A high performance computing approach to the simulation of fluid-solid interaction problems with rigid and flexible components

Arman Pazouki\textsuperscript{1,}\textsuperscript{*}, Radu Serban\textsuperscript{1,}\textsuperscript{†}, Dan Negrut\textsuperscript{1,}\textsuperscript{‡}

\textsuperscript{1} University of Wisconsin-Madison, 1513 University Avenue, Madison, WI-53706, USA

Abstract
This work outlines a unified multi-threaded, multi-scale High Performance Computing (HPC) approach for the direct numerical simulation of Fluid-Solid Interaction (FSI) problems. The simulation algorithm relies on the extended Smoothed Particle Hydrodynamics (XSPH) method, which approaches the fluid flow in a Lagrangian framework consistent with the Lagrangian tracking of the solid phase. A general 3D rigid body dynamics and an Absolute Nodal Coordinate Formulation (ANCF) are implemented to model rigid and flexible multi-body dynamics. The two-way coupling of the fluid and solid phases is supported through use of Boundary Condition Enforcing (BCE) markers that capture the fluid-solid coupling forces by enforcing a no-slip boundary condition. The solid-solid short range interaction, which has a crucial impact on the small-scale behavior of fluid-solid mixtures, is resolved via a lubrication force model. The collective system states are integrated in time using an explicit, multi-rate scheme. To alleviate the heavy computational load, the overall algorithm leverages parallel computing on Graphics Processing Unit (GPU) cards. Performance and scaling analysis are provided for simulations scenarios involving one or multiple phases with up to tens of thousands of solid objects. The software implementation of the approach, called Chrono::Fluid, is part of the Chrono project and available as an open-source software.

Keywords
fluid-solid interaction · high performance computing · smoothed particle hydrodynamics · rigid body dynamics · flexible body dynamics

1. Introduction

The last decade witnessed a paradigm shift in the microprocessor industry towards chip designs that provide strong support for parallel computing. Teraflop computing, i.e. computing at the rate of $10^{12}$ Floating-point Operation Per Second, until recently the privilege of a select group of large research centers, is becoming a commodity due to inexpensive GPU cards and many-core x86 processors. The steady improvements in processor speed and architecture, memory layout and capacity, and power efficiency have motivated a trend of re-evaluating simulation algorithms with an eye towards identifying solutions that map well onto these new hardware platforms. In this context, a scrutinizing of the existing Fluid-Solid Interaction (FSI) solutions reveals that they are mostly inadequate since they either rely on algorithms that do not map well on emerging computer architectures, or they are unable to capture phenomena of interest which may require the simulation of tens of thousands of

\textsuperscript{*}E-mail: pazouki@wisc.edu
\textsuperscript{†}E-mail: serban@wisc.edu
\textsuperscript{‡}E-mail: negrut@wisc.edu
rigid and deformable bodies that interact directly or through the fluid media. FSI problems are usually simulated in either an Eulerian-Lagrangian framework, where the Lagrangian solid phase moves with/over the Eulerian grid used for fluid discretization, or an Eulerian-Eulerian framework, where the solid phase is considered as an average of ensembles instead of a specific state. While the former approach delivers the flexibility required by general multibody dynamics problems, it does not provide a level of performance suitable for the simulation of dense fluid-solid problems such as suspensions. This can be traced back to the costly processes that overlap the two different representations of the fluid and solid phases.

The Lagrangian representation of the fluid flow dovetails well with the Lagrangian approach used for the motion of the solid phase. This approach has recently been employed in the context of multibody dynamics and particle suspension in [12, 25, 29–31, 33]. Specifically, the FSI methodology relying on Smoothed Particle Hydrodynamics (SPH) [16, 19, 22] and rigid body Newton-Euler equation of motion [10] was discussed in detail in [29, 30]. This or a similar framework was used for the investigation and validation of rigid body suspensions [30], as well as flexible beams interacting with fluid [12, 31, 33]. Additionally, support for the short-range solid-solid interaction was introduced in [30, 31], which was leveraged in the simulation of dense suspensions.

This contribution concentrates on the high performance computing approach required for the efficient simulation of FSI problems. Although the approach and algorithms explained herein can leverage any multi-threaded architecture, the CUDA library [26] was employed for the execution of all solution components on the GPU, with negligible data transfer between the host (CPU) and the device (GPU). The paper is organized as follows. The numerical methods adapted for the simulation of fluid and solid phases, fluid-solid coupling, and solid-solid short range interaction are explained in Sect. 2. The computational scheme, including the time integration algorithm, HPC-based implementation, and parallel neighbor search required for SPH are described in Sect. 3. Section 4 contains a performance analysis of the algorithm using Kepler-type GPU cards.

2. Numerical approach

The proposed FSI simulation framework combines SPH for the simulation of the fluid flow, the Newton-Euler formalism for rigid body dynamics, and ANCF for modeling flexible body dynamics. These algorithmic components are described in more detail in the following subsections, including a discussion on the approach adopted for resolving fluid-solid and solid-solid interactions.

2.1. Smoothed Particle Hydrodynamics

The term smoothed in SPH refers to the approximation of point properties via a smoothing kernel function \( W \), defined over a support domain \( S \). This approximation reproduces functions with up to 2nd order of accuracy, provided the kernel function: (1) approaches the Dirac delta function as the size of the support domain tends to zero, that is \( \lim_{h \to 0} W(r, h) = \delta(r) \), where \( r \) is the spatial distance and \( h \) is a characteristic length that defines the kernel smoothness; (2) is symmetric, i.e., \( W(r, h) = W(-r, h) \); and (3) is normal, i.e., \( \int_S W(r, h) dV = 1 \), where \( dV \) denote the differential volume. A typical spatial function \( f(x) \) is then approximated by \( \langle f(x) \rangle \) as

\[
\langle f(x) \rangle = \int_S f(x) \delta(x-x') dV \approx \int_S f(x) W(x-x', h) dV + O(h^2) \quad (1)
\]

To simplify notation, in the remainder of this document we use \( f(x) \) to represent \( \langle f(x) \rangle \). Using Eq. (1) and the divergence theorem, the spatial derivatives of a function can be mapped to the derivatives of the kernel function. For instance, the gradient of a function can be written as

\[
\nabla f(x) = \int_S \frac{f(x')}{\partial \mathcal{S}} W(x-x', h) ndA - \int_S f(x') \nabla W(x-x', h) dV \quad (2)
\]

where \( \partial \mathcal{S} \) is the boundary of the integration volume \( \mathcal{V} \) and \( dA \) is the differential area. By imposing an additional property for the kernel function, namely that (4) it approaches zero as \( r \) increases, i.e., \( \lim_{r \to \infty} W(r, h) = 0 \), the first term on the right hand side of Eq. (2) vanishes. Note that some additional considerations, which will be addressed later, are required for the SPH approximation near boundaries. The term particle in SPH terminology indicates the discretization of the domain by a set of Lagrangian particles. To remove the ambiguity caused by the use of the term rigid particles in the context of FSI problems, we use the term marker to refer to the SPH discretization unit. Each marker has mass \( m \) associated to the representative volume \( dV \) and carries all of the...
essential field properties. As a result, any field property at a certain location is shared and represented by the markers in the vicinity of that location. The value at a given unknown at a given location is calculated according to the distance between that location and the collection of markers in its vicinity using the field values at these markers and the expression of the kernel function $W$. This leads for the second approximation embedded in SPH, which can be expressed as

$$f(x) = \sum_{b} \frac{m_b}{\rho_b} f(x_b) W(x - x_b, h).$$

(3)

where $b$ is the marker index and $\rho_b$ is the fluid density, smoothed at the marker location $x_b$. The summation in Eq. (3) is over all markers whose support domain overlaps the location $x$. Several other properties of the kernel functions are provided in [16]. For instance, the kernel function should be a positive and monotonically decreasing function of $r$, which implies that the influence of distant markers on field properties at a given location is less than that of nearby markers. Moreover, for acceptable computational performance and to avoid a quadratic computational complexity, kernel functions have a compact domain of influence with a radius $R$, defined as some finite multiple of $\kappa$ of the characteristic length $h$, as shown in Figure 1. The methodology adopted herein relies on a cubic spline,

$$W(q, h) = \frac{1}{4\pi h^3} \times \begin{cases} (2 - q)^3 - 4(1 - q)^3, & 0 \leq q < 1 \\ (2 - q)^3, & 1 \leq q < 2 \\ 0, & q \geq 2 \end{cases}$$

where $q \equiv |r|/h$, has a support domain with radius $2h$.

2.1.1. SPH for fluid dynamics

For fluid dynamics, the continuity and momentum equations are given in Lagrangian form as

$$\frac{d\rho}{dt} = -\rho \nabla \cdot \mathbf{v}$$

(5)

and

$$\frac{d\mathbf{v}}{dt} = -\frac{1}{\rho} \nabla \rho + \frac{\mu}{\rho} \nabla^2 \mathbf{v} + \mathbf{f},$$

(6)

where $\mu$ is the fluid viscosity, while $\mathbf{v}$ and $\rho$ are the flow velocity and pressure, respectively. Within the SPH framework described earlier, Eqs. (5) and (6) are discretized at an arbitrary location $x = x_0$ within the fluid domain as [24]

$$\rho_0 = \frac{d\rho_0}{dt} = \rho_0 \sum_{b} \frac{m_b}{\rho_b} (\mathbf{v}_0 - \mathbf{v}_b) \cdot \nabla W_{0b},$$

(7)

and

$$\dot{\mathbf{v}}_0 = \frac{d\mathbf{v}_0}{dt} = -\sum_{b} m_b \left[ \left( \frac{\rho_b}{\rho_0} + \frac{\rho_0}{\rho_b} \right) \nabla W_{0b} + \Pi_{0b} \right] + \mathbf{f}_0.$$ 

(8)

In the above equations, quantities with subscripts $a$ and $b$ are associated with markers $a$ and $b$ (see Figure 1), respectively. It is important to note that these quantities are different from the corresponding physical quantities at locations $x_a$ and $x_b$. The viscosity term $\Pi_{ab}$ is defined as

$$\Pi_{ab} = -\frac{(\mu_a + \mu_b) x_{ab} \cdot \nabla W_{ab}}{\rho_a^2 (x_{ab}^2 + \epsilon h_{ab}^2)} \nabla_a W_{ab},$$

(9)

where $x_{ab} = x_a - x_b$, $W_{ab} = W(x_{ab}, h)$, $\nabla_a$ is the gradient with respect to $x_a$, i.e., $\partial/\partial x_a$, and $\epsilon$ is a regularization coefficient. Quantities with an over-bar are averages of the corresponding quantities for markers $a$ and $b$.

Alternative viscosity discretizations include:

1. the model suggested in [5]:

$$\Pi_{ab} = -\mu_a \rho_a \nabla a \cdot \nabla b \left( \rho_a \rho_b (\mu_a + \mu_b) (x_{ab}^2 + \epsilon h_{ab}^2) \right) \nabla a W_{ab};$$

2. direct discretization of $\nabla^2 \mathbf{v}$ [22, 24]; and

3. the class of artificial viscosity models introduced in [17, 19, 23].

However, using Eq. 9 is preferred since it has the following properties: (1) it ensures that the viscous force is along the shear direction $\mathbf{v}_{ab}$, instead of the particles center line $\mathbf{x}_{ab}$; (2) it is less sensitive to local velocities; (3) it allows for better computational efficiency by removing the nested loop required for the
computation of the $\nabla^2$ operator; and (4) it is stated in terms of physical properties, rather than model parameters like those in artificial viscosity, which are introduced primarily for numerical stabilization through damping. In the simulation of transient Poiseuille flow, although we managed to virtually obtain exact results using Eq. 9 [30], the error caused by implementing either $\Pi_{ab}$ or artificial viscosity were non-negligible. In the weakly compressible SPH model, the pressure $p$ is evaluated using an equation of state [22]

$$p = \frac{c_s^2 \rho_0}{\gamma} \left\{ \left( \frac{\rho}{\rho_0} \right)^\gamma - 1 \right\},$$

(10)

where $\rho_0$ is the reference density of the fluid, $\gamma$ tunes the stiffness of the pressure-density relationship, and $c_s$ is the speed of sound. The value $c_s$ is adjusted depending on the maximum speed of the flow, $V_{max}$, to keep the flow compressibility below some arbitrary value. We use $\gamma = 7$ and $c_s = 10V_{max}$, which allows 1% flow compressibility [22].

The fluid flow equations (7) and (8) are solved together with

$$\dot{x}_a = \frac{dx_a}{dt} = \mathbf{v}_a,$$

(11)

to update the position of the SPH markers.

The original SPH summation formula calculates the density according to

$$\rho_a = \sum_b m_b W_{ab}.$$

(12)

Equation (7), which evaluates the time derivative of the density, was preferred to the above since it produces a smooth density field, works well for markers close to the boundaries (the free surface, solid, and wall), and does not exhibit the large variations in the density field introduced when using Eq. (12) close to the boundaries. However, Eq. (7) does not guarantee consistency between a marker’s density and the associated mass and volume [3, 21, 24]. The so-called density re-initialization technique [6] attempts to address this issue by using Eq. (7) at each time step and periodically; i.e., every $n$ time steps, use Eq. (12) to correct the mass-density inconsistency. The results reported herein were obtained with $n = 10$. The Moving Least Squares method or a normalized version of Eq. (12) could alternatively be used to address the aforementioned issues, see [6, 7].

Finally, the methodology proposed employs the extended SPH approach (XSPH), which prevents extensive overlap of markers’ support domain and enhances flow incompressibility [20]. This correction takes into account the velocity of neighboring markers through a mean velocity evaluated within the support of a nominal marker $a$ as

$$\dot{\mathbf{v}}_a = \mathbf{v}_a + \Delta \mathbf{v}_a,$$

(13)

where

$$\Delta \mathbf{v}_a = \frac{1}{5} \sum_b \frac{m_b W_{ab}}{\rho_b} (\mathbf{v}_b - \mathbf{v}_a) W_{ab},$$

(14)

and $0 \leq \zeta \leq 1$ adjusts the contribution of the neighbors’ velocities. All the results reported herein were obtained with $\zeta = 0.5$.

The modified velocity calculated from Eq. (13) replaces the original velocity in the density and position update equations, but not in the momentum equation [6].

2.2. Rigid body dynamics

Once the fluid-solid interactions between individual markers; i.e., the right hand side of Eqs. (7) and (8), are accounted for, the total rigid body force and torque due to the interaction with the fluid can be obtained by respectively summing the individual forces and their induced torques over the entire rigid body. They are then added to any other forces, including external and contact forces. The dynamics of the rigid bodies is fully characterized by the Newton–Euler equations of motion (EOM), see for instance [10]:

$$\frac{d\mathbf{V}_i}{dt} = \frac{\mathbf{F}_i}{M_i},$$

(15)

$$\frac{d\mathbf{V}_i}{dt} = \mathbf{V}_i,$$

(16)

$$\frac{d\omega_i^T}{dt} = J^{-1} \left( \mathbf{T}_i^T - \mathbf{\Omega}_i^T \mathbf{\Omega}_i^T \right),$$

(17)

$$\frac{d\mathbf{q}_i}{dt} = \frac{1}{2} \mathbf{G}_i \mathbf{\omega}_i,$$

(18)

and

$$\mathbf{q}_i \mathbf{q}_i - 1 = 0,$$

(19)

where $\mathbf{F}_i$, $\mathbf{T}_i$, $\mathbf{X}_i$, $\mathbf{V}_i$, $\mathbf{\omega}_i \in \mathbb{R}^3$ denote the force, torque, position, velocity, and angular velocity associated to body $i$, $i = 1, 2, \ldots, n_w$, respectively. The quantity $\mathbf{q}_i$ denotes the rotation quaternion, while $M_i$ and $J_i$ are the mass and moment of inertia, respectively. Quantities with a prime symbol are represented in the rigid body local reference frame. Given $\mathbf{a} = [a_x, a_y, a_z]^T \in \mathbb{R}^3$ and $\mathbf{q} = [q_x, q_y, q_z, q_w]^T \in \mathbb{R}^4$, the auxiliary matrices $\mathbf{\tilde{a}}$ and $\mathbf{G}$ are defined as [10]

$$\mathbf{\tilde{a}} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix}$$

and

$$\mathbf{G} = \begin{bmatrix} -q_y & q_z & q_w & -q_x \\ q_z & -q_w & q_x & q_y \\ -q_w & -q_x & q_y & q_z \\ q_x & q_y & q_z & -q_w \end{bmatrix}.$$
2.3. Flexible body dynamics

The ANCF formulation [36], which allows for large deformations and large rigid body rotations, is adopted herein for the simulation of flexible bodies suspended in the fluid. While extension to other elastic elements is straightforward, the current Chrono::Fluid implementation only supports gradient deficient ANCF beam elements, which are used to model slender flexible bodies composed of \( n_s \) adjacent ANCF beam elements. The flexible bodies are modeled using a number \( n_s = n_a + 1 \) of equally-spaced node beam elements, each represented by 6 coordinates, \( e_j = [r^t_j, r^t_{j+1}]^T, j = 0, 1, \ldots, n_a; \) i.e., the three components of the global position vector of the node, and the three components of its slope. This is therefore equivalent to a model using \( n_s \) ANCF beam elements with \( 6 \times n_a \) continuity constraints, but is more efficient in that it uses a minimal set of coordinates. We note that formulations using gradient deficient ANCF beam elements display no shear locking problems [9, 34, 35] and, due to the reduced number of nodal coordinates, are more efficient than fully parametrized ANCF elements. However, gradient deficient ANCF beam elements cannot describe a rotation about its axis and therefore cannot model torsional effects.

Consider first a single ANCF beam element of length \( \ell \). The global position vector of an arbitrary point on the beam center line, specified through its element spatial coordinate \( 0 \leq x \leq \ell \), is then obtained as

\[
r(x, e) = S(x)e,
\]

where \( e = [e^T_0, e^T_1] \in \mathbb{R}^{12} \) is the vector of element nodal coordinates. With \( I \) being the 3 \( \times \) 3 identity matrix, the shape function matrix \( S = [S_1 \  S_2 \  S_3 \  S_4] \in \mathbb{R}^{1 \times 12} \) is defined using the shape functions [36]

\[
S_1 = 1 - 3 \xi^2 + 2 \xi^3,
S_2 = \xi \left( \xi^2 - 2 \xi + 1 \right),
S_3 = 3 \xi^2 - 2 \xi^3,
S_4 = \xi \left( -\xi^2 + \xi^3 \right),
\]

where \( \xi = x/\ell \in [0,1] \). The element EOM are then written as

\[
\text{Me} + Q^e = Q^p,
\]

where \( Q^e \) and \( Q^p \) are the generalized element elastic forces and applied forces, respectively, and \( \text{Me} \in \mathbb{R}^{12 \times 12} \) is the symmetric consistent element mass matrix defined as

\[
\text{M} = \int_0^\ell \rho_i AS^T S \, dx.
\]

The generalized element elastic forces are obtained from the strain energy expression [36] as

\[
Q^e = \int_0^\ell E A e^{T_1} \left( \frac{\partial e^{T_11}}{\partial e} \right) \, dx + \int_0^\ell E I k \frac{\partial \kappa}{\partial \epsilon} \, dx,
\]

where \( e^{T_1} = (r^{T}_{r} r_{s} - 1)/2 \) is the axial strain and \( \kappa = r_{r} \times r_{s}/r_{s}^{3} \) is the magnitude of the curvature vector. The required derivatives of the position vector \( r \) can be easily obtained from Eq. (21) in terms of the derivatives of the shape functions as \( r_{r}(x, e) = S_{r}(x)e \) and \( r_{s}(x, e) = S_{s}(x)e \).

External applied forces, in particular the forces due to the interaction with the fluid, see Sect. 2.4, are included as concentrated forces at a BCE marker. The corresponding generalized forces are obtained from the expression of the virtual work as

\[
Q^p = S^T(x_{a}) F - \int_{0}^{\ell} \rho_i A S^T \, dx.
\]

In the above expressions, \( \rho_i \) represents the element mass density, \( A \) is the cross section area, \( E \) is the modulus of elasticity, and \( I \) is the second moment of area.

The EOM for a slender flexible body composed of \( n_s \) ANCF beam elements are obtained by assembling the elemental EOMs of Eq. (23) and taking into consideration that adjacent beam elements share 6 nodal coordinates. Let \( \hat{e} = [e^T_0, e^T_1, \ldots, e^T_{n_s}] \) be the set of independent nodal coordinates; then the nodal coordinates for the \( j \)-th element can be written using the mapping

\[
\begin{bmatrix}
    e^T_j \\
    e^T_{j+1}
\end{bmatrix} = B_j \hat{e}, \quad \text{with} \quad B_j = \begin{bmatrix}
    0 & 0 & \cdots & 1_1 & 0 & \cdots & 0 \\
    0 & 0 & \cdots & 0 & 1_1 & \cdots & 0
\end{bmatrix}
\]

and the assembled EOMs are obtained, from the principle of virtual work, as follows. Denoting by \( \text{Me} \) be the element mass matrix of Eq. (24) for the \( j \)-th ANCF beam element, it can be written in block form as

\[
\text{Me} = \begin{bmatrix}
    M_{j,11} & & & \\
    & M_{j,22} & & \\
    & & \ddots & \\
    & & & M_{j,n_{s}n_{s}}
\end{bmatrix},
\]

where \( M_{j,ii} = M_{j,ii}^{T} \) and all sub-blocks have dimension \( 6 \times 6 \). Here, \( i \) denotes the left end of the beam element, i.e., the node
characterized by the nodal coordinates \( e_{j-1} \), while \( r \) corresponds to the node with coordinates \( e_r \). With a similar decomposition of a generalized element force into

\[
Q_i = \begin{bmatrix} Q_{i,1} \\ Q_{i,r} \end{bmatrix}
\]  
(30)

one obtains

\[
\dot{\bar{M}}e = \dot{\bar{Q}}^r - \dot{\bar{Q}}^t
\]  
(31)

where

\[
\dot{\bar{M}} = \begin{bmatrix} M_{i,1} & M_{i,r} \\ M_{i,r} & M_{i,rr} + M_{i,11} \\ & & M_{r,r} & \vdots \\ & & & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots \\ & & & & & M_{n,rr} \end{bmatrix}
\]  
(32)

\[
\dot{\bar{Q}}^t - \dot{\bar{Q}}^r = \begin{bmatrix} \sum Q^t_{i,j} + \sum Q^t_{i,r} \\ \sum Q^t_{i,j} + \sum Q^t_{i,r} \\ \vdots \\ \sum Q^t_{n,r} \end{bmatrix} - \begin{bmatrix} Q^r_{i,j} \\ Q^r_{i,j} + Q^r_{j,i} \\ \vdots \\ Q^r_{n,r} \end{bmatrix}
\]  
(33)

Inclusion of additional kinematic constraints, e.g., anchoring the beam at one end to obtain a flexible cantilever or fixing its position to obtain a flexible pendulum, can be done either by formulating the EOM as differential-algebraic equations or by deriving an underlying ODE after explicitly eliminating the corresponding constrained nodal coordinates. The latter approach was used in all simulations involving flexible cantilevers discussed in Sect. 4.

2.4. Fluid-solid interaction

The two-way fluid-solid coupling was implemented based on a methodology described in [29]. The state update of any SPH marker relies on the properties of its neighbors and resolves shear as well as normal inter-marker forces. For the SPH markers close to solid surfaces, the SPH summations presented in Eqs. (7), (8), (12), and (14) capture the contribution of fluid markers. The contribution of solid objects is calculated via Boundary Condition Enforcing (BCE) markers placed on and close to the solid surface as shown in Figure 2. The velocity of a BCE marker is obtained from the rigid/deformable body motion of the solid thus ensuring the no-slip condition on the solid surface. Including BCE markers in the SPH summation equations (7) and (8) results in fluid-solid interaction forces that are added to both fluid and solid markers.

\[
\begin{align*}
F_{ij}^{\text{sub}} &= \min \left\{ -6\pi \mu \left( \frac{a_i a_j}{a_i + a_j} \right)^2 \left( \frac{1}{s} - \frac{1}{\Delta s} \right), 0 \right\} \cdot v_{nj}, \\
F_{ij}^{\text{sub}} &= \sum_{k=1}^{N} f_{ij}^k,
\end{align*}
\]  
(34)

where \( a_i \) and \( a_j \) are the sphere radii, \( v_{nj} \) is the normal component of the relative velocity, and \( s \) is the distance between surfaces. For \( s > \Delta s \), \( F_{ij}^{\text{sub}} = 0 \), and the spheres are subject only to hydrodynamic forces.

Equation (34) provides a basic model for the estimation of the lubrication force in normal direction. The generalization of this model to non-spherical objects requires the calculation of the minimum distance and the curvature of the two contact surfaces. The calculation of the partial lubrication force between nonspherical surfaces follows the approach proposed in [8] for a lattice Boltzmann method but is amended to fit the Lagrangian formulation adopted herein. Accordingly, the force model provided in Eq. (34) is modified as

\[
F_{ij}^{\text{sub}} = \sum_{k=1}^{N} f_{ij}^k,
\]  
(35)

with

\[
f_{ij}^k = \min \left\{ -\frac{3}{2} \pi \mu b^2 \left( \frac{1}{s^2} - \frac{1}{\Delta s} \right), 0 \right\} \cdot \nu_{nj},
\]
where \( s^* \) and \( \nu^* \) denote the markers' relative distance and velocity, respectively, and the summation is over all interacting markers of the two solid objects.

3. HPC implementation

Chrono::Fluid [4], an open-source simulation framework for fluid-solid interaction, provides an implementation that executes all phases of the solution method on the GPU. A brief overview of the GPU hardware and programming model adopted is followed below by a detailed discussion of the critical kernels that implement the proposed modeling and solution approach.

3.1. GPU hardware and programming model

To a very large extent, the performance of today’s simulation engines is dictated by the memory bandwidth of the hardware solution adopted. Recent numerical experiments conducted by the authors revealed that only about 5 to 10% of the peak flop rate is reached by the large number of cores present on today’s architectures since these cores most of the time idle waiting for data from global memory or RAM. It is this observation that motivated the selection of the GPU as the target hardware for implementing Chrono::Fluid. At roughly 300 GB/s, the GPU memory bandwidth stands four to five times higher than what one could expect on a fast CPU.

To describe the hardware organization of the GPU, we consider here the NVIDIA GeForce GTX 680 [18, 27]. This GPU is based on the first generation of Kepler architecture, code name GK104, which is also implemented in Tesla K10. The Kepler architecture relies on a Graphics Processing Cluster (GPC) as the defining high-level hardware block. There are a total of 4 GPCs on the GK104. Each GPC includes two Stream Multi-processors (SM), each of which has 192 Scalar Processors (SP), for a total of 1536 SPs, and 3.1 TFlops processing rate.

In addition to the processing cores, the second important aspect of the GPU hardware and programming model adopted is followed below by a detailed discussion of the critical kernels that implement the proposed modeling and solution approach.

3.2. Time integration

Chrono::Fluid uses a second order explicit mid-point Runge-Kutta (RK2) scheme [2] for the time integration of the fluid and solid phases, the latter in its rigid or flexible representation. Algorithm 1 summarizes the steps required for the time integration of a typical FSI problem.

The external forces on the rigid and flexible bodies include the FSI forces captured via BCE markers at distributed locations on the solid surfaces. The distributed forces need to be accumulated into a single force and torque at the center of each rigid body, or point forces at node locations of each flexible body. The reduction (summation of the forces and torques) is handled by parallel scan operations available through the Thrust library [11], which exposes a scan algorithm that scales linearly. The RK2 integration scheme requires the calculation of the force on the SPH markers, rigid bodies, and deformable beams at time step \( k \). The variables \( N_w, N_r, \) and \( N_b \) denote the number of markers, rigid bodies, and flexible beams, respectively. The arrays \( \rho, x, v, \) and \( \dot{v} \) store the density, position, velocity, and modified velocity for all markers, respectively; for example, \( \rho = \{\rho_i, i = 0, 1, 2, \ldots, N_w - 1\} \).

The parallel execution paradigm best supported on GPUs is "single instruction multiple data" (SIMD). In this model, if for instance two arrays of length one million are to be added, one million threads are launched with each executing the same instruction, i.e., adding two numbers, but on different data - each thread adding two different numbers. Fortunately, SIMD computing is very prevalent in the solution methodology proposed, where each SPH marker is handled by a thread in the same way in which hundreds of thousands of other threads handle their markers using different data. While strongly leveraging the SIMD model owing to the fine grain parallelism that it exposes, the methodology adopted is prone to lead to memory access patterns that do not display high spatial and/or temporal locality. This is because the SPH markers move in time leading to less structured memory accesses that adversely impact the effective bandwidth reached by the code.
3.2.1. Multi-rate integration

Stable integration of the SPH fluid equations requires step-sizes which are also appropriate for propagating the dynamics of any rigid solids in the FSI system. However, integration of the dynamics of deformable bodies, especially as their stiffness increases, calls for very small time steps. To alleviate the associated computational cost, we use a dual-rate integration scheme using intermediate steps for the integration of the flexible dynamics EOMs, typically with Δt_{SPH}/Δt_{ANCF} = 10, although stiffer problems may require ratios of up to 50. This aspect is noteworthy given that typical FSI models involve a number of SPH markers orders of magnitude larger than the number of ANCF nodal coordinates required for the flexible bodies. Without a multi-rate implementation, the numerical solution would visit the fluid phase at each integration step, an approach that led to very large solution times. This aspect is further discussed in Sect. 4.

3.3. Proximity calculation and neighbor search

In the current implementation, each marker has a list of neighbor markers; i.e., markers that fall within its support domain. Given this set of lists, the calculation of \( v^e \equiv \hat{v}_e (\rho^e, x^e, v^e) \), \( \rho^e \equiv \hat{\rho}_e (\rho^e, r^e, \hat{v}^e) \), and \( \hat{v}_e \equiv \hat{\hat{v}}_e (\rho^e, r^e, \hat{v}^e) \) is straightforward. The loop iterations in Algorithms 1 and 2 have no overlap and can be executed in parallel. The computational bottleneck thus becomes the determination of the neighbor lists through proximity calculation, a step that requires about 70% of the entire computational budget and thus critically impacts the overall performance of the simulation. Herein, we adopt a spatial binning algorithm which is sub-optimal in terms of amount of net work but superior in terms of memory usage. In this approach, the computation domain is divided into a collection of bins. The bins have side lengths equal to the maximum influence distance of a marker, i.e. \( kh \). This localizes the search for the possible interacting markers to the bin and all of its 26 immediate 3D
neighbors. Figure 3 shows the binning approach in 2D. In the neighbor search approach adopted herein, neighbor lists are not saved in memory; instead, neighbors are evaluated whenever required. Alternatively [28], the principle of action and reaction could be leveraged to calculate and store the acceleration terms for each inter-penetration. Although the second algorithm reduces the amount of work by re-using the acceleration terms for each inter-penetration, it can not be applied to the SPH method due to the massive amount of memory required to store the data associated with all inter-penetration events. Another advantage of the adopted neighbor search approach is the coalesced memory access achieved by sorting and accessing the data based on the markers location. The steps required for accessing the neighbors are summarized in Algorithm 3.

Algorithm 3 Inner loop: accessing neighbor markers

1. Divide the solution space into \( n_b \) bins of \((\Delta_x, \Delta_y, \Delta_z)\) dimensions, where \( n_b = n_x \times n_y \times n_z \), and \((n_x, n_y, n_z)\) is the number of grid cells along \((x, y, z)\) axis.
2. Construct the hash array: \( s = \{s_a | a = 0, 1, 2, ..., N_a - 1\}\) according to \( s_a = i \times n_y \times n_z + j \times n_z + k \), where \((i, j, k)\) is the location of the bin containing the marker \(a\).
3. Sort \( s \) into \( s_{\text{sorted}} \) and obtain corresponding \( \rho_{\text{sorted}}, x_{\text{sorted}}, y_{\text{sorted}}, \) and \( v_{\text{sorted}} \).
4. Construct \( c_1^i = \{c_1^i | p = 0, 1, 2, ..., n_p - 1\} \) and \( c_2 = \{c_2^i | p = 0, 1, 2, ..., n_p - 1\}\), where \( c_1^i \) and \( c_2^i \) denote the two indices in \( s_{\text{sorted}} \) that bound the sequence of hash values \( s_a = p \).
5. Access markers data in bin \((i, j, k)\) by loading \( [c_1^i, c_2^i] \) portions of the sorted arrays \( \rho_{\text{sorted}}, x_{\text{sorted}}, y_{\text{sorted}}, \) and \( v_{\text{sorted}} \), as needed.

The data sorting in step 3 of Algorithm 3 is performed using the linear-complexity radix sort available in the Thrust library. As a result, this solution component does not affect the overall linear scaling of Algorithms 1 and 2. Working with the sorted arrays for the bulk of the computation has the additional advantage of increasing the memory spatial locality: SPH markers that share a neighborhood in the physical space, do so in their memory location as well.

4. Results

The simulation approach described in Sect. 3 was implemented to execute in parallel on GPU cards using the CUDA programming environment [26]. All simulations reported in this section were run on an NVIDIA GeForce GTX 680 GPU card described in sub-section 3.1. In all simulations, if present, the flexible beams were modeled using \( n_e = 4 \) ANCF beam elements, while the integrals appearing in the elastic forces \( \mathbf{Q}^e \) in Eq. (25) were evaluated using 5 and 3 Gauss quadrature points for the axial and bending elastic forces, respectively.

A snapshot from a typical FSI problem involving flexible bodies is shown in Figure 4. Additional Chrono::Fluid simulation examples can be found at [32]. Comprehensive validation studies of the Chrono::Fluid simulation package are provided in [31]. The focus here is on numerical experiments that illustrate the performance and scalability of the proposed algorithm and its parallel implementation.

We begin by analyzing the efficiency and scaling attributes of the solution for systems composed exclusively of rigid bodies, flexible bodies, or a fluid phase. In each of these three scenar-
Table 1. Time required for advancing the rigid body dynamics simulation by one time step as function of problem size (number of rigid bodies, $N_r$)

<table>
<thead>
<tr>
<th>$N_r$ ($\times 10^3$)</th>
<th>0.49</th>
<th>2.87</th>
<th>16.59</th>
<th>56.77</th>
<th>118.23</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ (ms)</td>
<td>5</td>
<td>8</td>
<td>16</td>
<td>44</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 2. Time required for advancing the flexible body dynamics simulation by one time step as function of problem size (number of flexible bodies, $N_f$)

<table>
<thead>
<tr>
<th>$N_f$ ($\times 10^3$)</th>
<th>0.78</th>
<th>3.51</th>
<th>17.55</th>
<th>56.94</th>
<th>115.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ (ms)</td>
<td>8</td>
<td>14</td>
<td>48</td>
<td>122</td>
<td>238</td>
</tr>
</tbody>
</table>

Table 3. Time required for advancing a fluid dynamics simulation by one time step as function of problem size (number of SPH markers, $N_m$)

<table>
<thead>
<tr>
<th>$N_m$ ($\times 10^6$)</th>
<th>0.06</th>
<th>0.32</th>
<th>0.93</th>
<th>1.79</th>
<th>4.13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ (ms)</td>
<td>27</td>
<td>121</td>
<td>331</td>
<td>538</td>
<td>1150</td>
</tr>
</tbody>
</table>

Figure 5. Scaling analysis of Chrono:Fluid for rigid body dynamics, flexible body dynamics, and fluid dynamics as function of problem size ($N_r$, $N_f$, and $N_m$, respectively). The coefficients $R^2$ are specified to each linear regression.

Table 4. Simulation time per time step for an FSI problem with fixed number of SPH markers and increasing number of rigid bodies

<table>
<thead>
<tr>
<th>$N_m$ $\approx$ 3.0 $\times 10^6$, $N_f$ = 0</th>
<th>0</th>
<th>36</th>
<th>120</th>
<th>480</th>
<th>1800</th>
<th>8400</th>
<th>33160</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ (ms)</td>
<td>906</td>
<td>919</td>
<td>923</td>
<td>925</td>
<td>926</td>
<td>926</td>
<td>921</td>
</tr>
</tbody>
</table>
Table 5. Simulation time per time step for an FSI problem with fixed number of SPH markers and increasing number of flexible bodies, for two different values of the multi-rate integration factor $\tau = \Delta t_{SPH}/\Delta t_{ANCF}$

<table>
<thead>
<tr>
<th>$N_w$</th>
<th>0</th>
<th>45</th>
<th>140</th>
<th>440</th>
<th>1152</th>
<th>2100</th>
<th>4704</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_f$</td>
<td>906</td>
<td>923</td>
<td>928</td>
<td>916</td>
<td>960</td>
<td>950</td>
<td>921</td>
</tr>
<tr>
<td>$N_r$</td>
<td>906</td>
<td>973</td>
<td>978</td>
<td>965</td>
<td>1066</td>
<td>1060</td>
<td>1060</td>
</tr>
</tbody>
</table>

Table 6. Simulation time per time-step for combined FSI problems of increasing size

<table>
<thead>
<tr>
<th>$N_w$ ($\times 10^6$)</th>
<th>0.08</th>
<th>0.16</th>
<th>0.29</th>
<th>0.62</th>
<th>0.95</th>
<th>1.54</th>
<th>2.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_f$ ($\times 10^3$)</td>
<td>0.17</td>
<td>0.52</td>
<td>1.12</td>
<td>4.48</td>
<td>7.84</td>
<td>14.56</td>
<td>24.64</td>
</tr>
<tr>
<td>$N_r$ ($\times 10^3$)</td>
<td>0.16</td>
<td>0.42</td>
<td>0.84</td>
<td>2.10</td>
<td>3.36</td>
<td>5.88</td>
<td>9.66</td>
</tr>
<tr>
<td>$t$ (ms)</td>
<td>45</td>
<td>74</td>
<td>120</td>
<td>230</td>
<td>343</td>
<td>522</td>
<td>820</td>
</tr>
</tbody>
</table>

5. Conclusions

This contribution discusses details of an HPC approach to the simulation of FSI problems. Relying on: (i) a Lagrangian/Lagrangian formalism for the simulation of the fluid and solid phases, and (ii) GPU parallelism, Chrono::Fluid, the software implementation of the proposed solution methodology, is suitable for the study of multi-phase/multi-component problems such as particle suspensions, polymer flow, and biomedical applications.

Chrono::Fluid has been shown to scale linearly in the dynamics of each phase independently as well as in the overall FSI solution implementation. The dynamics of the fluid phase, i.e. the time propagation of the SPH markers, controls the overall simulation time, and neither cross-phase communications nor solid phase simulation have a notable effect on the simulation time. This represents one advantage of the adopted unified FSI methodology over co-simulation and asynchronous approaches. Chrono::Fluid, whose preliminary validation is discussed in [30], is open source and freely available under a BSD3 license.

Several directions of future work are as identified as follows: (a) revisit the SPH marker numbering scheme in order to improve the spatial and temporal memory access patterns, thus increasing the effective bandwidth of Chrono::Fluid; (b) augment the current implementation for use on cluster supercomputers that adopt a non-uniform memory access model and rely on the Message Passing Interface standard; (c) investigate problems at macroscale such as vehicle fording operations, which call for large scale simulations at low Reynolds numbers and possibly in the presence of very large viscosity; and (d) gauge the potential of Chrono::Fluid in biomechanics applications such as blood flow in deformable arteries or heart, channel occlusion in stroke, etc.

Acknowledgments

Financial support for this work was provided by National Science Foundation award CMMI-0840442. Support for the second and third authors was provided in part by Army Research Office awards W911NF-11-1-0327 and W911NF-12-1-0395.
References


Lagrangian framework for the simulation of rigid and deformable bodies in fluid pp. 33–52 (2014)


