. For fiber m the coordinates of the centerline are given by $\mathbf{x}_m(s, t) = \mathbf{x}_m(t) + s\mathbf{t}_m(t)$, where \mathbf{x}_m is the center point and \mathbf{t}_m the unit tangent vector of the fiber and $m = 1, 2, \dots, M$.

If the fluid exerts a force per unit length, \mathbf{f}_m on fiber m , the slender body approximation for the velocity of the centerline of fiber m is given by

$$\begin{aligned} d(\dot{\mathbf{x}}_m + s\dot{\mathbf{t}}_m) &= \left[d(\mathbf{I} + \mathbf{t}_m \mathbf{t}_m^\top) + 2(\mathbf{I} - \mathbf{t}_m \mathbf{t}_m^\top) \right] \mathbf{f}_m(s) \\ &\quad + (\mathbf{I} + \mathbf{t}_m \mathbf{t}_m^\top) \bar{\mathbf{K}}[\mathbf{f}_m](s) + \mathbf{V}_m(s). \end{aligned} \tag{2.14}$$

Here d is a geometry parameter

$$d = -\ln \epsilon^2 e, \tag{2.15}$$

2.3. SLENDER FIBERS

and $\bar{\mathbf{K}}[\mathbf{f}_m](s)$ is an integral operator given by

$$\bar{\mathbf{K}}[\mathbf{f}_m](s) = \int_{-1}^1 \frac{\mathbf{f}(s') - \mathbf{f}(s)}{|s' - s|} ds'. \quad (2.16)$$

The contribution to the velocity of fiber m from the other fibers in the system is accounted for in $\mathbf{V}_m(s)$ as

$$\mathbf{V}_m(s) = \sum_{l=1}^M \int_{-1}^1 \mathbf{G}(\mathbf{R}_{lm}(s, s')) \mathbf{f}_l(s') ds', \quad (2.17)$$

where $\mathbf{R}_{lm}(s, s') = \mathbf{x}_m + s\mathbf{t}_m - (\mathbf{x}_l + s'\mathbf{t}_l)$ is the distance between one point on fiber m and one point on fiber l as illustrated in Fig. 2.4.

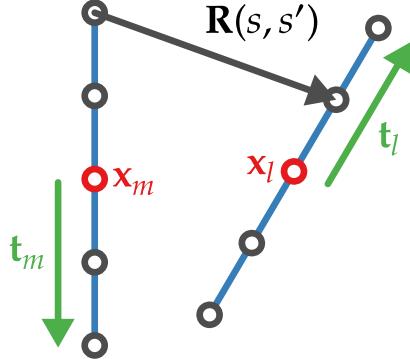


Figure 2.4: Two fibers with five quadrature points each. The interaction between two fibers is given by the term \mathbf{V}_m and depends solely on the distance between the two. The integral in Eqn. (2.17) can not be evaluated analytically and will be approximated by a quadrature rule.

The Green's function in this case is a linear combination of two fundamental solutions and reads

$$\mathbf{G}(\mathbf{R}) = \begin{cases} \mathbf{S}(\mathbf{R}) + \frac{r^2}{2} \mathbf{D}(\mathbf{R}) & \text{if } l \neq m \\ 0 & \text{if } l = m \end{cases} \quad (2.18)$$

with \mathbf{S} and \mathbf{D} as defined in Eqns. (2.7) and (2.8).

The equations are stated in non-dimensional form using the half-length of the fiber as a characteristic length, $L_c = L$. This will give use a characteristic velocity and time as

$$U_C = \frac{d\Delta\rho g V}{4\pi\mu_f L}, \quad T_C = \frac{2\pi\mu_f L^2}{d\Delta\rho g V}. \quad (2.19)$$

CHAPTER 2. THEORETICAL FOUNDATION

The unknowns in Eqns. (2.14) are the translational and rotational velocities, $\dot{\mathbf{x}}_m$ and $\dot{\mathbf{t}}_m$ and the force distribution along the fiber $\mathbf{f}_m(s)$. To close the formulation of Eqn. (2.14), we use the additional conditions stating that the integrated force and torque on each fiber must balance the external forces and torques applied to the fibers,

$$\mathbf{F}_m = \int_{-1}^1 \mathbf{f}_m(s) ds = \mathbf{F}_g, \quad \mathbf{T}_m = \int_{-1}^1 s(\mathbf{t}_m \times \mathbf{f}_m(s)) ds = 0. \quad (2.20)$$

For our simulation the external force is just gravity and the external torque is set to zero. Together with this the system is now closed and we can solve our rigid fiber simulation using Eqns. (2.14) and (2.20).

Chapter 3

Numerical algorithm and serial implementation

In the last chapter, we presented the theoretical foundation of the physics and mathematics involved in simulating rigid fibers. Based on the Stokes Equation, we introduced the framework of boundary integral methods to efficiently model the behavior of rigid fibers. Using this background we will now review the numerical approach used for the simulation.

We will separate the overall algorithm into four steps and discuss each step individually. Additionally, we will touch upon implementation details used in the original serial version. We close the chapter with a brief reflection of the performance characteristics of the serial implementation to guide the parallel implementation on the GPU.

3.1 Assemble system

In the previous chapter we obtained the final closed system and the Eqns. (2.14) and (2.20). By integrating Eqn. (2.14) from -1 to 1 we can derive two separate equations for $\dot{\mathbf{x}}_m$ and $\dot{\mathbf{t}}_m$. Using these two equations and some algebra we can obtain a system of equations for \mathbf{f}_m for all fibers from $m = 1, 2, \dots, M$, which only include computable quantities. For more details and an in depth discussion please refer to the original paper by Tornberg and Gustavsson [29].

In order to be able to solve these equations for \mathbf{f}_m we have to discretize them. For this we expand the force as a sum of $N + 1$ Legendre polynomials $P_n(s)$

$$\mathbf{f}_m = \frac{1}{2}\mathbf{F}_g + \sum_{n=1}^N a_m^n P_n(s), \quad (3.1)$$

where the coefficients a_m^n are vectors with three components. The number of Legendre polynomials N used for the force expansion is a parameter and is set to 5 in our simulation.

3.1.1 Closed linear system

Using the force expansion from Eqn. (3.1) in Eqn. (??) we get a closed linear system of equations for the coefficients a_m^n for $n = 1, \dots, N$ force expansions and $m = 1, \dots, M$ fibers.

For a 3-dimensional simulation, this results in a linear system of size $3MN \times 3MN$. The first step of the algorithm is thus to compute and assemble the linear system in memory. Writing the system in the standard form $\mathbf{A}\bar{\mathbf{a}} = \mathbf{b}$ gives the following structure for the matrix A and the right-hand side b

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \bar{A}_{12} & \cdots & \bar{A}_{1M} \\ \bar{A}_{21} & \mathbf{I} & \cdots & \bar{A}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{A}_{M1} & \bar{A}_{M2} & \cdots & \mathbf{I} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \vdots \\ \bar{b}_M \end{bmatrix}. \quad (3.2)$$

In this notation \bar{A}_{ml} describes the $3N \times 3N$ matrix encapsulating the contribution from the force coefficients on fiber l onto the force coefficients for fiber m .

3.1.2 Inner integral

For each \bar{A}_{ml} , a 3×3 matrix, Θ_{lm}^{kn} , where

$$\Theta_{lm}^{kn} = \int_{-1}^1 \left[\int_{-1}^1 \mathbf{G}(\mathbf{R}(s, s')) P_k(s') ds' \right] P_n(s) ds, \quad (3.3)$$

has to be evaluated for each force index $k, n = 1, 2, \dots, N$. G and R are defined as in Eqns. (2.18) and (??), respectively. This can be achieved by using a standard Gaussian quadrature approach for both the inner and outer integral. However,

3.2. SOLVE SYSTEM

as fibers get very close the number of quadrature points have to increase to accurately compute the fiber interactions. Unfortunately, doing so also increases the unknowns in the system, which in turn decreases the performance. For our simulation we use the same approach as the original paper. We divide each fiber into 8 subintervals and use a three-point gaussian quadrature on each interval. This results in a total of $3 \times 8 = 24$ quadrature points per fiber, which represents a good trade-off between accuracy and performance.

Another option is to find an analytical solution for the inner integral and only solve the outer integral numerically. We will not discuss the detailed derivation of the analytical solution, for an in-depth discussion please see [29]. In theory this approach allows for perfect accuracy for the inner integral, however in practice this is limited by the numerical precision of the simulation. The obtained formulas are recursive and thus sensitive to round off errors. To minimize the accumulation of round off errors, the original serial implementation uses a trick and switches the direction of the recursion, depending on how far apart the fibers are. This improves the practical accuracy and did not have a negative effect on performance.

Choosing between both options requires a careful examination of the accuracy and performance trade-off and is depend on the simulation setup. For the original serial implementation the combined numerical and analytical approach for evaluation proved to be the fastest and was thus chosen as the default. We will later explore how it applies to the new parallel GPU implementation.

3.2 Solve system

After having assembled the linear system $\mathbf{A}\bar{\mathbf{a}} = \mathbf{b}$ the next step to solve it. This can be done using standard linear equation solvers. This step is treated as a black box by the simulation, simply plugging in the matrix \mathbf{A} and the vector \mathbf{b} and getting back the solution vector $\bar{\mathbf{a}}$ representing the force coefficients.

The linear system can either be solved using a direct solver or an iterative method like GMRES. Unfortunately the matrix is neither symmetric nor sparse so we can't take advantage of specialized solvers for these cases [29]. As long as the fibers are sufficiently far apart and the matrix is not ill-conditioned, GMRES is able to solve the system in less than 10 iterations. This is the reason why the original serial implementation uses GMRES by default. How the different solvers perform on the GPU will be compared in Sec. 6.3.

3.3 Update velocities

The force coefficients obtained by solving the linear system can now be used to calculate the remaining unknowns. By using the force expansion in the two separate equations for $\dot{\mathbf{x}}_m$ and $\dot{\mathbf{t}}_m$. The required implementation is thus similar to the computations for the *Assemble System* step. The two integrals, again, can be evaluated either exclusively numerical or with the combined approach of solving the inner integral analytically.

3.4 Update fibers

The final step takes care of advancing the fibers forward in time. The equations do not impose any strict stability restrictions, so an explicit time-stepping scheme can be used. We also use the same second order multi-step method used in the original paper. The update for the position of the center coordinate \mathbf{x}_m is given by the following discretization in time

$$\frac{3\mathbf{x}_m^{i+1} - 4\mathbf{x}_m^i + \mathbf{x}_m^{i-1}}{2\Delta t} = (2\dot{\mathbf{x}}_m^i - \dot{\mathbf{x}}_m^{i-1}), \quad (3.4)$$

where the time step is denoted by Δt and superscripts denote the numerical approximation of $\mathbf{x}_m(t_i)$. In order to compute the next state this method requires both the previous and the current state. As there is no previous time step for t_0 , the first step \mathbf{x}_m^1 is computed by a simply first order forward Euler method. The calculation for the orientation vector \mathbf{t}_m is computed using the same discretization by replacing the translational velocity $\dot{\mathbf{x}}_m$ with the rotational velocity $\dot{\mathbf{t}}_m$. Additionally, we must renormalize the orientation vector so that it maintains its unit length.

At the end of this step the state of the fibers can optionally be written to an external file for post processing and visualization in other tools. After completing this step, the algorithm starts again from the top with the *Assemble System* step. This cycle repeats until a specified number of time steps have been executed.

3.5 Algorithm summary

The original paper implemented this algorithm using Fortran. The computation were all performed in double precision and executed using a single thread on the

3.5. ALGORITHM SUMMARY

CPU. In summary the four steps of the algorithm are at each timestep $t = t^n$ given the fiber configuration in terms of position, x_m , and orientation, t_m ,

1. Assemble System

Computes all interactions between fibers and yields the linear system $\mathbf{A}\bar{\mathbf{a}} = \mathbf{b}$

2. Solve System

Solves linear system, treated as black box

3. Update Velocities

Computes velocities $\dot{\mathbf{x}}$ and $\dot{\mathbf{t}}$

4. Update Fibers

Yields new fiber configuration at $t = t^{n+1}$

Empirical results show that the majority of the required computation time is spent on the 1. *Assemble System* and 3. *Update Velocities* steps. The time required for advancing the simulation state in step 4. *Update Fibers* is completely negligible. We will see later in Chapter 6 that the same holds true for the new parallel implementation.

Since we won't be writing our own implementation of linear solvers on the GPU we have only limited influence on the performance of the 2. *Solve System* step. We instead treat it as a black box and rely on the efficiency of pre-existing libraries. The only control we have is the choice which library and solver to use. For direct solvers the computation time only depends on the number of unknowns and is approximately constant throughout the simulation. The performance of iterative solvers on the other hand is highly depend on the condition number of the matrix and thus unpredictable. In line with the original paper we will both test a direct solver and iterative solvers to get a better understanding of their respective performance behavior.

Consequently, the most important step to optimize is 1. *Assemble System*. Fortunately it is well suited for parallelization. The fibers can be partitioned naturally across the compute units, where each unit is responsible for a subset of fibers. The focus of the optimization will thus lie on the *Assemble System*. As the required computation for the 3. *Update Velocities* step is similar it will also indirectly benefit from any optimizations. Additionally, we will also look at how the two different options for solving the integral in Eqn. 3.3 perform in the parallel

CHAPTER 3. NUMERICAL ALGORITHM AND SERIAL IMPLEMENTATION

environment. Therefore, the next goal is to parallelize each algorithm step on the GPU.

Chapter 4

GPU Programming

In the previous chapter the numerical algorithm and its serial implementation was presented. It discussed various implementation details which have to be considered to arrive at the most efficient implementation.

To further increase the efficiency modern GPUs offer a massively parallel architecture to accelerate many different applications. We will begin with a short introduction to general purpose computing on GPUs. Afterwards, different Application Programming Interfaces (API) for programming on GPUs are discussed regarding their advantages and disadvantages. The chosen API, CUDA, will be presented with its programming model at the end of this chapter.

4.1 General Purpose Computing on GPUs

In the beginning Graphics Processing Units were highly specialized pieces of hardware developed to exclusively improve the performance of real-time 3D graphics. However, in recent years GPUs have started to be used for running arbitrary code instead of being limited to graphics related computations. This allows for impressive performance increases across a wide range of different general purpose applications. A good overview of the evolution of GPU computing can be found in Owens et al. [21].

The deciding factor for the performance is how well the application can be parallelized to take advantage of the massively parallel architecture of GPUs. This massive parallelism has lead to potentially large performance advantages of GPUs over CPUs as illustrated in Fig. 4.1. It shows the year over year increase in the theoretical floating point operations per second (FLOP/s). The FLOP/s

number is calculated by combining information of the number of compute units, frequency and memory bandwidth for both the CPU and GPU models. This does not necessarily translate to direct real world performance increases but tries to visualize the potential GPUs have.

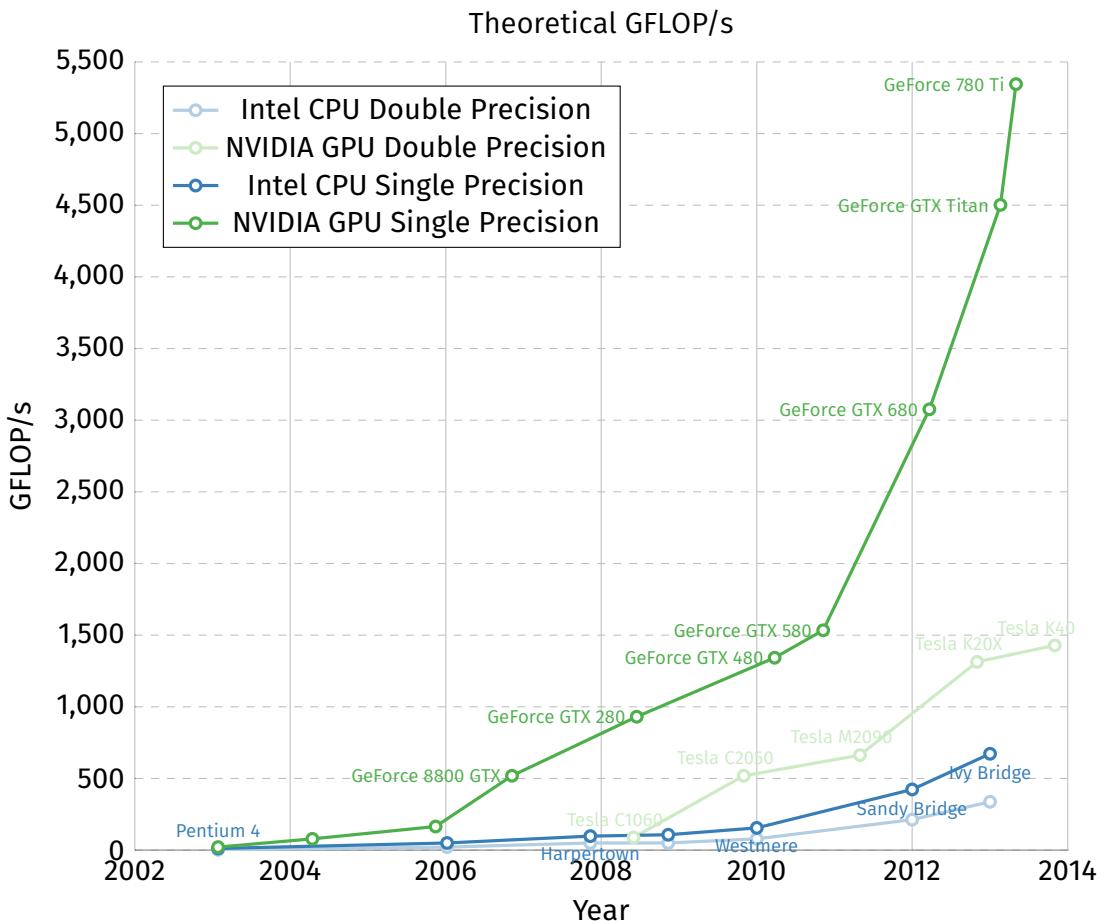


Figure 4.1: Year over year increase in theoretical floating-point operations per second for CPUs and GPUs [20].

The huge difference in performance between a GPU and a CPU mainly derives from the number of independent compute units. How a compute unit is exactly defined differs between architectures and vendors. On a CPU the number of compute units usually refers to the number of CPU cores. Each individual CPU cores is very fast, however, CPUs usually only have four, eight or maybe sixteen cores. In contrast to this, GPUs can have several hundreds of independent compute units. Each GPU compute unit can perform calculations in parallel and thus provides

4.2. CUDA VS. OPENCL

the opportunity to yield big performance improvements for high-throughput type computations. This is one reason why general purpose computing on GPUs was introduced to the world of supercomputers. Over time, a growing number of supercomputers started supplementing their computing power with GPUs and some even rely exclusively on GPUs for their computations.

In order to take advantage of these new massively parallel architectures new API had to be developed. The two proposed APIs are OpenCL and CUDA. OpenCL is an open and cross platform standard maintained by the Khronos Group [13]. The same group is also responsible for its graphics focused counterpart OpenGL. OpenCL is not exclusive to GPUs, but instead tries to be a general abstract layer for different parallel architectures. This allows OpenCL code to be run not only on GPUs but also on CPUs. CUDA on the other hand is developed by NVIDIA exclusively for their line of GPUs.

4.2 CUDA vs. OpenCL

Choosing between OpenCL and CUDA is the first decision to be made when starting to implement a new project on GPUs. The main advantage of OpenCL is the ability to run on many different devices. All major players in the computing space provide an implementation on top of their platforms. Both Intel and AMD provide the API for their CPU and both AMD and NVIDIA have drivers available for their GPUs. However, this advantage can also be a disadvantage as the achievable performance might suffer from the abstraction across all platforms. The OpenCL framework is potentially not optimized for a particular device specific architecture. CUDA on the contrary is in theory highly optimized to achieve the best possible performance on NVIDIA's GPUs. In practice the difference can possibly be mitigated by spending the extra time to fine-tune the OpenCL implementation to the hardware's specific needs. Another disadvantage of OpenCL is the potentially outdated and inconsistent driver support for the various devices. This is especially true for NVIDIA who seem to have stopped updating OpenCL, still only supporting OpenCL 1.1 which was released back in 2010. Their main focus is on pushing CUDA and updating it to support all the feature in their new GPUs.

For this thesis we chose to go with NVIDIA's CUDA framework mainly because of the available hardware both at the workstation computers and at the local computing cluster. Additionally, this project does not need the cross-platform capability as the main focus is on pure performance in a highly specialized setup

and simulation scenario. The application will not be widely distributed and only used for internal purposes.

4.3 CUDA programming model

The abbreviation CUDA stands for Compute Unified Device Architecture and was introduced by NVIDIA in 2006 as a general purpose parallel computing platform. It leverages the highly parallel architecture of modern NVIDIA GPUs to solve many different computational problems, which can lead to potentially large performance improvements compared to traditional CPUs.

The CUDA platform allows developers to use a variety of different options to program the GPU. The easiest way is to link to any CUDA-accelerated library and simply using the libraries interfaces from any software environment. For more advanced uses extensions to various programming languages exist like C/C++, Fortran and even managed languages like Java, Python and many more. This allows for easy and fast integration into any software environment the developer is comfortable with. Fig. 4.2 illustrated the different components of the overall CUDA platform.

The basic building blocks of the CUDA Programming Model from a development perspective are kernels. CUDA kernels are the equivalent of normal C functions. However, the major difference is that instead of being executed just once, kernels are executed in parallel by N different threads. These CUDA threads are distributed and run across the available compute units of the GPU. To illustrate how a very basic kernel invocation looks, Listing 4.1 shows a code sample for a very simple vector addition.

CUDA kernels It is important to remember that each kernel invocation is executed independently and no ordering is guaranteed. It is therefore essential to avoid any order-dependent operations or shared memory access. There are however, ways to allow for shared memory access which will be briefly touched upon later in the practical implementation of the simulation.

Thread hierarchy In order to efficiently distribute the different threads across the compute units of the GPU, CUDA defines a thread hierarchy. As discussed previously a GPU consists of many independent compute units. On NVIDIA GPUs these units are referred to as Streaming Multiprocessors (SMs). During execution

4.3. CUDA PROGRAMMING MODEL

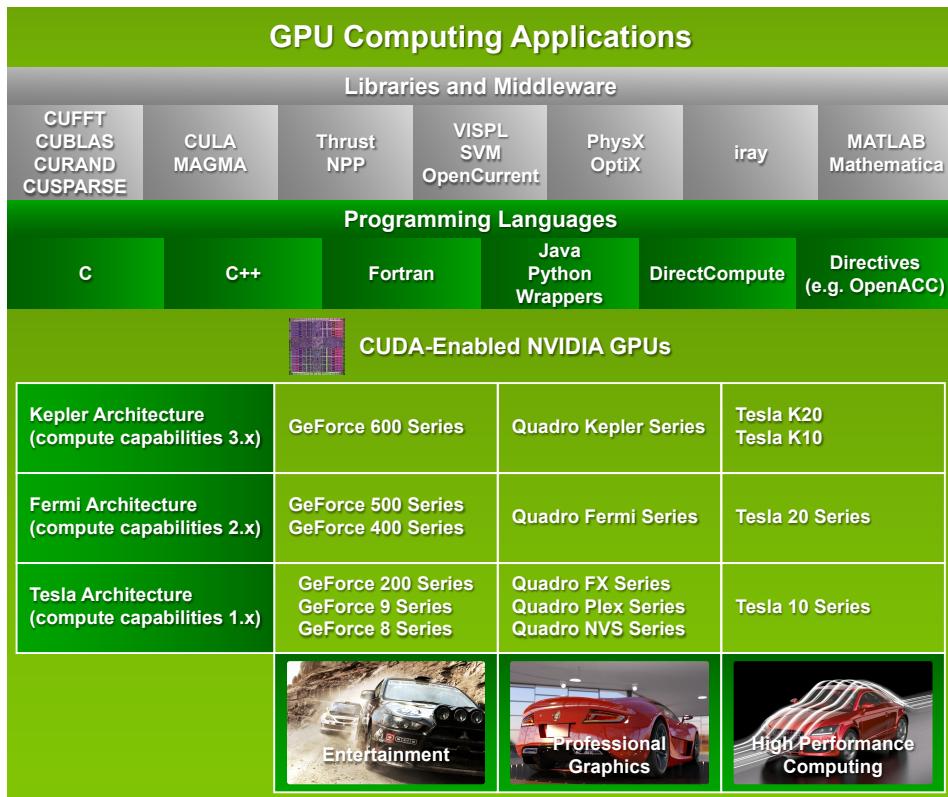


Figure 4.2: Overview of the CUDA platform [20].

```

1 // Kernel definition
2 __global__ void VecAdd(float *A, float *B, float *C)
3 {
4     int i = threadIdx.x;
5     C[i] = A[i] + B[i];
6 }
7
8 int main()
9 {
10    ...
11    // Kernel invocation with N threads
12    VecAdd<<1,N>>(A,B,C);
13    ...
14 }
```

Listing 4.1: Pseudocode for CUDA vector addition.

of the application each SM is tasked with running a distinct set of threads. In CUDA these sets of threads are called thread blocks. Each thread block is distributed to all the available SMs, which allows for automatic scalability depending on the number of SMs available as illustrated in Fig. 4.3. Thus the developer only has to divide the workload into appropriately sized blocks of threads and invoke the kernel.

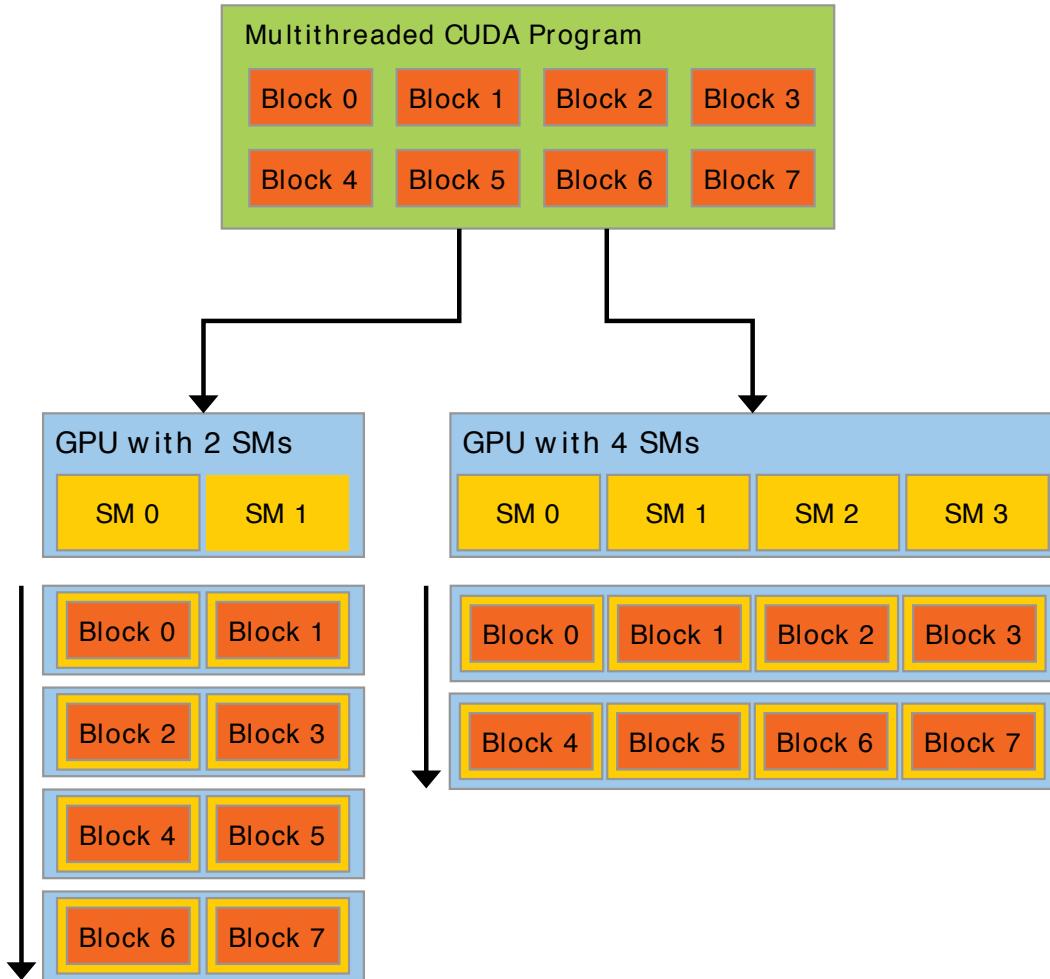


Figure 4.3: Automatic scaling of blocks across an arbitrary number of Streaming Multiprocessors [20].

How to choose the optimal size of a block in order to maximize the performance is not an easy question to answer and it is highly dependent on the particular task and implementation. In practice the size is often chosen by running benchmarks with various sizes to determine the optimal configuration.

4.3. CUDA PROGRAMMING MODEL

In order to make programming easier, CUDA blocks can be addressed using either a one-dimensional, two-dimensional, or three-dimensional thread index. For example in the case of a calculation involving matrices, it is more natural to think about parallelizing each element given by the row and column index instead of a single one-dimensional index. This is illustrated in the code sample in Listing 4.2

```
1 // Kernel definition
2 __global__ void MatAdd(float A[N][N], float B[N][N], float C[N][N])
3 {
4     int i = threadIdx.x;
5     int j = threadIdx.y;
6     C[i][j] = A[i][j] + B[i][j];
7 }
8
9 int main()
10 {
11     ...
12     // Kernel invocation with one block of N * N * 1 threads
13     int numBlocks = 1;
14     dim3 threadsPerBlock(N, N);
15     MatAdd<<numBlocks, threadsPerBlock>>(A,B,C);
16     ...
17 }
```

Listing 4.2: Pseudocode for CUDA matrix addition, illustrating 2D thread blocks.

Finally, as the resources of each Streaming Multiprocessor are limited, there is an upper bound of how many threads a block can contain. Currently this maximum number of threads is 1024. This means that the maximum size of matrices, that can be added with the two-dimensional thread block code sample in Listing 4.2 is $32 \times 32 = 1024$. To solve this problem CUDA introduces another layer above blocks called a grid. A grid organizes thread blocks again into either one, two, or three dimensions. The number of thread blocks in a grid is unlimited and thus solely dependent on the size of the workload. Fig. 4.4 shows an example configuration of a 2D grid with 2D blocks.

Memory hierarchy In addition to the *Thread Hierarchy* CUDA also implements a 3 layer memory hierarchy as illustrated in Fig. 4.5. Each level of this hierarchy differs in size, latency and scope. The sizes refers to amount of bytes that can be stored in the level and latency to the time it between requesting data and

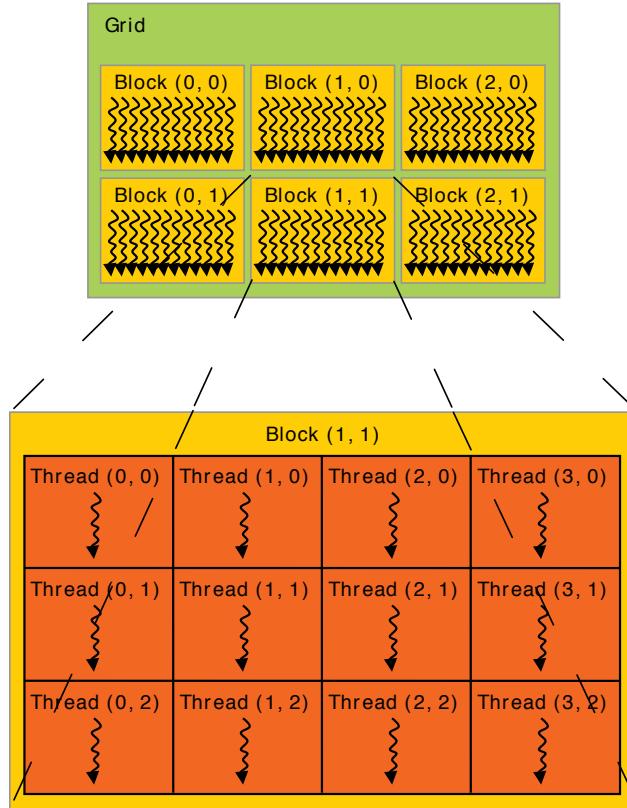


Figure 4.4: CUDA block addressing using 2D grid with 2D thread blocks [20].

being able to actually use it. Additionally each level is confined to a specific scope, restricting from where the data can be accessed.

The first level contains the fastest and smallest memory. This is a private memory where each thread has its own area and only this thread can read from and write to this area. How much space each thread has is determined by how many threads are executed on each SM. The more memory an individual thread requires the fewer threads can run in parallel on a SM. Optimizing this trade-off is important for achieving a good performance. On the newest GPUs each SM has $65536 * 32\text{-bit} = 256\text{KB}$ to divide among the threads.

The second level is called shared memory, or sometimes local memory. Although it is slower than the private memory, it has the advantage that all threads executing on an SM have simultaneous access to it. This allows different threads to cooperate and share work. Depending on the characteristics of the algorithm, this can save a lot of computing time. Therefore it is important to analyze whether

4.3. CUDA PROGRAMMING MODEL

shared memory can actually improve the performance. The size of this memory level on current GPUs is around 64KB to 96KB.

The third level contains the largest memory area which is called the global memory. This is the amount of memory used on the packaging of the GPUs to advertise the product. Nowadays it is usually 2GB or 4GB. It is accessible by all threads and stores all data relating to the current application. However, accessing it can be quite slow compared to the other levels of the memory hierarchy. Additionally, one has to be careful to avoid memory conflicts when accessing the same memory location from different threads at the same time. This is especially important since this can further reduce the performance. Taking everything into account, it can be said that it is an important optimization step to make access to global memory as efficient as possible. This can be achieved by intelligently packing data, or caching data in a lower level of the hierarchy.

Another option to avoid the slow the performance of the global memory is to make sure that the required data for each individual thread from global memory is small enough. The associated latency when reading data from global memory can be hidden when enough calculation are performed on this data. The GPU is then able to quickly switch to another set of threads and continue the computation while waiting on data from global memory.

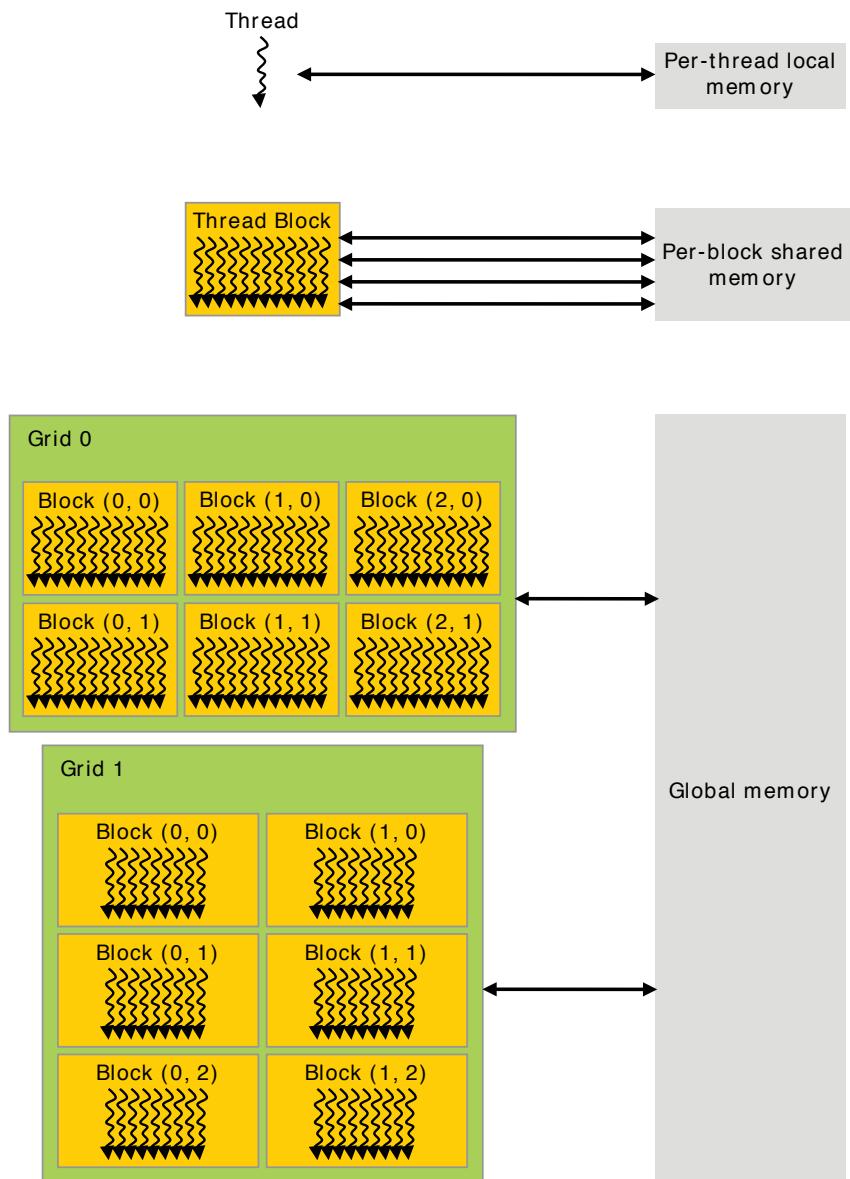


Figure 4.5: CUDA memory hierarchy [20].

Chapter 5

Parallel implementation

Based on the previous two chapters describing the serial Fortran implementation and GPU programming we will now look at the algorithm in more detail and show, how it was adapted to take advantage of multi-core architectures. The main focus of this thesis is the implementation of the simulation on a modern massively parallel NVIDIA GPUs using CUDA. In order to gain a better understanding of the achievable performance improvements we additionally back-ported the finished GPU code of the algorithm to multi-core CPUs using the OpenMP framework.

We will begin by introducing the software tools and libraries utilized. Afterwards, we will explain the practical implementation of the algorithm in CUDA. This is followed by the discussion of several potential optimization approaches to further improve the performance of the simulation. The chapter ends with a brief explanation of OpenMP and how the code was parallelized on the CPU.

5.1 Development environment

The rigid fiber simulation algorithm developed as part of this thesis is only loosely based on the original serial Fortran implementation. This was done to ensure a clean starting point and avoid difficulties in adapting the existing code for parallel execution, as it was never intended to be run across multiple cores. This also provided the opportunity to learn from the shortcomings of the old code to not only parallelize it but also improve the efficiency in general.

The development was done exclusively on a Linux workstation running Ubuntu as this will also be the exact same runtime environment used in the later experimental usage of the resulting application. The build system for compiling and

linking the final application was CMake [14]. It was chosen because it is a widely used open-source and cross-platform build system, which allows for easy integration of the various required libraries in a well documented and straightforward manner. Under the hood the build system uses NVIDIA's CUDA platform tools to compile the code. For this NVIDIA includes the CUDA compiler, called *nvcc*, capable of compiling C/C++ code together with the CUDA specific extensions.

In order to facilitate easier usage of the application both during development and later real-world usage a Python wrapper script is also available. The script completely automates the building process and dynamically customizes the application code to support three different modes of operation. The first is a simple *run* mode which takes the supplied parameters and executes the simulation. The second mode is *validate*, which allows for a fully automated way to test and validate different algorithm variations against a known correct simulation run. The *validate* mode automatically computes the error and makes debugging of changes easier. The last mode is *benchmark* which executes a particular simulation multiple times and collects and aggregates the timings for each simulation step as well as the total time.

5.2 Linear solvers

In addition to the CUDA platform, the application also requires support libraries for different linear solvers. During a simulation, a linear system of equations (Eqn. (3.2)) has to be solved in each timestep. This can be one of the most time consuming parts of the algorithm. Hence, the overall simulation time relies heavily on the ability to solve large and dense systems as fast as possible. Therefore, it is very important to weight all the available options carefully when choosing a particular linear solver algorithm or library.

In this work we have compared the performance of a direct solver to certain iterative solvers for solving the system. For the direct solver, the MAGMA library is used [15] and for the iterative we use the ViennaCL library [26]. Both libraries will be introduced below.

MAGMA / CuBLAS / OpenBLAS MAGMA [15] is a dense linear algebra library that provides features similar to standard LAPACK functions but for multicore architectures. It also has features to support hybrid algorithms to run code across multiple GPUs or CPUs at the same time. However, these features were not explored in this

5.3. KERNELS

thesis. Instead the focus was on a high performance single GPU implementation of a direct linear system solver.

MAGMA provides access to various high-level algebra routines, but the underlying math functions utilize the platform specific implementations of the BLAS levels. For CUDA this is implemented directly by NVIDIA in the form of the CuBLAS libraries [18]. Additionally, MAGMA tries to further improve their performance by combining GPU with CPU based algorithms. Thus a CPU based BLAS implementation is also needed. For this the OpenBLAS library [31] was chosen which is the most up-to-date and high performance library available apart from Intel's Math Kernel Library [9]. OpenBLAS takes full advantage of multicore systems and is also used for the comparison of our Fortran CPU implementation against the CUDA GPU implementation.

ViennaCL ViennaCL [26] is an open-source linear algebra library developed at the University of Vienna. The library provides an abstraction layer across many different parallelization methods in order to facilitate consistent and easy to use support for BLAS level 1-3 and iterative solvers. This unique feature allows the developer to easily switch between different APIs and architectures for parallelization. Currently, the library supports OpenMP, OpenCL and most importantly for this thesis - CUDA.

ViennaCL's focus is on solving sparse matrices with the implemented iterative solvers. However, it also has basic support for solving dense matrices using a variety of different iterative solvers. As the rigid fiber simulation exclusively relies on dense matrices this makes it an ideal candidate for benchmarking. For this thesis the BiCGStab as well as the GMRES iterative solvers were used and tested.

5.3 Kernels

The overall parallel algorithm is very similar to the serial version, however, each simulation step is separated into different kernels. Each kernel is invoked in a serial manner, this means CUDA guarantees that all data modified in a kernel is available before the next kernel is executed. These kernels are then distributed across the GPU. All calculations are done using single precision floating point numbers, as NVIDIA limits high performance double precision computation to their server class GPUs. The CUDA pseudocode for the algorithm is illustrated in Listing 5.1. It begins by parsing the simulation parameters and the initial fiber

configuration. After that, the required memory is allocated on the GPU. Finally a simple loop executes all time steps and the four main algorithm sub steps — 1. *Assemble System*, 2. *Solve System*, 3. *Update Velocities* and 4. *Update Fibers* — as described in Sec. 3.5 are run on the GPU.

```

1 int main()
2 {
3     // Parsing algorithm parameters and initial fiber positions
4     readParameters();
5     readFiberConfiguration();
6     allocateGPUMemory();
7     ...
8
9     for (int step = 0; step < max_timestep; step++)
10    {
11        AssembleSystem<<numBlocks, threadsPerBlock>>(...);
12        SolveSystem<<numBlocks, threadsPerBlock>>(...);
13        UpdateVelocities<<numBlocks, threadsPerBlock>>(...);
14        UpdateFibers<<numBlocks, threadsPerBlock>>(...);
15    }
16    ...
17 }
```

Listing 5.1: Pseudocode for parallel algorithm on the host.

The application requires two general configuration files as an input. The first file is referred to as the parameters file which contains the different configuration variables and constants used throughout the algorithm. These include, for example the number and size of the time steps as well as the number of force expansion terms and quadrature points. Additionally, this file is also used to configure the iterative solvers like specifying the number of restarts for GMRES or the solution tolerance for BiCGStab and GMRES.

Each of the parallelized sub steps are now discussed in more detail. The purpose of each kernel as well as the required input and outputs are described below.

Assemble System The *Assemble System* kernel is the most important step of the algorithm, as it is the most time consuming step. This was already discussed in Sec. 3.5 and it will also be shown in the benchmark tests presented in Chapter 6.

The goal of the assembly step is to build the matrix and right-hand side for the linear system of equations in the memory. As an example Listing 5.2 shows a pseudocode for the *Assemble System* step. In this example the code is

5.3. KERNELS

```

1 __global__ void AssembleSystem1D(
2     in float *positions,
3     in float *orientations,
4     out float *a_matrix,
5     out float *b_vector)
6 {
7     const int i = blockIdx.x * blockDim.x + threadIdx.x;
8
9     if (i >= NUMBER_OF_FIBERS) return;
10
11    for (int j = 0; j < NUMBER_OF_FIBERS ++j)
12    {
13        for (int force_index_j = 0;
14             force_index_j < NUMBER_OF_TERMS_IN_FORCE_EXPANSION;
15             ++force_index_j)
16        {
17            computeInnerIntegral(...);
18
19            for (int force_index_i = 0;
20                 force_index_i < NUMBER_OF_TERMS_IN_FORCE_EXPANSION;
21                 ++force_index_i)
22            {
23                // Only 1D thread block
24                // Each thread updates unique memory locations, thus
25                // no need for atomics
26                setMatrix(...)
27                setVector(...)
28            }
29        }
30    }
31 }
```

Listing 5.2: Pseudocode for the assemble system step with a 1D thread block.

parallelized using a one-dimensional thread block as described in Sec. 4.3. This means that the code is parallelized for each fiber and that each thread calculates the contributions to this fiber from all the other fibers. Looking at the matrix in Eqn. (3.2) each thread is thus responsible for $3 * N$ rows of the matrix, where N is the total number of terms in the force expansion from Eqn. (3.1). Alongside the one-dimensional thread implementation other options exist. They will be discussed in the optimization Sec. 5.4.

The kernel requires two inputs, the current position of each fiber and its orientation. Using these combined with the equations outlined in Chapter 2 and Chapter 3 the matrix and right-hand side elements are computed and used in the next step to solve the linear system they define.

Solve System As this thesis does not aim to implement generic linear solvers, this step is treated as a black box. In the *Assemble System* kernel two arrays containing the matrix and right-hand side of the linear system have been computed. These two arrays are now passed to the library that implements the linear solver. In case of the direct solver it is the MAGMA library and for the two tested iterative solvers, BiCGStab and GMRES, it is the ViennaCL library. Both libraries are able to directly use the already allocated memory regions and no additional allocations have to be performed. In order to conserve memory space the resulting solution vector is stored in the same memory location as the right-hand side and is passed on to the subsequent steps.

Update Velocities By solving the linear system of equations we obtain the coefficients in the force expansion as shown in Sec. 3.1.1. The *Update Velocities* kernel accumulates the exerted forces for all fibers and updates both the translational and the rotational velocities of the fibers simultaneously by computing the right-hand side in Eqns. (??) and (??). A two-dimensional thread block version of the kernel is illustrated in Listing 5.3. In a two-dimensional thread block version each individual kernel invocation is responsible for a single pair of fiber interactions. In case of two different threads calculate the interactions for the same fiber, it might happen that they try to write their results to the same memory location. When this happens, it could lead to incorrect results for the velocities. Fortunately CUDA provides atomic functions to circumvent this issue which will be discussed in more detail in Sec. 5.4.3.

Update Fibers The final simulation step takes care of advancing the position and orientation of the fibers in time. The pseudocode in Listing 5.4 implements the second-order multi-step method (Eqn. (3.4)) introduced in Sec. 3.4. As it will be shown during benchmarking in Chapter 6, the required time for this kernel is minuscule compared to the other steps. The kernel only scales linearly with the number of fibers and in addition has a perfectly aligned memory access resulting in close to optimal usage of the GPU hardware.

5.4 Optimizations

During the development of the parallel GPU simulation great care was taken to continuously optimize the code both on an algorithmic level as well as on

5.4. OPTIMIZATIONS

```
1 __global__ void UpdateVelocities2D(...)  
2 {  
3     const int i = blockIdx.x * blockDim.x + threadIdx.x;  
4     const int j = blockIdx.y * blockDim.y + threadIdx.y;  
5  
6     if (i >= NUMBER_OF_FIBERS) return;  
7     if (j >= NUMBER_OF_FIBERS) return;  
8     if (i==j) return;  
9  
10    for (int quadrature_index_i = 0;  
11         quadrature_index_i < TOTAL_NUMBER_OF_QUADRATURE_POINTS;  
12         ++quadrature_index_i)  
13    {  
14        for (int quadrature_index_j = 0;  
15             quadrature_index_j < TOTAL_NUMBER_OF_QUADRATURE_POINTS;  
16             ++quadrature_index_j)  
17        {  
18            force = computeForce(coefficients, ...)  
19            computeDeltaVelocities(force)  
20        }  
21    }  
22  
23    // 2D thread block  
24    // Each thread responsible for an interaction pair, thus  
25    // result is written to the same memory location  
26    // Using atomics to avoid conflicts  
27    atomicAdd(&(translational_velocities[i].x),  
28                delta_translational_velocity.x);  
29    atomicAdd(&(translational_velocities[i].y),  
30                delta_translational_velocity.y);  
31    atomicAdd(&(translational_velocities[i].z),  
32                delta_translational_velocity.z);  
33  
34    atomicAdd(&(rotational_velocities[i].x),  
35                delta_rotational_velocity.x);  
36    atomicAdd(&(rotational_velocities[i].y),  
37                delta_rotational_velocity.y);  
38    atomicAdd(&(rotational_velocities[i].z),  
39                delta_rotational_velocity.z);  
40 }
```

Listing 5.3: Pseudocode for the updating velocities simulation step.

```

1 __global__ void UpdateFibers(...)
2 {
3     int i = blockIdx.x * blockDim.x + threadIdx.x;
4
5     if (i >= NUMBER_OF_FIBERS) return;
6
7     next_positions[i] = 4/3 * current_positions[i]
8         - 1/3 * previous_positions[i]
9         + 2/3 * Timestep
10        * (2 * current_translational_velocities[i]
11           - previous_translational_velocities[i]));
12
13    next_orientations[i] = 4/3 * current_orientations[i]
14        - 1/3 * previous_orientations[i]
15        + 2/3 * Timestep
16        * (2 * current_rotational_velocities[i]
17           - current_rotational_velocities[i]));
18
19    normalize(next_orientations)
20 }

```

Listing 5.4: Pseudocode for the updating fibers simulation step.

an implementation level. Numerous small code-level optimizations have been performed based on the original serial code like precomputing as much data as possible, avoiding variable allocations or unnecessary copy operations in performance critical sections of the code.

Additionally, more advanced optimization of the code were made like rearranging calculations inside loops to avoid executing redundant calculations and consolidating multiple loops into one. Finally, techniques such as loop unrolling and faster math functions, like the reciprocal of the square root, where also tested and included.

Throughout the optimization phase a benchmark suite was run after each step. This ensures that optimizations were only included if they had a measurable impact on the overall performance of the simulation. Moreover, with this approach potential performance regressions could be identified early and be avoided.

Many optimizations performed during this process are applicable to both the CPU and GPU implementations, since they showed performance improvements for both. However, some optimizations and algorithm variations are uniquely suited to the GPU hardware. Below we will look into three different optimizations on the GPU in more detail. The performance results for each will then later be discussed in Chapter 6.

5.4. OPTIMIZATIONS

5.4.1 Numeric vs. analytic integration

In the original paper [29] it was observed that the analytical integration of the inner integral (Eqn. (3.3)) yielded a performance increase compared to purely numerical integration. Generally, using analytical integration should not only be preferred because of being faster but more importantly also because of being more accurate. However, for numerical precision reasons the actual implementation of the analytical integration can't achieve this theoretical level of accuracy. Especially for configurations with fibers that are very far apart the recursive implementation suffers from round-off errors and numerical instabilities. The steps taken to minimize these instabilities potentially affect the performance.

Based on these considerations and the fact that the computation of the integrals is a very performance critical part of the implementation, exploring the performance implications of both approaches on the GPU is of great interest. In contrast to the serial CPU implementation of the original paper, our implementation on the GPU is actually faster when solving both integrals numerically. We will discuss the reason and consequences of this in more detail in Sec. 6.2.1.

5.4.2 Shared memory

As described in Sec. 4.3, CUDA code is subject to a highly specialized memory hierarchy. Whereas traditional CPUs only have small caches and a large main memory pool, CUDA introduces the concept of a shared local memory space. The access time to this local memory is orders of magnitudes faster compared to accessing the global GPU memory. Additionally, local memory can be shared among the threads running on Streaming Multiprocessor and potentially save time by avoiding to constantly access the slow global memory. In order to test this, a shared memory version was implemented and tested for the *Assemble System* step, since it is the most performance critical kernel.

To understand the idea, imagine the 2D thread block implementation of the kernel. Then each thread block is responsible for many pairs of fiber interactions, e.g. fibers $[1, \dots, 8]$ each interacting with fibers $[9, \dots, 16]$. In total, these are $8 \times 8 = 64$ interactions. Each kernel invocation is responsible for one pair and has to load the position and orientation for the two interacting fibers. However, on closer inspection it is obvious that one does not need to load each fiber every time. As soon as fiber 1 has been loaded into shared memory it can be reused for

all the interactions with fibers 9 through 16 avoiding the unnecessary and slow access to global memory.

How this affects performance is not always easy to tell, as various factors can influence the result. One factor, for example, might be that the amount of necessary data is small enough to utilize memory caches and that repeated access to global memory can then automatically be avoided. Another factor is the performance characteristic of the kernel. Optimizing for shared memory usage only makes sense if on the one hand the kernel is memory bound, meaning that most of the execution time is spent waiting for data. If on the other hand, the kernel is compute bound the GPU is able to use time efficiently by performing pending computations while waiting for memory access. In this case forcing threads in a block to wait for shared data can actually decrease the overall performance of the kernel.

However, in general efficient exploitation of shared memory can be a huge advantage for parallel GPU implementations. This is especially true when comparing the performance to CPUs, as they do not have an equivalent fast and comparatively large memory space. Unfortunately, during our testing in Sec. 6.2.2, we did not see a performance advantage by using shared memory for our rigid fiber simulation.

5.4.3 Thread block dimension

There are many different factors that determine the performance of a particular GPU algorithm. This is especially true with regard to optimally taking advantage of the specific underlying GPU architecture, which change even between different models of graphics cards. How to best utilize the hardware depends on specific memory access patterns, avoiding too much register usage and choosing optimal settings for the thread block size. Each graphics card can have a different number of streaming multiprocessors with only limited resources and taking advantage of this can result in performance increases.

For this thesis we looked at the thread block dimension in particular and how choosing a different approach of parallelizing affects the performance. Doing so, the focus was on the *Assemble System* step, which is the most performance critical step. However, the results were then also transferred to the *Update Velocities* step.

The kernel invocations for one-dimensional and two-dimensional thread block approach are visualized in Fig. 5.1 using the matrix structure in Eqn. (3.2). In the one-dimensional case each invocation is responsible for computing contributions

5.4. OPTIMIZATIONS

to all sub matrices \bar{A}_{ml} . Thus in total this results in M kernel invocations, one for each fiber. For the two-dimensional case each invocation is only responsible for the contribution to a single sub matrix. Here the total number of invocations is $M \times M$. The three-dimensional case divides the calculation for the sub matrix further, one additional invocation per force index for a total of $M \times M \times N$ invocations.

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \bar{A}_{12} & \cdots & \bar{A}_{1M} \\ \bar{A}_{21} & \mathbf{I} & \cdots & \bar{A}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{A}_{M1} & \bar{A}_{M2} & \cdots & \mathbf{I} \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} \mathbf{I} & \bar{A}_{12} & \cdots & \bar{A}_{1M} \\ \bar{A}_{21} & \mathbf{I} & \cdots & \bar{A}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{A}_{M1} & \bar{A}_{M2} & \cdots & \mathbf{I} \end{bmatrix}$$

(a) 1D
(b) 2D

Figure 5.1: Illustration of 1D and 2D thread block dimension using simulation system matrix.

The most straight forward approach is to use the one-dimensional thread block which means that the algorithm is parallelized with regard to a single fiber. In this case a single kernel invocation is responsible for multiple rows of the resulting linear system matrix, as shown in Fig. 5.1a. Additionally, this approach does not have any memory access conflict as each kernel only writes to the memory location belonging to its unique fiber. The potential disadvantage for a one-dimensional thread block, however, is that the resulting code can be more resource intensive for each single kernel and potentially hinder the performance on each multiprocessor.

For a two-dimensional thread block each kernel invocation is responsible for a pair of interacting fibers as shown in Fig. 5.1b. While this decreases the necessary resources, it also requires atomic functions to handle the case when two threads try to update the right-hand side value for the same fiber. By using the so-called atomic functions CUDA streamlines and serializes the memory access. For example the `atomicAdd` function in Listing 5.3 for the *Update Velocities* step accumulates the contributions to a fiber from all the other fibers. This ensures that each update to the memory location is handled in a serial manner, guaranteeing the correct value in memory. Of course this implies a potential performance degradation, however, newer GPUs with new CUDA versions have been very well optimized to

only have a minimal impact. Benchmarking for the rigid fibers simulation shows that using a 2D thread block and the associated performance increases outweigh the potential performance hit of using atomics.

Three-dimensional thread blocks are the maximum allowed dimensions for a CUDA thread block. They are a further extension of the two-dimensional thread block, as now each kernel invocation is not responsible for the complete interaction but only the interaction resulting from a specific point of the force expansion. This results in even more potential memory conflicts and also increases the total number of thread blocks which have to be distributed.

It is not clear, how exactly the performance is affected by each decision for the thread block dimension. Only trial-and-error benchmarking combined with metrics from CUDA can find the optimal setting for the specific algorithm. We will show in Sec. 5.4.3, that for our GPU implementation the two-dimensional approach is the most efficient one.

5.5 OpenMP

The goal of this thesis is to implement a high performance rigid fiber simulation on the GPU using CUDA. In order to better understand to what degree this goal was achieved, it is crucial to be able to do a fair comparison. The original serial implementation is not an ideal candidate as it does differ in a number of ways. First of all, it is purely serial and doesn't take advantage of todays multicore CPUs. Furthermore, it was implemented in double precision and because of arbitrary restrictions NVIDIA places on its consumer GPUs, they are only suited for single precision. That is why the CUDA implementation is strictly single precision. Finally, the primary focus of the original Fortran implementation was a correctly implemented algorithm and not performance.

For these reasons and in order to have a fairer comparison of the performance differences between the GPU and CPU implementation, a completely new and rewritten parallel CPU code was also implemented. For the parallelization on the CPU the OpenMP library was chosen [2].

After having implemented a parallel algorithm for the GPU, the conversion to the OpenMP-based CPU implementation was relatively straightforward. All optimizations done for the GPU implementation were also applied to the new CPU code when applicable. In order to parallelize the BLAS functions required for the linear solver the already included OpenBLAS library was chosen. OpenBLAS [31] is

5.5. OPENMP

an open-source and highly optimized library and automatically parallelizes BLAS functions using pthreads across all available CPU cores. In contrast to the GPU implementation, the OpenMP version is only parallelized using a similar approach as the one-dimensional thread block on GPU as discussed in Sec. 5.4.3. This means that each core calculates the interactions for one fiber with all other fibers or put differently each core calculates the entire matrix row belonging to one fiber, as illustrated in Fig. 5.1a. As the underlying number of independent threads is much lower on CPUs, different parallelization dimensions did not have an impact during testing.

The end results of the practical implementation for this thesis is a highly optimized CUDA implementation for NVIDIA GPUs and additionally a parallelized and optimized Fortran OpenMP implementation for CPUs. In the next chapter we will present a number of performance metrics and comparisons between the GPU and CPU implementations.

Chapter 6

Benchmarks

The previous chapter introduced the parallel implementation of the numerical algorithm using NVIDIA’s CUDA framework. It presented a practical overview of the GPU implementation and discussed various strategies for optimizing the algorithm.

In this chapter we will present a number of benchmark tests. The main goal of the benchmarks is to measure the performance of the GPU implementation and compare it to the performance of the OpenMP-based CPU implementation. In addition to the comparison between the two different implementations, benchmark tests are also performed to investigate the different optimizations presented in the previous chapter.

6.1 Methodology

The methodology used for the benchmark suite is the same for all presented benchmarks. This ensures comparable results and the fairest comparison possible. A brief overview of the hardware and benchmark scheme used are described in the next sections.

6.1.1 Hardware

All benchmarks are run on the same workstation with specifications listed in Tab. 6.1. These hardware components can be considered a balanced system. This is done to come as close as possible to a fair comparison between the CPU and GPU, as comparing a high-end CPU against a low-end GPU would only be of limited value.

Workstation	
Processor	Intel Core i7 4770
Graphics	NVIDIA GTX 970 4GB
RAM	16GB DDR3
Operating System	Ubuntu Linux 12.04 LTS
CUDA Driver	CUDA 6.5.16

Table 6.1: Benchmark Hardware Specification.

The Intel Core i7 4770 processor is an 4-Core CPU based on Intel's Haswell Architecture. With its 8 parallel threads and 3.4 GHz it was one of the top-of-line processor from 2013/14 and is currently still available for around 300\$. The NVIDIA GTX 970 4GB is part of NVIDIA newest lineup of graphics cards based on the Maxwell Architecture. The main advantage of these new cards is the large memory of 4 GB allowing for larger simulations. With a current price of slightly above 300\$ it fills the middle price class for all Maxwell cards. Overall this can be considered to be a balanced system.

6.1.2 Benchmark scheme

In order to generate statistically significant and reproducible execution times, all benchmarks tests for both the CPU and GPU implementation uses exactly the same run-scheme.

For each benchmark run we start with a configuration of M fibers randomly distributed in space. In order to exclude configurations where fibers overlap and/or intersect, the random process is modified such that there always is a fixed minimum distance between two fibers. Furthermore, on order to ensure a fair comparison between e.g. different linear solvers, the average distance between fibers are kept fixed for all runs independently of the number of fibers. As will be shown in Section , the number of iterations of the iterative solvers depend on the average distance between the fibers. As the average distance become smaller, the condition number of the matrix increases and more iterations are needed for the iterative solver to converge. Thus by keeping the average distance fixed, we minimize the influence of the number of iterations in the iterative solver on the comparison.

Using the semi-random fiber configuration, the simulation is run for 10 time steps. To avoid remaining outliers in the configuration potentially causing varia-

6.1. METHODOLOGY

tions in the timings the first time step is excluded. The first time step is thus used as a simple warmup step for the simulation. So, the final average time for each run is measured using the last 9 time steps.

To measure how the execution time depends on the number of fibers in the simulation, all tests are run with varying number of fibers starting from 100 fibers up to 2000 fibers using an increment of 100.

In addition, every run is repeated a number of times where each run uses a new random fiber configuration. The final execution time is computed as the average time over the runs. The execution time is measured using the built-in CUDA timing events for the GPU implementation and the Fortran *SYSTEM_CLOCK* function is used for the CPU implementation.

To further improve the statistical significance of the result the benchmark scheme dynamically adjusts the number of runs measured for each test. If the relative standard error (RSE) of timings collected is too high after a minimum number of runs, more runs are scheduled. This repeats until the relative standard error falls below 20% and reliable timings have been obtained. The algorithm for collecting the benchmark results is illustrated using pseudocode in Listing.

```
1 for(int N = 100; N <= 2000; N += 100)
2 {
3     array timings;
4
5     int runs = 4;
6     while (runs <= MAX_RUNS)
7     {
8         for(int r = 0; r < runs; ++r)
9         {
10             generateRandomInitialFiberConfiguration();
11             run(10); // execute 10 timesteps
12             timings.add(getTiming());
13         }
14
15         rse = calculateRelativeStandardError();
16
17         if (rse <= 0.2) break;
18
19         iterations = timings.count();
20     }
21
22     reportTimings();
23 }
```

Listing 6.1: Pseudocode for benchmark scheme.

6.2 Optimizations

We now look at the performance results for the different optimizations previously outlined in Sec. ???. Where applicable, the results will be compared between the OpenMP and the CUDA version of the algorithm.

6.2.1 Numeric vs. Analytic Integration

The first benchmark tests the performance of the two different approaches to compute the inner integral from Eqn. (3.3). It can be solved either numerically or analytically. Fig. 6.1 illustrates the performance timings for the *Assemble System* step of the parallel OpenMP version. Inline with the observations made by the authors of the original serial implementation [29] the parallel version of the analytic integration is always faster than the numeric integration.

In contrast to the expected results of the OpenMP implementation the CUDA implementation shows a very different picture. When we look at the graph for the CUDA implementation in Fig. 6.2, we can observe that the results are reversed. The numerical integration outperforms the analytical integration by a large margin increasing with the number of fibers.

The reason for this result lies in the scheduling and execution of work on the GPU. All code inside a thread block (more precisely a warp) is always executed in lockstep. This means each line of code is executed for each thread in parallel. However, if the code encounters a branch in the execution path, like a simple *if* statement, the threads diverge. First, all threads for which the condition is true are executed while the other threads have to wait. Then all threads for which the condition is false are executed while the others are not used. Finally, after all divergent code paths have been executed the code continues in lockstep. This issue is referred to as Branch Divergence and should be avoided as much as possible when writing parallel GPU Code [19].

To confirm that Branch Divergence is the reason for the slowdown of the analytic integration on the GPU we look at the metrics of the CUDA profiler *nvprof*. The metric *Warp Execution Efficiency* shows the ratio of the average active threads per warp to the maximum number of threads per warp. The metrics for both the numerical and analytical integration of the Spherical fiber setup used in Sec. 7.3 can be seen in Tab. 6.2.

On the one hand the numerical integration is almost 100% efficient, meaning all warps execute in complete lockstep. The analytical integration on the other

6.2. OPTIMIZATIONS

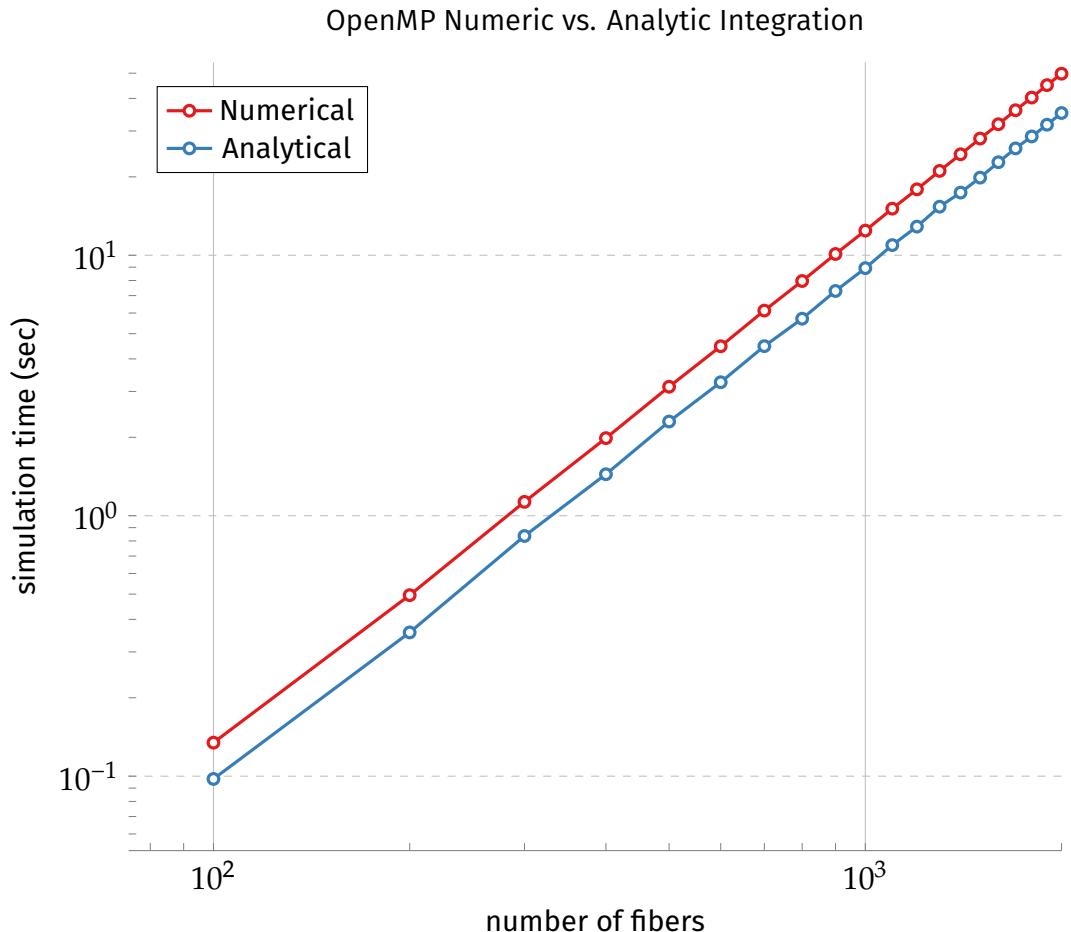


Figure 6.1: Benchmark computing integrals using OpenMP.

Algorithm	warp_execution_efficiency
Numerical	99.01%
Analytical	53.79%

Table 6.2: Warp Exection Efficiency of Numerical vs. Analytical Integration.

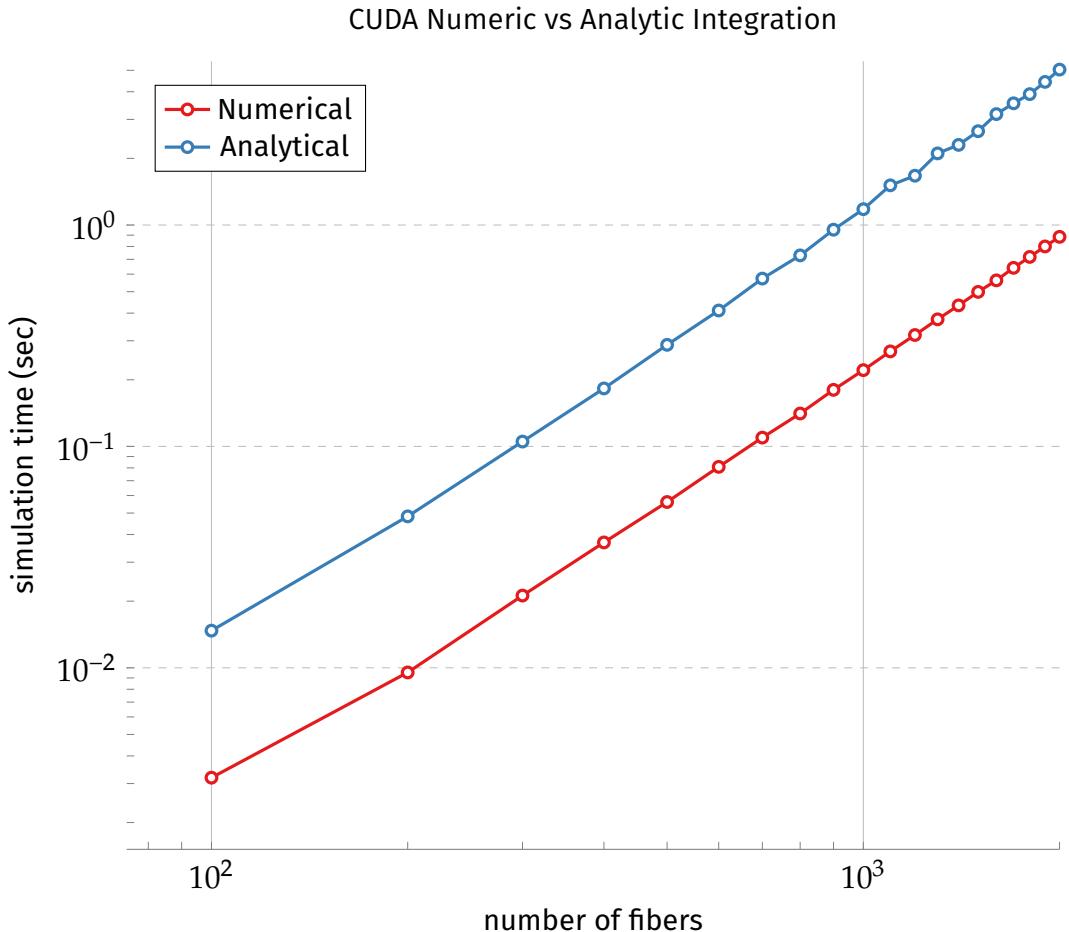


Figure 6.2: Benchmark computing integrals using CUDA.

hand is only 50% efficient, meaning that most of the time only half of the threads actually perform work while the other half is just waiting. This results in the observed performance difference. As already discussed in Sec. 5.4.1 the analytical solution of the inner integral potentially suffers from numerical instabilities. Closer inspection of the source code reveals that the steps taken to minimize these instabilities are responsible for the Branch Divergence. The steps involve a simple *if* statement, that switches between two code path depending on how far apart the two fibers are. Unfortunately, this workaround is unavoidable to ensure numeric stability and explains the decreased performance on the GPU.

6.2. OPTIMIZATIONS

6.2.2 Shared memory

The second optimization tried to use shared memory to reduce the amount of data that had to be transferred between the compute units and the global memory. For this each Streaming Multiprocessor has a small amount of locally shared memory. This memory can be accessed from all threads on the SM and in case data can be shared, it only has to be transferred slowly from global to shared memory once. Afterwards, it can be accessed from the faster local memory.

During testing and benchmarking the shared memory implementation of the *Assemble System* step as described in Sec. 5.4.2, no performance effect could be observed. Even though data can theoretically be shared among different threads it does not result in saved execution time.

The reason for this is that the *Assemble System* step is compute bound and not memory bound. This means that the time it takes to execute the computations, e.g. solving the integrals, takes substantially longer than reading and writing to global memory. This can be explained by looking at the 2D thread block version, where each kernel is responsible for a pair of fibers. Each kernel only needs to read 4 vectors from global memory, the position and orientation of both fibers. Assuming single precision this is a total of just $4 * 3 * 4\text{bytes} = 48\text{bytes}$ per kernel invocation.

While waiting for this data from global memory, CUDA is able to quickly switch between different sets of threads and continue the computation there. Thus the only waiting time occurs when the very first set of data has to be loaded. After the first data has arrived computations can be performed on it. While these longer running computations are happening, other threads can already issue data loading requests. Once the first computation finishes and the next set of threads is executed the data has already been fetched. As there was no advantage in using shared memory in the compute bound *Assemble System* step, we opted for the simpler implementation without it for all our simulations.

6.2.3 Thread block dimension

The next optimization looked at was the Thread Block Dimension on the GPU. Choosing the best option is a trade-off between the resources used and the overhead caused by an increased amount of memory writes to the same location. Writing to the same memory location from different threads would result in undefined behavior and avoiding this requires potentially slow atomic functions.

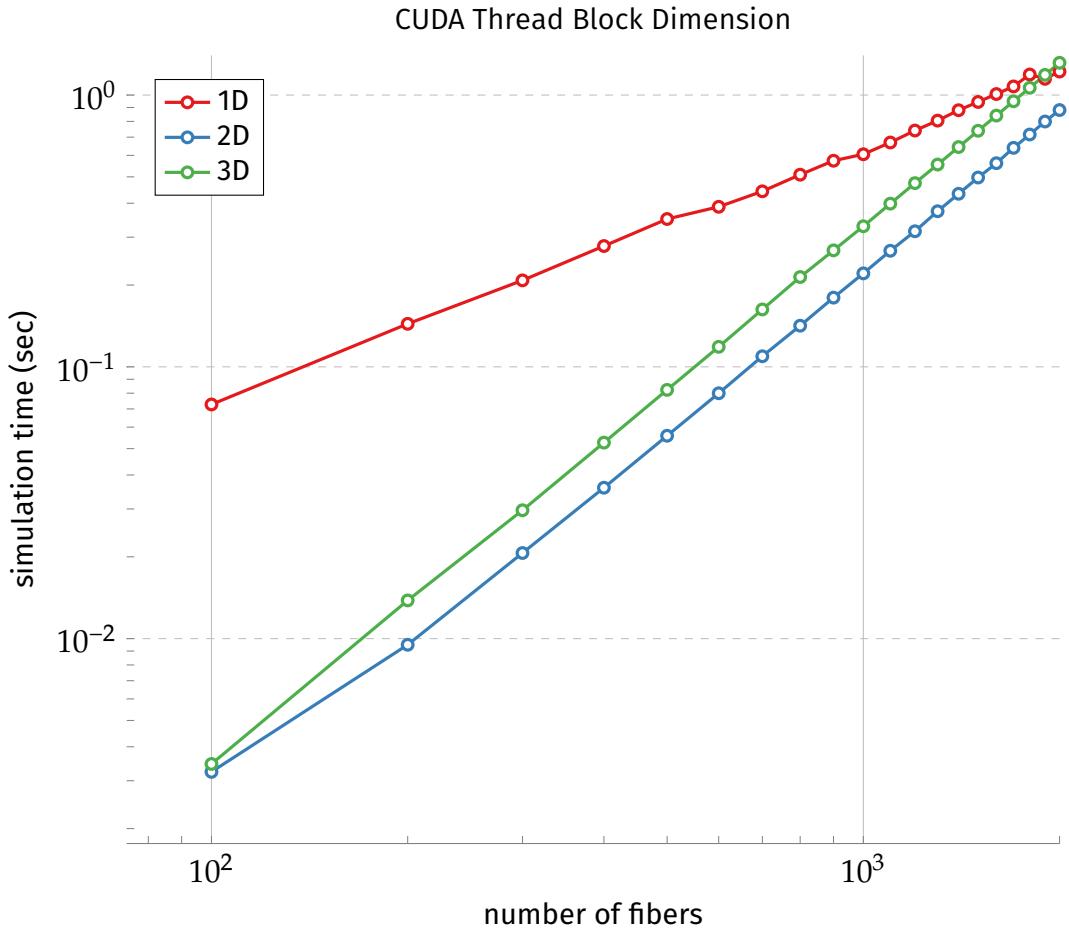


Figure 6.3: Benchmark of assemble system step for different thread block dimensions.

The results in Fig. 6.3 indicate that the best option for this particular GPU is a two-dimensional thread block. The three-dimensional thread block is always slower and the performance gap grows with the number of fibers. The reason for this performance gap is the increased usage of atomics in the three-dimensional case. The overall usage of atomic functions can be inspected with the NVIDIA profiler *nvprof* and the profiling metric *Atomic Transactions*. This metric simply counts the total number of atomic transactions performed by calling atomic functions. Tab. 6.3 lists the *Atomic Transactions* counts for both the two-dimensional and three-dimensional simulation for a simple sedimenting sphere setup. The required *Atomic Transactions* in the three-dimensional case are almost two times larger than for the two-dimensional case. These additional transactions incur a

6.3. LINEAR SOLVERS

performance penalty, because they serialize the access to memory and threads have to wait while other threads finish writing to memory.

Algorithm	atomic_transactions
2D	1,269,325
3D	2,350,670

Table 6.3: Atomic transactions of 2D vs. 3D thread block dimensions.

Fig. 6.3 also shows that the one-dimensional approach is slower than either the two-dimensional or three-dimensional approach. However, it appears to scale linearly whereas the other two scale exponentially. It can already be observed that the performance of the one-dimensional thread block becomes faster than the three-dimensional thread block for close to 2000 fibers. Unfortunately, the hardware of the workstation does not have enough memory to simulate more fibers, allowing the one-dimensional approach to overtake the two-dimensional approach. Thus at least for our simulation we always use the two-dimensional approach.

6.3 Linear solvers

Next we compare the performance for different linear solvers. We explored two different types of solvers. The first type is a direct solver of the linear equations which is both implemented for OpenMP and the CUDA implementation. The second type are iterative solvers. Unfortunately, we can't make use of the iterative solvers main advantage to efficiently solve sparse matrices as the system for the rigid fiber simulation is a dense matrix. The time required for solving the linear system can be a very large part of the overall runtime, depending choice of solver and the fiber configuration as discussed in Sec. 3.5. It is therefore very important to find the optimal solver for this particular solver to arrive at the best performing algorithm overall.

On the CPU side we used the direct solver provided by the OpenBLAS library, which is fully parallelized. For GMRES we used the single precision Fortran implementation from Frayssé et al. [4] which takes extensive advantage of the underlying BLAS functions parallelized by OpenBLAS. The original paper quotes that they

choose GMRES because it was faster than a direct solver for their tests [29]. The benchmark results for this thesis are illustrated in Fig. 6.4.

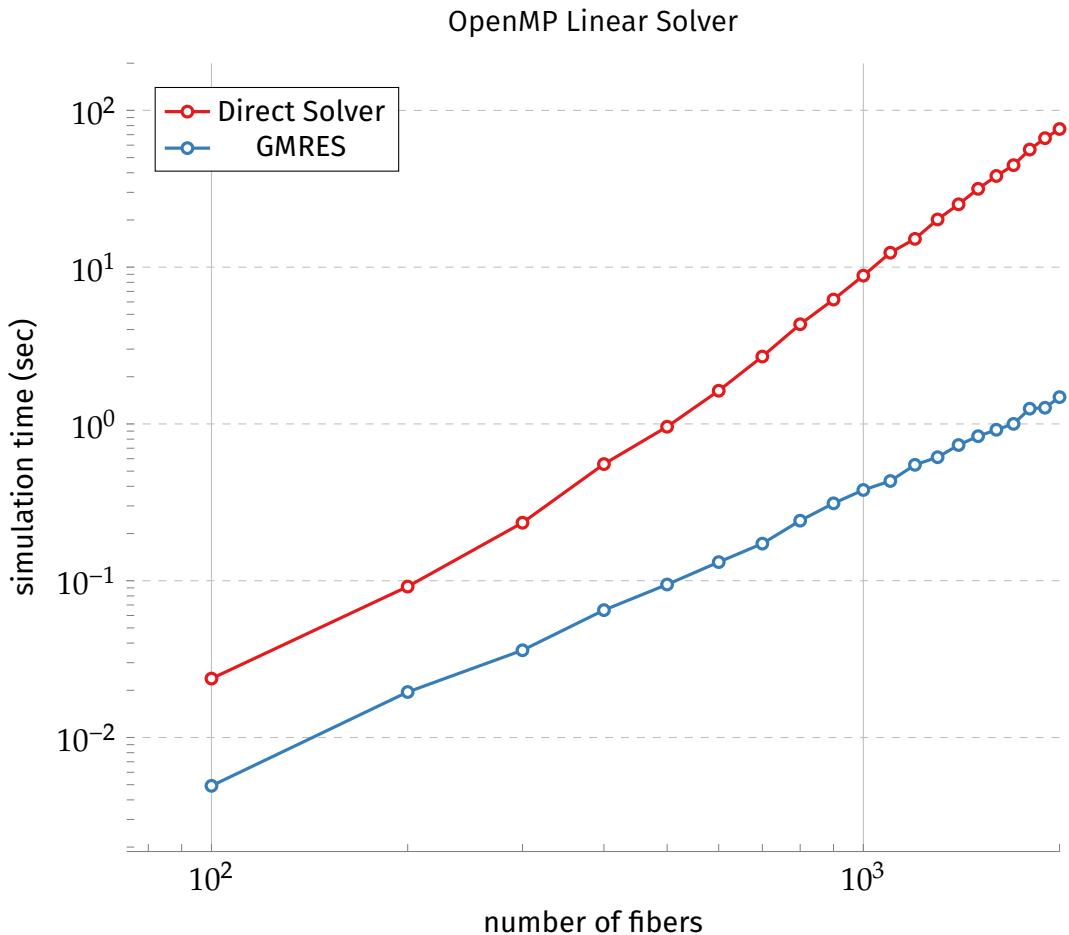


Figure 6.4: Benchmark of solve system step for different Fortran solvers.

We observe exactly the same, GMRES is faster by a wide margin as reported by the original authors [29]. From an advantage of just 40× for 1000 fibers this increases to 300× for the maximum number of 2000. GMRES clearly scales better with the number of fibers than the direct solver. The observed number of iterations remain almost the same regardless of the number of fibers. However, there is one caveat to this clear victory, which is the minimal and average distance between the fibers. As their concentration increases so do the required GMRES iterations. It is likely that for a very complex setup with many more time steps, that at one point fibers come very close together. This might cause GMRES to have problems finding the solution. This effect will be tested later in Sec. 7.2.

6.3. LINEAR SOLVERS

On the GPU side we used the direct solver provided by the MAGMA library. For the iterative solvers both GMRES as well as BiCGStab were compared as they can be easily exchanged and tested using the ViennaCL library. The benchmark results for the CUDA solvers are illustrated in Fig. 6.5.



Figure 6.5: Benchmark of solve system step for different GPU linear solvers.

The same fact that the iterative solvers are faster than the direct solver holds true for the GPU. However, compared to the CPU solvers the difference between the performance of the direct solver and the iterative solvers is not as large. At close to 2000 fibers the iterative solvers are only about 25x faster than the direct solver. Looking at the difference between the two iterative solvers, BiCGStab and GMRES they perform almost exactly the same. Any small differences can be attributed to small measuring uncertainties. Regardless of these clear results,

it is always important to keep in mind, that this particular performance ratio only holds true for the specific fiber concentration which was benchmarked. For other concentrations the iterative solvers might need more iterations to find the solution and might even perform worse than the linear solver.

6.4 Individual steps

The next benchmark explores the relative time taken by each step of the algorithm, which were described in Sec. 3.5 and Sec. 5.3. This gives valuable insight into the time allocation of the overall run time. Thereby helping to figure out which steps are the best ones to optimize. The results for the OpenMP implementation can be seen in Fig. 6.6 and for CUDA in Fig. 6.7. On the CPU the *Assemble System* and *Update Velocities* steps solve the inner integral in Eqn. (3.3) analytically. On the GPU we use a purely numerical approach instead. This is done to use the fastest algorithm variation on both the CPU and GPU as we showed in Sec. 6.2.1.

The results for the OpenMP implementation show that for 2000 fibers the *Assemble System* step is responsible for 78% and the *Update Velocities* step for 21% of the overall time. The two other steps *Solve System* and *Update Fibers* barely register with just 1%. However, as discussed in the previous Sec. 6.3 this only holds true for sufficiently low concentrations of fibers.

For the CUDA implementation the *Assemble System* step remains the largest block with 72% of the time. However, now the *Solve System* step takes 20%. *Update Velocites* drops to just 7% and *Update Fibers* is negligible with below 1%.

This difference in the relative time between the steps can be attributed to the comparatively slow GMRES implementation on the GPU. In absolute terms for 2000 fibers both the CPU and GPU take 0.25 seconds for solving the system. But, when transferring the same relative distribution from the CPU to GPU this makes the GPU GMRES implementation 40× too slow. The other steps all perform relatively better.

The reason for this most likely lies in the non-optimized code for the GMRES solver in ViennaCL. The authors openly state that the major goal for their library is easy of use and not pure performance [25]. The parallel BLAS functions from OpenBLAS on the other hand have been highly optimized and tested for a long time.

This point is reinforced if we look at the same numbers for the linear solvers. Here the absolute time for *Solve System* step on the GPU is just 7 seconds

6.5. GPU VS. CPU

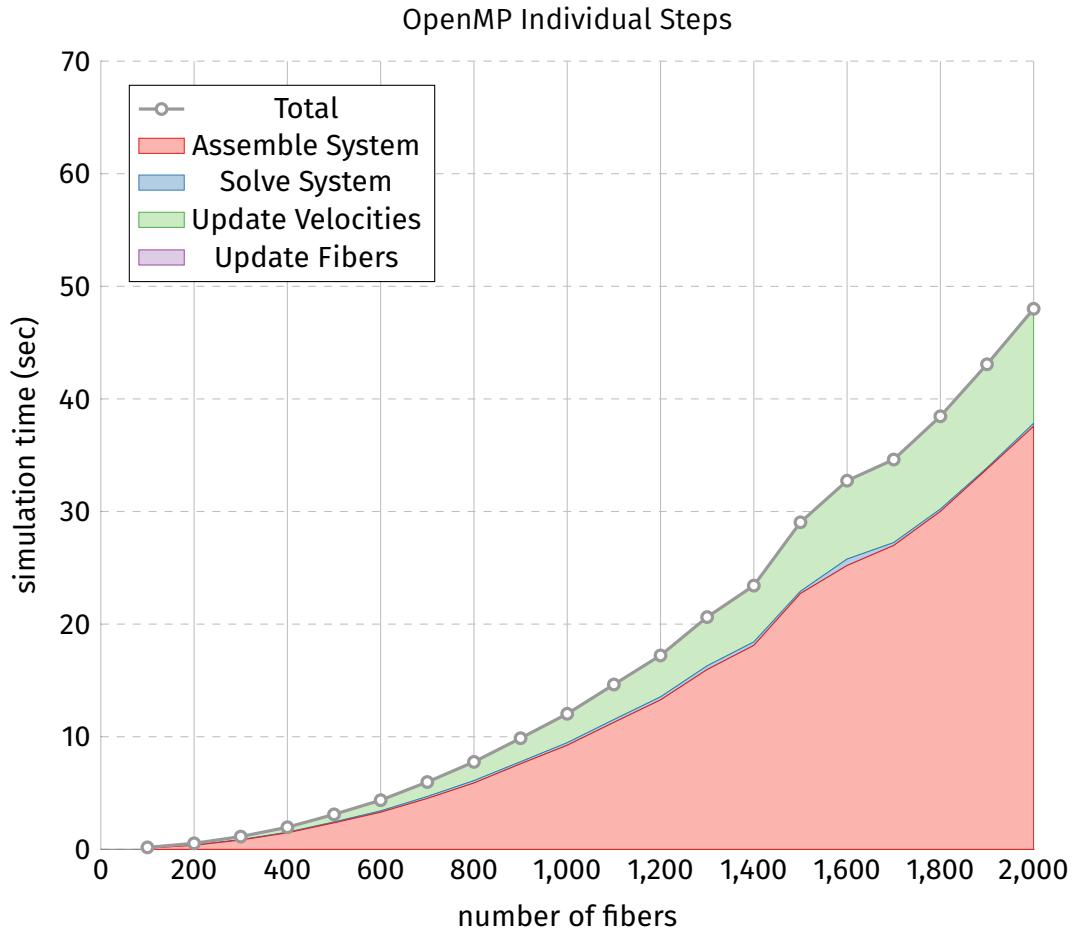


Figure 6.6: Benchmark for individual steps with OpenMP.

compared to the 76 seconds on the CPU. Again transferring the relative distribution from the CPU to the GPU yields just a small factor of 1.4x. This illustrates that the highly optimized code of the MAGMA library performs roughly on the same level as the optimized code from OpenBLAS.

6.5 GPU vs. CPU

The final benchmark compares the CPU and GPU performance. How to make a fair comparison of the simulation performance between the CPU and GPU is a hotly debated topic in the research literature [16][6]. The underlying architectures of the two approaches are completely different and thus hard to compare. Some try to extrapolate relative performance from the underlying FLOPs by taking processor

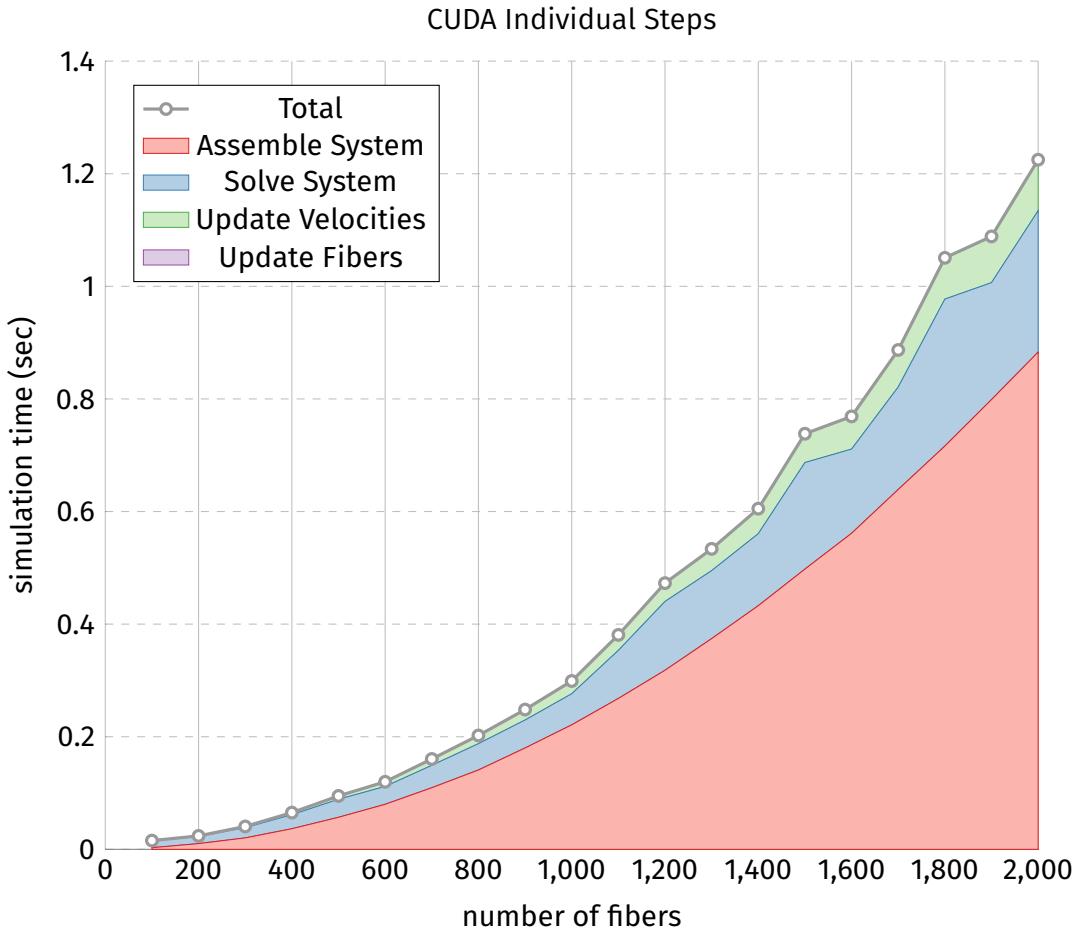


Figure 6.7: Benchmark for individual steps with CUDA.

count, frequency and memory bandwidth into account [16]. However, due to intricate hardware details this approach is also not applicable to all scenarios. Thus in the super computing community metrics like performance-per-dollar or even performance-per-watt have become the main focus [12].

Exploring this question in more detail is out of the scope of this thesis. Nevertheless we try to make a best effort to do fair comparison of CPU and GPU performance. In order to come as close as possible given these complexities and constraints, we used both a current CPU and GPU which can be considered a balanced system at the time of writing. Additionally, we implemented the parallel OpenMP version. It is directly based on the parallel CUDA version with the sole purpose to have as few difference between the two implementations as possible. Furthermore, the final benchmark for each uses the fastest possible algorithm

6.5. GPU VS. CPU

variant as determined by all previously performed benchmarks to compare the best variant on each architecture.

The results for average time required to take a single time step is illustrated in Fig. 6.8. For OpenMP the algorithm uses the analytical integration of the inner integral. For CUDA, we used the numerical integration and the thread block dimension was chosen to be 2-dimensional.

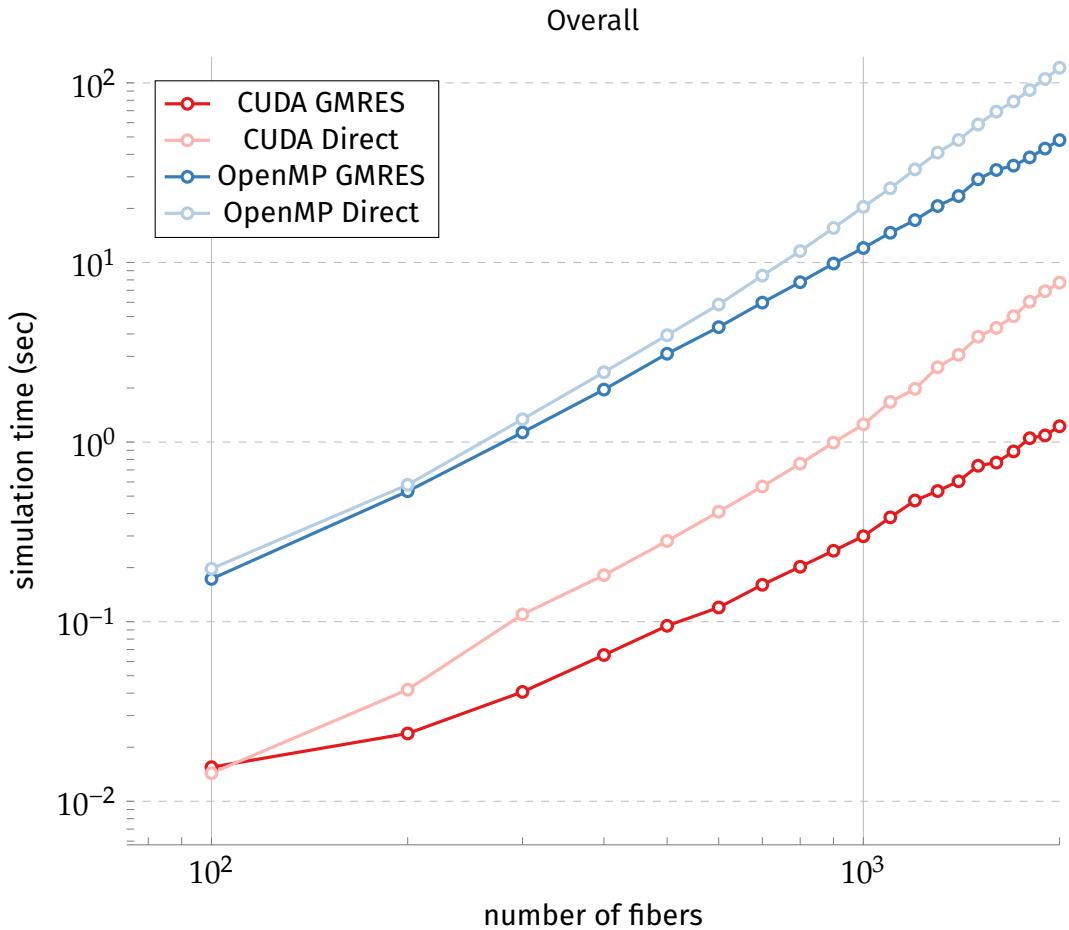


Figure 6.8: Benchmark of overall timestep for both OpenMP and CUDA.

The required simulation time on the GPU outperforms the CPU by a wide margin. The GPU version is faster for any number of fibers. The speedup factors for the presented simulation variations are listed in Tab 6.4. CUDA maintains a relative performance of 40x. The only advantage the CPU version has is the potentially larger memory as 4GB on the GPU limits the number of fibers to roughly 2000, so for more fibers OpenMP is the only option.

M = 2000	OpenMP Direct	OpenMP GMRES	CUDA Direct	CUDA GMRES
OpenMP Direct	1x	—	—	—
OpenMP GMRES	3x	1x	—	—
CUDA Direct	16x	6x	1x	—
CUDA GMRES	99x	39x	6x	1x

Table 6.4: Overall speedup factor for 2000 fibers.

We acknowledge that these numbers and performance increases are not necessarily fair. A different CPU and GPU combination from the one used in this thesis might perform differently. However, the relative performance should stay roughly the same. In the end, the only thing that really matters for the researcher working with rigid fibers is the time it takes to simulate large systems on the available workstation. There is no need to wait for computing time at a large computing cluster. Instead simulation can be run simply on a desktop computer allowing to rapidly iterate on the tests. The observed performance increase of $40\times$ is a difference between a whole day of waiting for the simulation results and a quick 30 minute result during the day. The saved time also translates directly into the ability to simulate many more fibers than ever before. Before the largest simulation possible in a reasonable time frame where around 500 to 800 fibers, in contrast one single time step for 2000 takes at most 8 seconds on the GPU. Therefore the implementation on the GPU and the optimizations of the simulation code is of great value to the research of rigid fiber simulations.

Chapter 7

Experiments

The previous chapter showed how much faster the CUDA implementation is compared to the same algorithm implemented on the CPU. This significant increase in performance allows for new experiments.

We will now perform a number of experiments to explore the numerical precision and validate the results against prior research. We begin by looking at tumbling orbits, a very delicate experiment requiring great precision. This is followed by a quick look at the effect of the fiber concentration on the required GMRES iterations to settle on the solution. Next we reproduce a very interesting physical phenomenon of letting a sphere of fibers sediment and validate our results against similar experiment by others. Furthermore, we will have a brief exploration of the effects of the number of fibers and the concentration of fibers on the sphere simulation. All experiments were run with the purely numerical, 2-dimensional CUDA algorithm implemented in the thesis. The systems were solved using the direct solver from MAGMA in order to avoid variations in the run time.

7.1 Tumbling orbits

The first example was used to verify that the single precision GPU version is able to replicate the result of the original double precision code. For this we looked at a very simple problem where a small number of fibers are set up with perfectly symmetrical positions and orientations. All fibers are evenly distributed on a circle and aligned vertically with gravity. During the simulation the fibers begin to rotate from their vertical orientation towards the horizontal position. Afterwards,

they continue rotating back into the vertical position. This motion is referred to as tumbling orbits and as long as there are no disturbances or numerical precision issues it repeat forever.

This simple but very interesting problem has also been simulated by Gustavsson and Tornberg [8]. Additionally an even more simplified version with only two fibers was studied both numerically and experimentally by Jung et al. [11]. This example is thus ideally suited to test and verify the numerical precision of the GPU simulation. A visualization of the GPU simulation of 16 fibers evenly distributed around a circle with a radius of 0.55, is shown in Fig. 7.1.

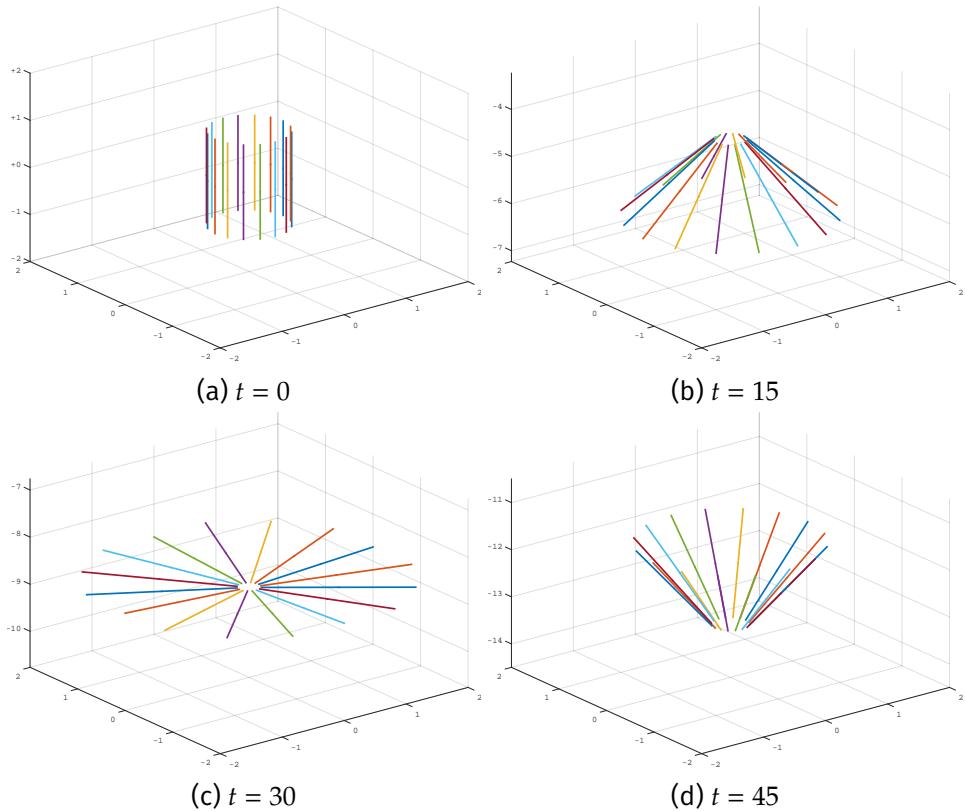


Figure 7.1: Visualization Tumbling Orbits.

In order to analyze the periodic movement of the fibers we now look at the sedimentation velocity caused by gravity. In the initial setup the fibers are aligned vertically and are dropping with the maximum velocity. As they rotate into the horizontal orientation the velocity is decreasing and reaches its minimum once the fibers are perpendicular to the direction of gravity. Afterwards on their way back to vertical orientation the velocity increases again.

7.2. FIBER CONCENTRATION EFFECT ON GMRES ITERATIONS

Fig. 7.2 shows a graph of the sedimentation velocity of a single fiber over time for both the single precision GPU code and the original double precision Fortran code. As the same force acts on all fibers and they perform the same motion they all have the same sedimentation velocity. For this particular setup the maximum velocity is 3.8 and the minimum velocity is 2.2. The graphs clearly shows the periodical rotation the fiber perform. This result perfectly captures the expected result obtained from prior simulation and experiments.

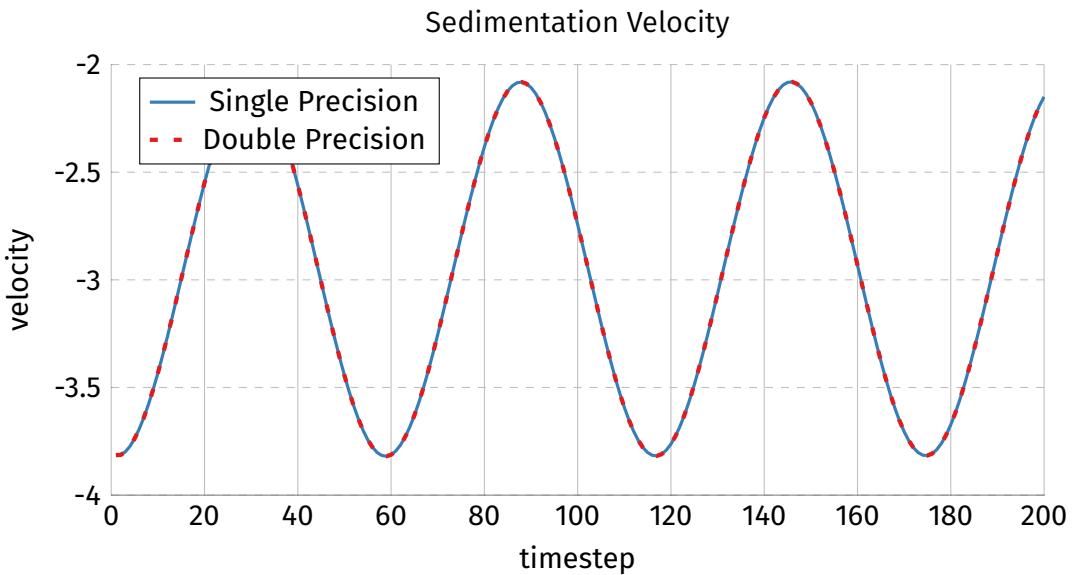


Figure 7.2: Comparison of sedimentation velocity for single- and double-precision simulation.

7.2 Fiber concentration effect on GMRES iterations

As alluded to in the discussion about the performance of direct solvers versus iterative solvers in Sec. 6.3 we will now investigate how the concentration of fibers affects the GMRES iterations. In this context the concentration of fibers refers to the average distance between each fiber and its closest neighbor. Thus we use the average minimal pair-wise distance between the fibers as a measurement of fiber concentration. Fig. 7.3 illustrates the number of iterations GMRS needs to solve the system for a given pair-wise distance. Please note that the x-axis as well as the y-axis are logarithmic to improve clarity.

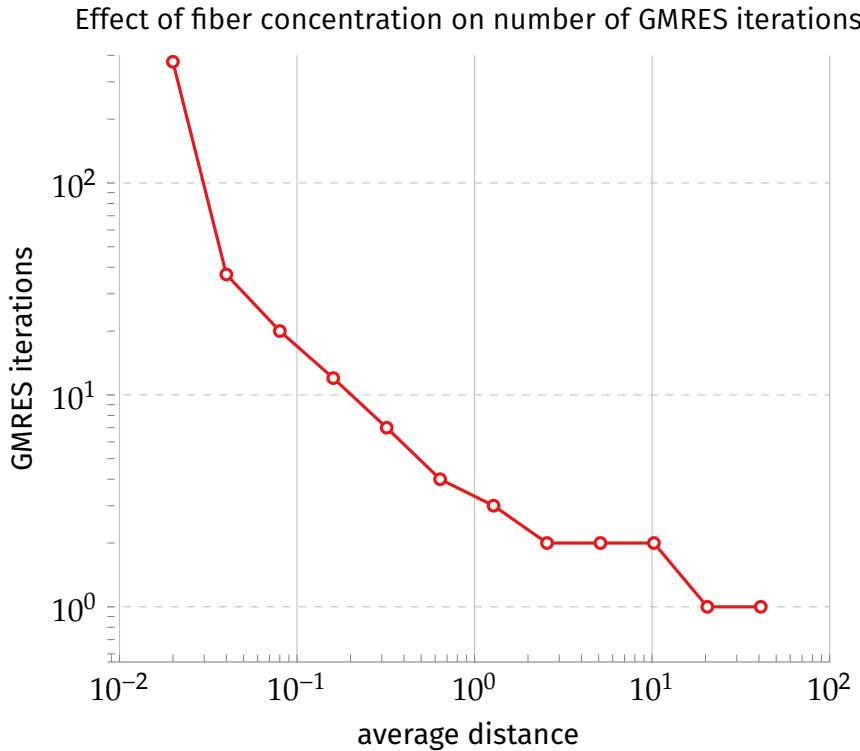


Figure 7.3: Effect of fiber concentration on GMRES iterations.

We can see that for an increasing average pair-wise distance, i.e. a lower fiber concentration, the number of iterations decreases. The average iteration count for the concentration (average distance of 0.4) we used during our benchmarks in Chapter 6 lies between 3 and 10. However, as the average pair-wise distance decreases we observe an exponential increase in the number of iterations.

To find the reason for this we have again have to look at the integral in Eqn. (3.3) which has to be solved for each fiber pair during the *Assemble System* step. The required Stokeslet computations in Eqn. (2.7) include the distance between the fibers in the denominator. As the distance gets very small the value of the Stokeslet becomes large. These large terms can then cause the assembled matrix to become ill-conditioned and thus increase the number of iterations GMRES requires to settle on a solution.

This result is important to keep in mind when simulating long running complex fiber configuration. Here the probability that any two fibers get close to each other is very high. If this is the case solving the system using GMRES takes longer than expected. In practice it might thus be beneficial to switch to the direct solver,

7.3. SEDIMENTING SPHERE

as it has a predictable runtime. This is especially true because the performance difference between direct and iterative solvers on the GPU is relatively small.

7.3 Sedimenting sphere

The next example showcases a more chaotic system with a large number of interacting fibers. This experiment is studied by several papers, e.g. [17][22][1]. It is especially interesting because the observed results only occur if enough fibers are simulated. Our GPU simulation is able to efficiently handle up to 2000 fibers and is thus ideally suited for studying this example. For the experiment 2000 fibers are initially distributed inside a sphere with a concentration of 0.4. Both their positions and orientations are randomly generated inside this sphere. An illustration of an example run can be seen in Fig. 7.4.

This sphere of packed fibers is then allowed to sediment due to gravity. Beginning from the spherical shape the interacting fibers slowly begin to form a continuously turning torus. Even though the behavior is more chaotic due to the large number of fibers and the random initial setup, this turning torus resembles the result for the tumbling orbits in Sec. 7.1.

After some time the torus breaks and the fibers are split into multiple smaller clusters. Each cluster continues to sediment separately and slowly forms its own smaller torus again. However, due to tiny variations these toruses can be harder to see.

The calculation of the presented example only takes 7.7 seconds with the GPU simulation. Consequently it is possible to perform the 500 time steps of the simulation in a little bit over 1 hour. Simulating this many fibers in such a short time will allow new research of this interesting phenomenon. One interesting research question is for example determining what influences the stability of the torus.

7.3.1 Sphere break-up

To begin investigating this question we first define the stability of the torus in terms of the break up time. How to determine the exact break up time of the torus is quite challenging. This problem and question was also examined by Park et al. [22]. They define the break up as the moment when the torus starts to bend prior to actually breaking. Additionally, they describe an algorithm that determines which fibers are part of the torus. Consequently, they compute metrics

CHAPTER 7. EXPERIMENTS

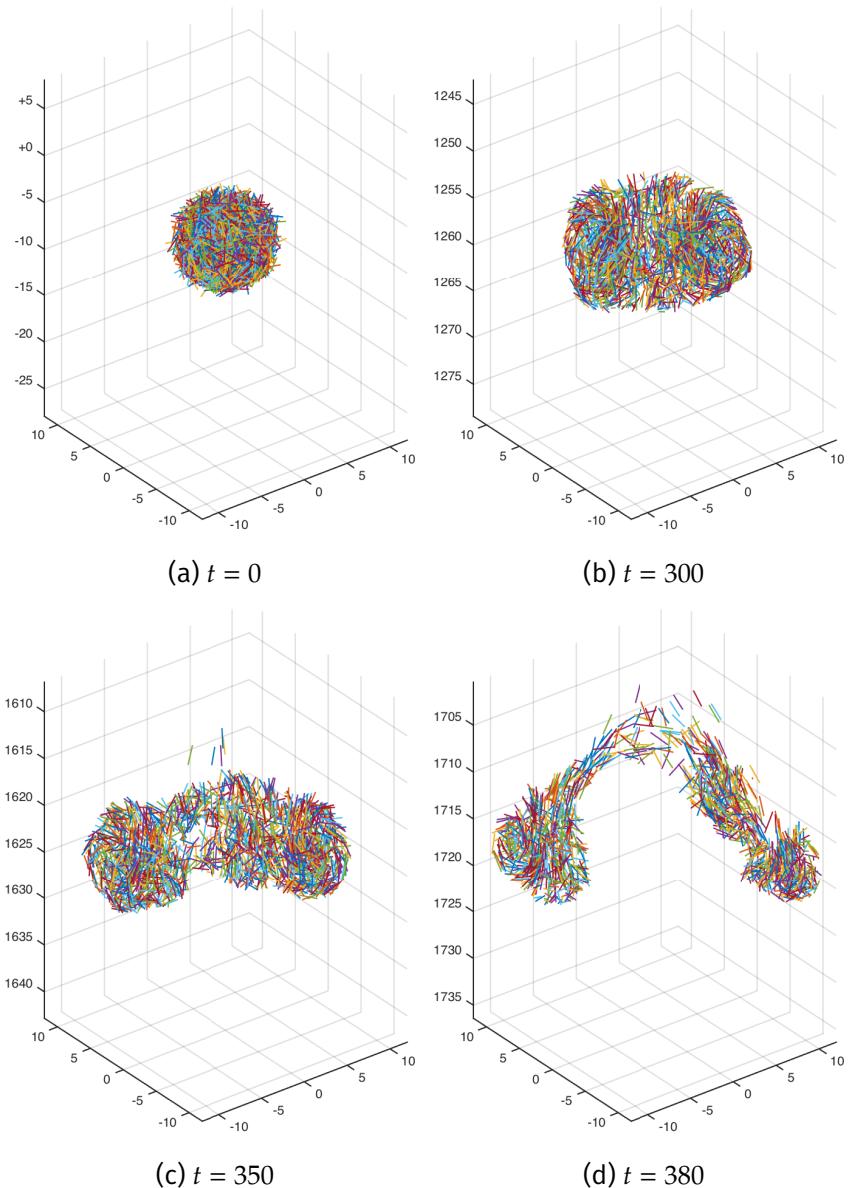


Figure 7.4: Visualization Sedimenting Sphere.

7.3. SEDIMENTING SPHERE

such as the remaining fibers, the radius and velocity of the sedimenting torus. Unfortunately, they don't define a metric that determines the exact break up time. Thus we came up with our own measure of the break up time based on their work.

We use the same definition for the torus as described by Park et al. [22]. First we compute the initial horizontal radius R_0 of the sphere. Then as the sphere sediments and starts forming the torus a small number of fibers are separate. For each time step the center of mass of all fibers belonging to the torus is calculated. Any fibers with a distance larger than R_0 in the direction of gravity are excluded from the active set of fibers belonging to the torus. For the remaining fibers we then calculate the radius R of the torus in each horizontal direction and the mean sedimentation velocity V_z . The torus metrics are illustrated for a sample run with 2000 fibers and an average distance of 0.4 in Fig. 7.5. The horizontal radii R_x and R_y of the torus and the percentage of fibers remaining in the torus are shown in Fig. 7.5a and Fig. 7.5b. The curves show the expected loss of fibers and the increasing torus radius.

We observe that the standard deviation of the radius in the direction of gravity in Fig. 7.5c continuously decreases as time passes. The standard deviation then reaches its minimum a few moments before the break up can be identified visually. We can now define this minimum of the standard deviation of the radius as a metric for the break up time. We verified this metric manually against many simulations and it showed great agreement with the visual break up. Using this metric we are now able to investigate the effect of some simulation parameters on the stability and break up time of the torus.

A surprising insight was that the standard deviation of the radius R_z seems to oscillate in Fig. 7.5c. This oscillation in the direction of gravity means that the torus is periodically expanding and shrinking until it breaks up. The same effect can also be seen in the mean sedimentation velocity V_z of the torus in Fig. 7.5d. Since the periodical change in velocity is similar to the motion observed in the simple tumbling orbits experiment in Sec. 7.1, this result strongly supports the idea that the turning torus with 2000 fibers resembles a more chaotic version of the simplified tumbling orbits.

7.3.2 Fiber concentration effect on break-up time

The first parameter we explored was the concentration of the fibers. Concentration again is defined in terms of the average distance between a fiber and its closest neighbor. For our experiments we fixed the number of fibers at 2000 and used

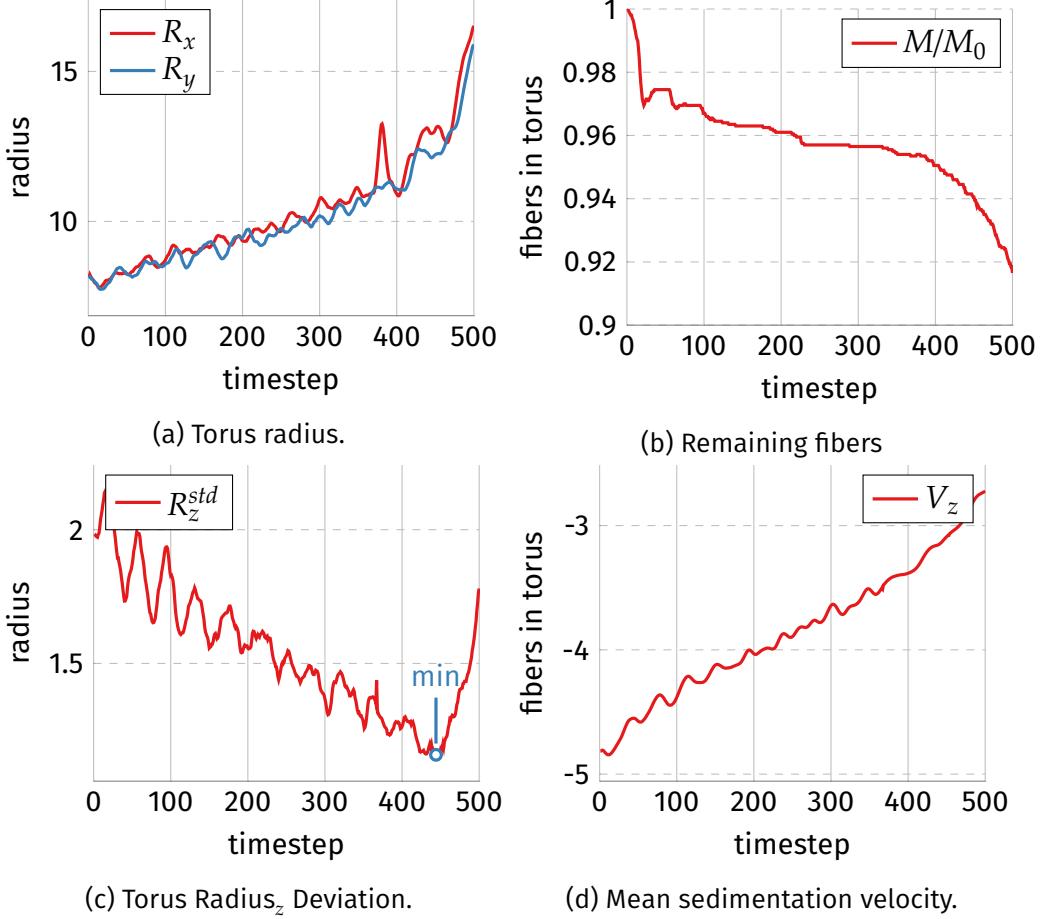


Figure 7.5: Torus composition.

the exact values for all other parameters only changing the concentration. Fig. 7.6 indicates that there is a clear correlation between the fiber concentration and the time until breakup. This matches the results found by Park et al. [22].

7.3.3 Number of fibers effect on break up time

The second parameter we looked at was the number of fibers. For this experiment only the number of fibers was changed. Every other parameter was fixed. For the concentration we chose the same value of 0.4 as during our benchmarks in Chapter 6. The resulting chart is shown in Fig. 7.7. It shows a positive correlation for the number of fibers and the break up.

These two results show that both the concentration and the number of fibers have an effect on the break up time of the torus. Future research should explore

7.3. SEDIMENTING SPHERE

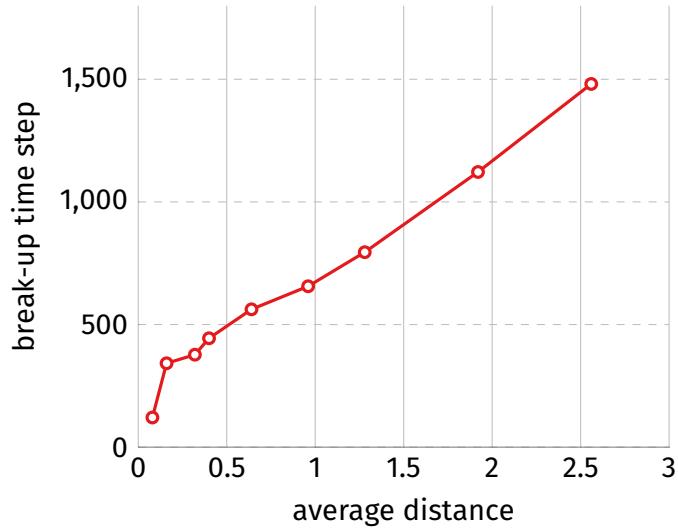


Figure 7.6: Effect of fiber concentration on torus break up time.

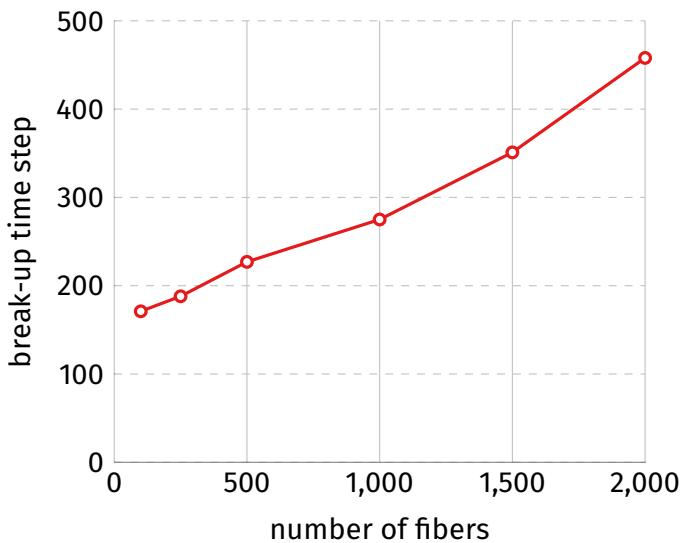


Figure 7.7: Effect of number of fibers on torus break up time.

these correlations closer as our tests only looked at two parameters in isolation. How other parameters, like external forces, the numerical accuracy and the initial fiber distribution affect the stability of the torus are promising questions.

7.4 Mixed density sphere

The final experiment we performed resembles the setup investigated by Bülow et al. [1]. In this setup instead of identical fibers, we look at the mixing of fibers with different densities. We divide the total number of fibers into two different density groups and arrange them inside a sphere, which again sediments due to gravity. To separate the two groups visually, they are colored in red and blue. The blue group has the lower density and the red group the higher density. The ratio between the densities is set to blue/red = 0.75/1.0.

Fig. 7.8 shows different timesteps of a simulation where the two density groups are separated into the top and bottom half of the sphere. Since the fibers are clearly separated, this particular configuration is referred to as an unmixed sphere. The higher density group is on top and the lower density group is at the bottom (Fig. 7.8a). After a few timesteps the higher density fibers have almost completely dropped through the lower density fibers (Fig. 7.8b). The falling high density fibers cause a large number of low density fibers to be ejected from the sphere. Both groups then proceed to almost independently form the characteristic torus observed in the previous experiment in Sec. 7.3 (Fig. 7.8c). The higher density fibers, however, form a larger torus with a lower fiber concentration. This lower concentration also causes them to lose velocity, which in turn allows the lower density fibers to drop through the larger outer torus (Fig. 7.8d). Both toruses are not very stable and soon afterwards break apart.

In contrast to the configuration with an unmixed sphere, Fig. 7.9 shows a similar simulation, starting with a randomly mixed sphere (Fig. 7.9a). The beginning of the simulation behaves similar to the unmixed case, but the separation does not happen as fast and far fewer lower density fibers are lost from the sphere (Fig. 7.9b). Both the larger higher density and the smaller lower density torus form again. Yet in contrast to the unmixed case both groups stay closer together. The close proximity prevents the groups from completely separating and causes the higher density torus to be sucked down by the passing lower density torus (Fig. 7.9c). During the following timesteps the two toruses appear to form an interlocked

7.4. MIXED DENSITY SPHERE

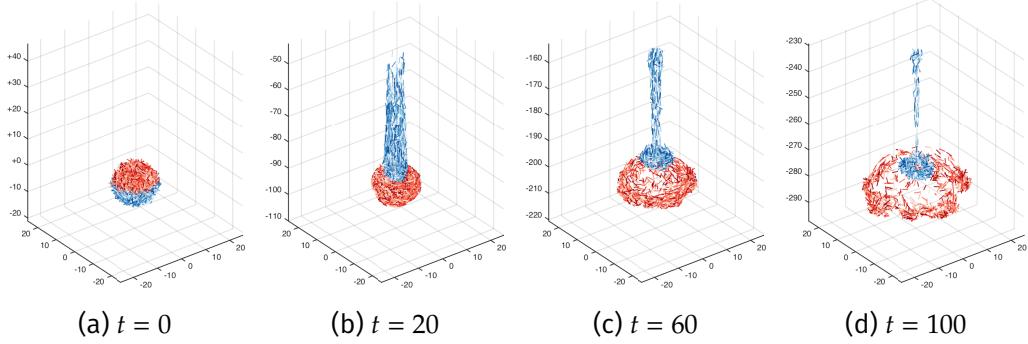


Figure 7.8: Unmixed sphere. Starting the sedimentation with the high density fibers (red) as the top half and the low density fibers (blue) as the bottom half of the sphere.

torus, which continues to sediment (Fig. 7.9d). Eventually the interlocked torus breaks up and fibers scatter.

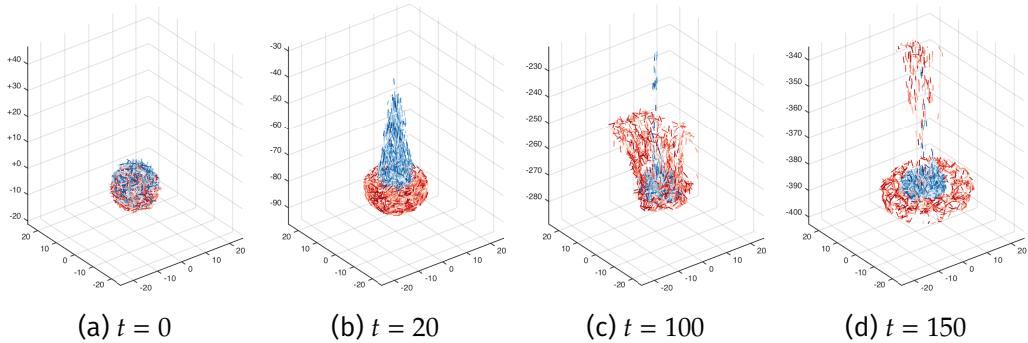


Figure 7.9: Mixed sphere. Starting the sedimentation with the high density fibers (red) and the low density fibers (blue) randomly mixed inside the sphere.

The results for the mixed and unmixed sphere with rigid straight fibers do largely agree with the results of spherical particles from Bülow et al. [1]. However, both simulation methods and setups are very different which makes a comparison very challenging. Furthermore our experiments are only a first step and many more experiments have to be carried out to get a deeper understanding of the behavior of rigid fibers in a mixed sphere. Exploring the impact of initial fiber configuration, the density ratio and the fiber concentration are all exciting future research opportunities.

Chapter 8

Conclusion

In this thesis we have developed a completely new parallel GPU implementation for the numerical simulation of rigid fibers suspension using CUDA. This is faster and therefore many more fibers can be simulated than before. Furthermore being able to simulate more fibers enables the researcher to perform more extensive and in-depth studies of the various properties of the flow and simulation.

Based on the theoretical foundation of the simulated model and the original serial Fortran implementation we rewrote the algorithm to take advantage of massively parallel computational power available in modern GPUs. We outlined the required steps to implement the algorithm using CUDA. We investigated a number of different optimizations to further improve the performance. These explored optimizations included variations of the algorithm, GPU specific implementation details and the choice of linear solver. To reach the goal of comparing the performance between CPU and GPU the new parallel algorithm was backported to the CPU using OpenMP. Thereby improving the fairness of the comparison as much as possible.

During extensive benchmarking we showed that variations perform differently depending on the underlying architecture. Optimizations developed for the CPU implementation actually slow down the GPU implementation due to diverging execution branches. We also discovered that off-the-shelf GPU implementations of iterative solvers can be slower compared to their highly optimized CPU alternatives. Hopefully future research in this area will allow for even better performance of iterative solvers on the GPU.

Overall our GPU simulation on the NVIDIA GTX 970 is able to outperform the CPU version on an Intel Core i7 4770 by a factor of 20 \times to 40 \times . This allows the GPU

simulation to handle up to 2000 fibers, which is more than double then before, and only takes 8 seconds to advance one time step on a desktop computer.

To test and validate our new parallel implementation, we performed a number of simulations of known experiments for sedimenting fibers. Using the test-case of tumbling orbits we showed that the numerical precision of our single precision implementation is able to reproduce the result obtained from the former double precision Fortran simulations. We also explored how the performance of iterative solvers is affected by the fiber concentration, showing that if fibers come too close the number of iterations increases and slows down the simulation. Our last experiment looked at the known phenomena that a sphere of fibers sedimenting form a torus and eventually break up into smaller clusters. Our results show an excellent agreement with both experimental and numerical work performed previously. Additionally, we showed that the stability of the torus is positively correlated with the increasing distance between fibers and the total number of fibers. Further research in this area is helped greatly by the fact that new case can rapidly be explored using our fast GPU simulation.

Currently the number of fibers in our simulation is limited by the amount of memory available on the GPU. In future we would like to research how to lift this restriction. One possible research direction is to avoid storing the matrix completely. Instead the required computation would have to be carried out iteratively and on-demand. This would allow for a practical unlimited number of fibers with the cost of an largely increased computation time. If this trade off is worth it remains to be seen.

Another exciting possibility is the move to a multi-device setup. Using multiple GPUs simultaneously would enable much larger systems, if the workload can be split efficiently. MAGMA, the library used for the GPU direct solver, offers a couple of interesting options in this area. The most challenging part to efficient parallelization is solving the system and broadcasting the result across multiple device. The other steps of the algorithm can be naturally partitioned and have already been implemented across multiple CPUs using OpenMPI.

Finally, to further improve the simulation it might be worth looking at the numerical algorithm itself. Adapting a different algorithm based on the fast summation method might not only improve the performance but also reduce the required memory allowing for faster and larger simulation. How this affects the accuracy and behavior of the simulation is an interesting future research question.

Bibliography

- [1] F. Bülow, H. Nirschl, and W. Dörfler. “On the settling behaviour of polydisperse particle clouds in viscous fluids”. In: *European Journal of Mechanics - B/Fluids* 50 (Mar. 2015), pp. 19–26.
- [2] Leonardo Dagum and Ramesh Menon. “OpenMP: an industry standard API for shared-memory programming”. In: *Computational Science & Engineering, IEEE* 5.1 (1998), pp. 46–55.
- [3] Xijun Fan, N. Phan-Thien, and Rong Zheng. “A direct simulation of fibre suspensions”. In: *Journal of Non-Newtonian Fluid Mechanics* 74.1998 (1998), pp. 113–135.
- [4] V Frayssé et al. *A set of GMRES routines for real and complex arithmetics on high performance computers: Technical Report TR*. Tech. rep. PA/03/03, CERFACS, 2003.
- [5] Thomas Götz. “Interactions of Fibers and Flow: Asymptotics, Theory and Numerics”. PhD thesis. 2000, p. 117.
- [6] Chris Gregg and Kim Hazelwood. “Where is the data? Why you cannot debate CPU vs. GPU performance without the answer”. In: *ISPASS 2011 - IEEE International Symposium on Performance Analysis of Systems and Software*. IEEE, Apr. 2011, pp. 134–144.
- [7] E Lisabeth Guazzelli and John Hinch. “Fluctuations and instability in sedimentation”. In: *Annual Review of Fluid Mechanics* 43.1 (Jan. 2010), pp. 97–116.
- [8] K. Gustavsson and a. K. Tornberg. “Gravity induced sedimentation of slender fibers”. In: *Physics of Fluids* 21.12 (2009), pp. 1–15.
- [9] Intel Corporation. *Intel Math Kernel Library*. 2014. URL: <https://software.intel.com/en-us/intel-mkl>.

BIBLIOGRAPHY

- [10] C.G. Joung, N. Phan-Thien, and X.J. Fan. "Direct simulation of flexible fibers". In: *Journal of Non-Newtonian Fluid Mechanics* 99. February 2001 (2001), pp. 1–36.
- [11] Sunghwan Jung et al. "Periodic sedimentation in a Stokesian fluid". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 74.3 (Sept. 2006).
- [12] S. Kamil, J. Shalf, and E. Strohmaier. "Power efficiency in high performance computing". In: *2008 IEEE International Symposium on Parallel and Distributed Processing* (2008), pp. 1–8.
- [13] Khronos Group. *OpenCL - The open standard for parallel programming of heterogeneous systems*. 2014. URL: <https://www.khronos.org/opencl/>.
- [14] Kitware. *CMake*. 2014. URL: <http://www.cmake.org/>.
- [15] Innovative Computing Laboratory. *MAGMA: MAGMA User Guide*. 2014. URL: <http://icl.cs.utk.edu/projectsfiles/magma/doxygen/>.
- [16] Victor W. Lee et al. "Debunking the 100X GPU vs. CPU myth". In: *ACM SIGARCH Computer Architecture News* 38 (2010), p. 451.
- [17] Bloen Metzger, Maxime Nicolas, and Élisabeth Guazzelli. "Falling clouds of particles in viscous fluids". In: *Journal of Fluid Mechanics* 580 (May 2007), p. 283.
- [18] NVIDIA Corporation. *CuBLAS*. 2014. URL: <https://developer.nvidia.com/cuBLAS>.
- [19] NVIDIA Corporation. *CUDA C Best Practices Guide*. 2014. URL: <http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/>.
- [20] NVIDIA Corporation. *CUDA C Programming Guide*. 2014. URL: <http://docs.nvidia.com/cuda/cuda-c-programming-guide/>.
- [21] J.D. Owens et al. "GPU Computing". In: *Proceedings of the IEEE* 96.5 (May 2008), pp. 879–899.
- [22] Joontaek Park et al. "A cloud of rigid fibres sedimenting in a viscous fluid". In: *Journal of Fluid Mechanics* 648 (Apr. 2010), p. 351.
- [23] Charles S. Peskin. "The immersed boundary method". In: *Acta Numerica* 11 (2002).

BIBLIOGRAPHY

- [24] Constantine Pozrikidis. *Boundary Integral and Singularity Methods for Linearized Viscous Flow*. Vol. 36. 1992, pp. 140–141.
- [25] K. Rupp, F. Rudolf, and J. Weinbub. “ViennaCL - A High Level Linear Algebra Library for GPUs and Multi-Core CPUs”. In: *Intl. Workshop on GPUs and Scientific Applications*. 2010, pp. 51–56.
- [26] Karl Rupp. *ViennaCL - The Vienna Computing Library*. 2014. URL: <http://viennacl.sourceforge.net/doc>.
- [27] Paal Skjetne, Russell F Ross, and Daniel J Klingenberg. “Simulation of single fiber dynamics”. In: *The Journal of Chemical Physics* 107 (1997), pp. 2108–2121.
- [28] John M. Stockie and Sheldon I. Green. “Simulating the Motion of Flexible Pulp Fibres Using the Immersed Boundary Method”. In: *Journal of Computational Physics* 147 (1998), pp. 147–165.
- [29] Anna Karin Tornberg and Katarina Gustavsson. “A numerical method for simulations of rigid fiber suspensions”. In: *Journal of Computational Physics* 215.1 (June 2006), pp. 172–196.
- [30] Anna Karin Tornberg and Michael J. Shelley. “Simulating the dynamics and interactions of flexible fibers in Stokes flows”. In: *Journal of Computational Physics* 196.1 (May 2004), pp. 8–40.
- [31] Zhang Xianyi. *OpenBLAS*. 2014. URL: <http://www.openblas.net/>.
- [32] Satoru Yamamoto and Takaaki Matsuoka. “Dynamic simulation of fiber suspensions in shear flow”. In: *The Journal of chemical physics* 102.5 (1995), pp. 2254–2260. URL: <http://link.aip.org/link/?JCPA6/102/2254/1>.

BIBLIOGRAPHY

Parameters for tumbling orbits experiment.

Parameter	Value
Number of fibers	16
Concentration	0.2146
Timestep	0.1
Slenderness	0.01
Number of terms in force expansion	5
Number of quadrature intervals	8
Number of quadrature points per interval	3

Parameters for sedimenting sphere experiment.

Parameter	Value
Number of fibers	2000
Concentration	0.2
Timestep	0.1
Slenderness	0.01
Number of terms in force expansion	5
Number of quadrature intervals	8
Number of quadrature points per interval	3