2011

A near real-time framework for extracting tip-sample forces in dynamic atomic force microscopy

David Ray Busch
Iowa State University

A near real-time framework for extracting tip-sample forces in dynamic atomic force microscopy

by

David Ray Busch

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Mechanical Engineering

Program of Study Committee:

Baskar Ganapathysubramanian, Major Professor

Pranav Shrotriya

Srinivas Aluru

Iowa State University

Ames, Iowa

2011

DEDICATION

I would like to dedicate this thesis to my wife Sharon and to my Family, without whose support and guidance I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Firstly I would like to thank Dr. Baskar Ganapathysubramanian for his guidance, patience, and support throughout this research and my graduate studies. His encouragement has and continues to inspire me to do great things. I would like to thank my committee members Dr. Pranav Shrotriya and Dr. Srinivas Aluru for guidance and support as well. Additionally, I would like to thank the research groups of Dr. Qingze Zou and Dr. Sriram Sundararajan for their guidance on the experimental aspects of the atomic force microscope. Finally, I would like to thank my wife Sharon for her guidance, patience, and support throughout my graduate studies.

ABSTRACT

The atomic force microscope (AFM) is a versatile, high-resolution tool used to characterize the topography and material properties of a large variety of specimens at nano-scale. The interaction of the micro-cantilever tip with the specimen causes cantilever deflections that are measured by an optical sensing mechanism and subsequently utilized to construct the sample topography. Recent years have seen increased interest in using the AFM to characterize soft specimens like gels and live cells. This remains challenging due to the complex and competing nature of tip-sample interaction forces (large tip-sample interaction force is necessary to achieve favorable signal-to-noise ratios). However, large force tends to deform and destroy soft samples. In situ estimation of the local tip-sample interaction force is needed to control the AFM cantilever motion and prevent destruction of soft samples while maintaining a good signal-to-noise ratio. This necessitates the ability to rapidly estimate the tip-sample forces from the cantilever deflection during operation. This work proposes a first approach to a near real-time framework for tip-sample force inversion. The inverse problem of extracting the tip-sample force as an unconstrained optimization problem. A fast, parallel forward solver is developed by utilizing graphical processing units (GPU). This forward solver shows an effective 30000 fold speed-up over a comparable CPU implementation, resulting in milli-second calculation times. The forward solver is coupled with a GPU based particle-swarm optimization implementation. The proposed framework is demonstrated over a series of tip-sample interaction models of increasing complexity. Most of these inversions are performed in sub-second timings, showing potential for integration with on-line AFM imaging and material characterization.

CHAPTER 1. Introduction

1.1 Motivation

The ability to characterize soft materials on the micro/nano-scale has significant implications to several areas in science ranging from fundamental studies in polymer physics (1; 2; 3) to applied bio-engineering (4; 5), where understanding nanoscale behavior and evolution is essential.

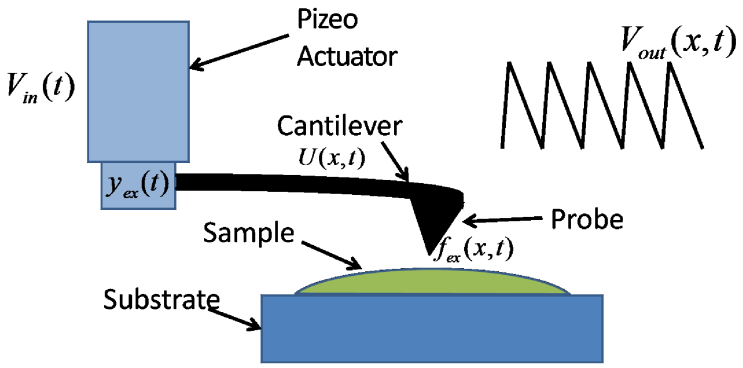

Figure 1.1 AFM diagram

Dynamic AFM imaging (6) is a very effective technique to interrogate surface topography of soft samples (1; 7), particularly for live biological samples in their physiologically friendly liquid environment (8; 9)[1]. Dynamic atomic force microscopy (dAFM) or intermittent contact mode AFM utilizes a micro-cantilever fixed-free beam to interrogate samples. The cantilever base is driven by a piezo-actuator to oscillate, causing the free tip to tap (i.e., come into intermittent contact with the sample). The oscillation amplitude and phase with respect to the cantilever base are measured and the amplitude is maintained around a set-point value via

[1]For instance, by using dynamic AFM imaging, time evolving phenomena like crystallization of polymers (2) and the dehydration process of collagen (5; 15) have been experimentally revealed for the first time.

feedback control. The measured phase and amplitude data are then utilized to construct the sample topography and also related to the material properties of the sample (10; 11).

To give an example of the capabilities of the AFM, fig. 1.2 shows an image of Bacillus atrophaeus (12) taken by the AFM. The image show a wide view, then images of hydrated and dehydrated Bacillus atrophaeus respectively. Such images and data gathered provide the opportunities to make more accurate models and develop a stronger understanding of the life-cycles of living cells. Ultimately leading to new methods of detection and treatment of viral and bacterial infections(12).

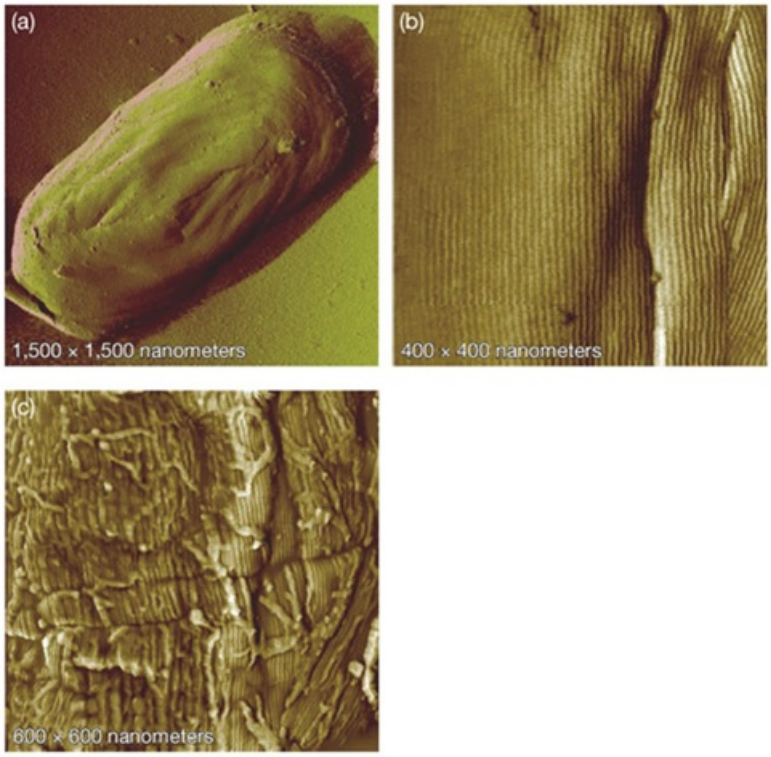

Figure 1.2 a) Wide image of Bacillus atrophaeus; b) Close image of hydrated Bacillus atrophaeus; c) Close image of dehydrated Bacillus atrophaeus; Image taken from (12)

Although by using dynamic-mode imaging, the detrimental sliding force on the sample has been largely reduced, the applied normal (tapping) force can still be large and result in not only imaging distortion, but more seriously, sample deformation and damage that can completely modify the sample (7). Large normal force, however, is needed in dynamic-mode AFM imaging to ensure imaging quality (i.e., high signal to noise ratio). The requirement of rapid scanning

(high-speed) imaging of specimens further exacerbates these problems (13; 14). *The challenge in tackling these hurdles lies in the need to maintain a small tip-sample interaction force during the scanning process.* Therefore, as a key first step to tackle this challenge, estimating the tip-sample interaction force – accurately and in real-time – is essential.

1.2 Background

Current methods for tip-sample force inversion generally require significant post-processing time and are, thus, incapable of addressing sample deformation and destruction in real-time (16). Off-line inverse problems have been formulated to estimate tip-sample interaction forces using conjugate gradient optimization (17) with limited success. The availability of newer computing methods, such as general purpose graphical processing unit computing, opens up the possibility of near real-time inversion.

1.3 Research objectives

Overall the goal of this project is to integrate analysis and manipulation of AFM cantilever tip-sample interactions, numerical computation and optimization and advanced control together to expand the capabilities of the AFM. Recent years have witnessed significant progress in both the understanding and manipulation of the non-linear cantilever tip-sample dynamics (in particular, in the dAFMs operation) (10; 11; 14; 16; 18; 19; 20; 21; 22; 23; 24) and the development of control techniques towards high-speed AFM imaging (see recent reviews (24; 25; 26; 27) and references therein) and material property measurements (28; 29; 30; 31). However, seldom have we seen efforts to marriage these two sides of work. We believe that a dynamics-control integrated approach is needed in developing the next-generation AFM technology for understanding and manipulation of dynamic evolutions of soft samples.

This paper focuses on formulating and implementing a parallel computational framework for fast inversion of tip-sample forces by using the hardware and software capabilities of GPU's and Compute Unified Device Architecture (CUDA), respectively. To the authors best knowledge, this is the first time that near real-time (sub-second) inversion of tip-sample forces has been

showcased. Other contributions include:

1. Formulating the problem of estimating the tip-sample interaction force as an inverse problem posed as an unconstrained optimization problem

2. Developing an ultra-fast predictive model for AFM dynamics based on parallel algorithms implemented on GPUs

3. Applying gradient-free optimization techniques to quickly find a solution to the optimization problem

4. Showcasing a hierarchy of models for inversion

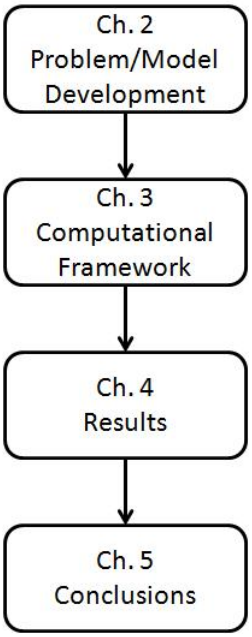

Figure 1.3 Thesis layout

To display the progress toward these objectives the paper's organization is displayed in fig. 1.3 and is as follows; In Chapter 2 we formulate the problem definition and pose the forward and inverse problems. Chapter 3 details the computational developments and algorithms along with some timing and complexity studies. In Chapter 4 we showcase the fast inversion framework on several examples of increasing complexity. Chapter 5 with give conclusions and suggest future work.

CHAPTER 2. Problem/Model development

2.1 Introduction

The physics of the sensing process demonstrates the difficulty in extracting the tip-sample interactions from the measured deflection data. A schematic of the AFM sensing process is shown in Fig. 2.1.

$$V_{in}(t) \rightarrow y_{ex}(t) \rightarrow f_{ex}(x,t) \rightarrow U(x,t) \leftrightarrow V_{out}(x,t)$$

Figure 2.1 Dynamic atomic force microscope force interaction process.

An AC input voltage ($V_{in}(t)$), usually a sinusoidal wave is sent to a piezo-electric actuator attached to the base of the cantilever (see Fig. 1.1), resulting in the oscillation of the base of the cantilever, $y_{ex}(t)$. Then as the vibrating tip is brought into intermittent contact with the sample, the tip-sample interaction force ($f_{ex}(x,t)$) is induced, which in-turn, results in the change of the oscillation (or vibration) pattern at the cantilever tip, $U(x)$. The tip deflection is measured using an optical sensing scheme. We approach the problem of extracting the tip-sample interaction force $f_{ex}(x,t)$ in two stages – first solve the forward dynamics problem of calculating cantilever deflections given a tip-sample interaction force and subsequently solving associated inverse problem of calculating interaction forces given a cantilever deflection.

2.2 The forward problem: Quantification of force effects on AFM cantilever

The forward problem quantifies the relationship between the cantilever base displacement $y_{ex}(t)$ and the tip sample interaction $f_{ex}(x,t)$ with the tip deflection $U(x,t)$. Formally, the

definition of the forward problem is as follows:

FP: *Given the cantilever properties,* (E, I, μ)*, the cantilever base displacement* $y_{ex}(t)$*, and the parametrized tip-sample interaction force, calculate the cantilever deflection* $U(x, t)$.

We model the cantilever forward dynamics by using Euler-Bernoulli (EB) beam theory. The choice is driven by the following rational:(a) the EB model provides a more accurate description of AFM cantilever dynamics than the conventional simple harmonic oscillator model (32); (b) from an inverse problem standpoint the EB model is computationally more tractable then finite element formulations (16; 32), with minimal loss in fidelity (14; 33); and (c) all assumptions made in EB beam theory is satisfied by an AFM cantilever (18; 19). In the following subsection the details of EB model will be shared.

2.2.1 Cantilever model

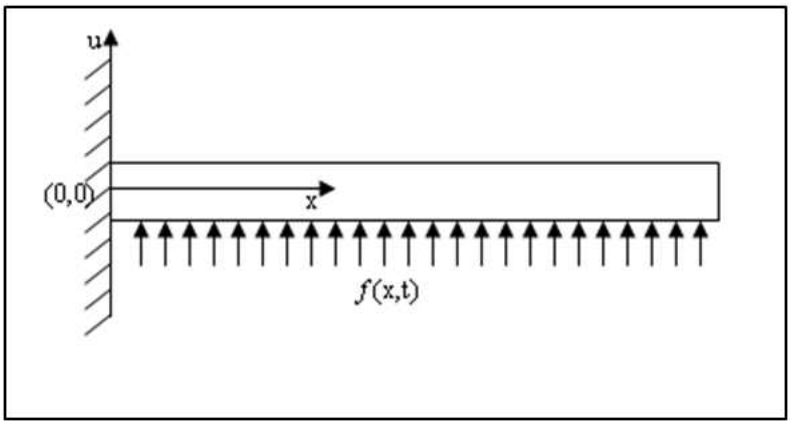

Figure 2.2 Cantilever model coordinate system.

The development and discussion of the cantilever model follows from the work of Meirovitch's book "Fundamentals of Vibrations" and selections from Massachusetts Institute of Technology Open Course Ware Mechanics and Materials II 2.001 (34; 35). Consider a fixed/free beam with coordinates system given as having the x axis extending from the fixed end through the free end of the beam and a u axis, perpendicular to the x axis, with origin at the fixed end of the beam. This beam will also be under an interaction force of $f(x, t)$. For a visual description of this coordinates set-up see Fig. 2.2. For this model, the following symbols are used:

$u =$ displacement of cantilever from neutral axis,

$x =$ distance along cantilever from base,

$t =$ time,

$E =$ elasticity,

$I =$ inertia,

$\mu =$ mass per unit length,

$f =$ forces interacting with cantilever,

$y =$ displacement on the u axis of cantilever base from its neutral axis.

Also, the following simplifying assumptions are being made:

1. The cross-section of the beam is uniform.
2. The length of the beam L is very large compared the width b and b is very large compared to h height.
3. The cantilever is fully elastic.
4. The amplitude of vibration of the cantilever is much smaller than the length or width.

To develop an equation to describe the motion of this beam, consider an infinitesimal piece of it. As shown in figure 2.3, this infinitesimal piece, under force f, will obey the laws of linear and angular momentum. Consider linear momentum. Equation 2.1 shows in this case that the acceleration of the beam (in the case of linear momentum) is caused by the force on the beam plus the change in momentum.

$$\mu dx \frac{\partial^2 u(x,t)}{\partial t^2} = Q(x+dx,t) - Q(x,t) + f(x,t)dx \tag{2.1}$$

In consideration of angular momentum:

$$0 = Q(x+dx,t)dx + M(x+dx,t) - M(x,t) - f(x,t)dx\frac{dx}{2} \tag{2.2}$$

Using Taylor series to the first term, equations 2.1 and 2.2 can be rewritten as follows:

$$\mu dx \frac{\partial^2 u(x,t)}{\partial t^2} = \frac{\partial Q(x,t)}{\partial x}dx + f(x,t)dx \tag{2.3}$$

$$0 = (Q(x,t) + \frac{\partial Q(x,t)dx}{\partial x})dx + \frac{\partial M(x,t)}{\partial x}dx - f(x,t)dx\frac{dx}{2} \tag{2.4}$$

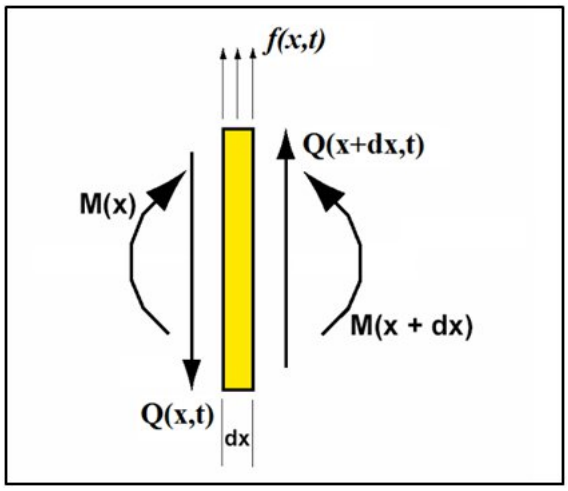

Figure 2.3 Infinitesimal piece of the cantilever. Figure adapted from (35) for this paper.

Using the fact that the cantilever piece is infinitesimal in size, the second order dx terms would be sufficiently small to neglect. Rewriting equation 2.4 neglecting second order terms yields:

$$\frac{\partial M(x,t)}{\partial x} = -Q(x,t),\ 0 < x < L \tag{2.5}$$

The bending moment $M(x,t)$ and sheer force of $Q(x,t)$ can be related to the bending displacement by:

$$M(x,t) = EI\frac{\partial^2 u(x,t)}{\partial x^2} \tag{2.6}$$

$$Q(x,t) = -EI\frac{\partial^3 u(x,t)}{\partial x^3} \tag{2.7}$$

If eqn. 2.4 is divided through by dx and eqns. 2.6 and 2.7 are considered, the following is the result:

$$\mu\frac{\partial^2 u(x,t)}{\partial t^2} = -\frac{\partial^4 u(x,t)}{\partial x^4}dx + f(x,t)dx, \tag{2.8}$$

which is the Euler-Bernoulli beam equation. To eqn. 2.8, a viscous damping term is added resulting in:

$$\mu\frac{\partial^2 u(x,t)}{\partial t^2} + 2\zeta\frac{\partial u(x,t)}{\partial t} = -\frac{EI\partial^4 u(x,t)}{\partial x^4}dx + f(x,t)dx \tag{2.9}$$

with boundary conditions (BC's) of:

$$u(0,t) = u'(0,t) = u''(L,t) = u'''(L,t) = 0, \tag{2.10}$$

with general initial conditions. To solve this problem, we consider the homogeneous problem of equation 2.9. The resulting equation becomes a differential eigenvalue problem:

$$\frac{\partial^4 u}{\partial x^4} + \frac{\mu}{EI}\frac{\partial^2 u}{\partial t^2} + \frac{2\zeta}{EI}\frac{\partial u}{\partial t} = 0, \tag{2.11}$$

with BC's from equation 2.10. The solution is via separation of variables,

$$u(x,t) = \Phi(x)\eta(t), \tag{2.12}$$

which yields,

$$\eta\frac{\partial^4 \Phi}{\partial x^4} + \frac{\mu\Phi}{EI}\frac{\partial^2 \eta}{\partial t^2} + \frac{2\zeta\Phi}{EI}\frac{\partial \eta}{\partial t} = 0 \tag{2.13}$$

grouping like terms,

$$\frac{\beta}{\Phi}\frac{\partial^4 \Phi}{\partial x^4} + \frac{1}{\eta}\frac{\partial^2 \eta}{\partial t^2} + \frac{2\zeta}{\beta\mu\eta}\frac{\partial \eta}{\partial t} = 0 \tag{2.14}$$

where $\beta = \frac{EI}{\mu}$. Next we assume that each variable has solution of the form $-\lambda^4$, focusing on the spatial variable first, we have:

$$\frac{\beta}{\Phi}\frac{\partial^4 \Phi}{\partial x^4} = -\lambda^4 \tag{2.15}$$

Equation 2.15 has a solution of the form:

$$\Phi(x) = C\sin(\omega x) + D\cos(\omega x) + E\sinh(\omega x) + F\cos(\omega x), \tag{2.16}$$

where $\lambda/\beta = \omega$ and C, D, E, F are constants. Applying the boundary conditions of equation 2.10 to equation 2.16 yields:

$$\Phi(0) = 0 \Rightarrow 0 = D + F, \tag{2.17}$$

$$\Phi'(0) = 0 \Rightarrow 0 = C + E, \tag{2.18}$$

$$\Phi''(L) = 0 \Rightarrow 0 = -C\sin(\omega L) - D\cos(\omega L) + E\sinh(\omega L) + F\cosh(\omega L), \tag{2.19}$$

$$\Phi'''(L) = 0 \Rightarrow 0 = -C\cos(\omega L) + D\sin(\omega L) + E\cosh(\omega L) + F\sinh(\omega L), \tag{2.20}$$

Substituting equations 2.17 and 2.18 into 2.19 and 2.20 gives us:

$$0 = -C(\sin(\omega L) + \sinh(\omega L)) - D(\cos(\omega L) + \cosh(\omega L)), \tag{2.21}$$

$$0 = -C(\cos(\omega L) + \cosh(\omega L)) - D(-\sin(\omega L) + \sinh(\omega L)). \tag{2.22}$$

Equations 2.21 and 2.22 imply:

$$0 = -\sin^2(\omega L) + \sinh^2(\omega L) = \cos^2(\omega L) + 2\cos(\omega L)\cosh(\omega L) + \cosh^2(\omega L), \tag{2.23}$$

after simplifying,

$$0 = \cos(\omega L)\cosh(\omega L) + 1. \tag{2.24}$$

Equation 2.24 is our characteristic equation for this differential eigenvalue problem. Solution for ω in equation 2.24 develops unique solutions to the spatial domain of this problem. The first three such solutions are approximately $\omega_1 \approx 1.8751, \omega_2 \approx 4.6941, \omega_3 \approx 7.8548$.

With AFM cantilevers, sometimes considering the mass of the tip is important. When considering the tip mass to be significant the following boundary condition is changed:

$$\Phi'''(L) + \frac{M}{m}\omega^4\Phi(L) = 0 \tag{2.25}$$

where M is the is the magnitude of the mass. The change of boundary condition displayed in eqn. 2.25 only affects the characteristic eqn. 2.24 and changes it to:

$$C\omega^3[-\cos(\omega L) - \cosh(\omega L) - \frac{\sin(\omega L) + \sinh(\omega L)}{\cos(\omega L) + \cosh(\omega L)}(\sin(\omega L) - \sinh(\omega L)) \tag{2.26}$$
$$+\frac{M}{m}\omega(\sin(\omega L) - \sinh(\omega L) - \frac{\sin(\omega L) + \sinh(\omega L)}{\cos(\omega L) + \cosh(\omega L)}\cos(\omega L) - \cosh(\omega L))] = 0$$

The rest of the discussion with be in the context of cantilevers without lumped mass, however should still apply will substituting eqn. 2.26 for the characteristic eq.

Using equations 2.21 and 2.22 we have:

$$D = -\frac{\sin(\omega L) + \sinh(\omega L)}{\cos(\omega L) + \cosh(\omega L)}C, \tag{2.27}$$

thus using equations 2.16, 2.17, 2.18, and 2.27 we have:

$$\Phi_i(x) = C_i[\sin(\omega_i x) - \sinh(\omega_i x) - \frac{\sin(\omega_i L) + \sinh(\omega_i L)}{\cos(\omega_i L) + \cosh(\omega_i L)}\cos(\omega_i x) - \cosh(\omega_i x)], \tag{2.28}$$

for the i_{th} solution to the characteristic equation. We can use C_i to change equation 2.28 in to a form most suitable to our needs. To that end, we will choose C_i such that,

$$\int \Phi_i^2(x)dx = 1. \tag{2.29}$$

It will be useful at this point to define the inner product of two functions f and g as,

$$\langle f, g \rangle = \int_0^L f g dx. \tag{2.30}$$

In order to solve for the temporal dependence of this model we will multiply both sides of equation 2.9 by $\Phi_j(x)$, taking the inner product of both sides, and using the orthogonality property of $< \Phi_i(x)\Phi_j(x) >= \delta_{ij}$ yields:

$$\ddot{\eta}_i(t) + 2\zeta\omega_i\dot{\eta}_i(t) + \omega_i^2\eta_i(t) = F_i(t), \tag{2.31}$$

where $F_i(t) =< f(x,t)F_i(x)dx >$, note: $\frac{d^4\Phi_i(x)}{dx^4} = \omega_i^4\Phi_i(x)$. With equations 2.28 and 2.31 we are able to mathematically solve this model with solutions of the form:

$$u(x,t) = \sum_1^\infty \Phi_i(x)\eta_i(t), \tag{2.32}$$

with $\Phi_i(x)$ representing the i^{th} modal function and $\eta_i(t)$ representing the i^{th} modal coefficient. The modal functions provide a basis for a "Cantilever Space" to describe the motion of a cantilever. Then the modal coefficients yield a position in this "Cantilever Space." In other words the modal function act like basis vectors and the modal coefficients tell us what multiple of these basis vectors describe a given deflection.

2.2.2 Tip-sample interaction models

Solving Eq. 2.31 and hence the original beam dynamics equation 2.9 to obtain the cantilever displacement U(x, t) requires that the tip-sample interaction force to be known. Tip-sample interactions are usually parametrized to account for different types of forces. The simplest tip-sample interaction is an elastic response that can be modeled by Hooke's law (36):

$$f_{spring} = -k(u - h) \tag{2.33}$$

where $u - h$ is the distance the cantilever tip has pressed into the sample (h is the datum), and k is the local stiffness of the sample. More complex materials respond in a visco-elastic manner, dissipating some of the energy of the tip-sample interaction (16). This response is

modeled using a spring-damper system given by:

$$f_{spring-damper} = -k(u-h) - \zeta_s \dot{u} \tag{2.34}$$

In the following paragraphs, complex models will be developed. The primary references for these paragraphs are from works by Israelachvili (37), Johnson (38), and an online class by Reifenberger and Raman (39). Elastic contact is commonly modeled using Hertzian contact mechanics (40) and has successfully match with experiments for low surface energies (41). For our model we use the contact between two spheres of radii R_{tip} the radius of the tip and $R_{sample} = \infty$ treating the sample as an infinite half-plane. Using the geometry as specified results in an equilibrium equation of:

$$w_{tip} + w_{sample} = \delta - \frac{1}{2R_{tip}} r^2 \tag{2.35}$$

where w's are the constrained displacements of points within the contact radius, δ is the summation of the displacement of distant points of the two spheres and r is the distance from the center of contact. Hertzian pressure is given by:

$$p = p_0 \sqrt{1 - (\frac{r}{a})^2} \tag{2.36}$$

where p_0 is the maximum pressure and a is the contact radius. What is currently unknown in eqn. 2.35 are the values for w's. Seeking an expression for w's we use eqn. 2.36 with the law of cosines to describe the affect of pressure from one point on another resulting in:

$$p(s,\phi) = \frac{p_0}{a} \sqrt{a^2 - r^2 2rs\cos(\phi) - s^2} \tag{2.37}$$

$$w = \frac{1-\nu^2}{\pi E} \int\int p(s,\phi) ds d\phi \tag{2.38}$$

$$\Rightarrow w = \frac{1-\nu^2}{E} \frac{\pi p_0}{4a} (2a^2 - r_2) \tag{2.39}$$

where ν is the Poisson's ratio of the material, s is the radius from one point to another, and ϕ is the angle from one point to another. Substituting eqn. 2.39 into eqn. 2.35 yields:

$$\frac{\pi p_0}{4aE^*}(2a^2 - r^2) = \delta - \frac{1}{2R_{tip}} r^2 \tag{2.40}$$

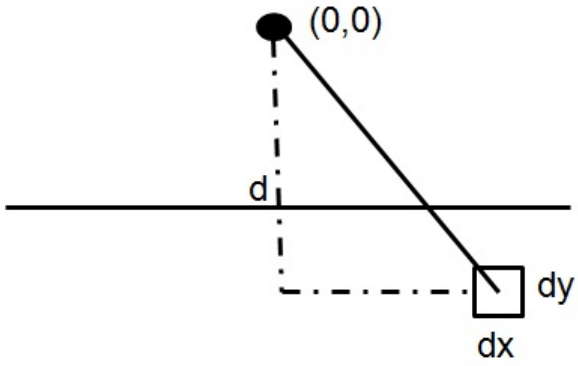

Figure 2.4 Diagram of tip point interaction with infinitesimal volume of the sample.

with $E^* = (\frac{1-\nu_{tip}^2}{E_{tip}} + \frac{1-\nu_{sample}^2}{E_{sample}})^{-1}$, $\delta = \frac{\pi p_0 a}{2E^*}$, and $a = \frac{2R_{tip}\pi p_0}{4E^*}$. To summarize and complete the model, the elastic deformation caused by the cantilever tip tapping on the sample leads to a force exchange which can be modeled when the tip and sample are in contact as:

$$f_{hertz} = \begin{cases} \frac{4E^* R_{tip}^{1/2}(h-u)^{3/2}}{3}, & \text{if } u \leq h \\ 0, & \text{if } u > h \end{cases} \tag{2.41}$$

where $E^* = (\frac{1-\nu_{tip}^2}{E_{tip}} + \frac{1-\nu_{sample}^2}{E_{sample}})^{-1}$, ν is Poisson's ratio, R_{tip} is the cantilever tip radius, and h is the location of the top of the sample.

To add adhesion to Hertzian contact, Van der Waal forces are added to create the DMT model(42) which applies to low adhesion situations and matches with experiments(43; 44). Van der Waal (VdW) forces add an attractive component due to atom-atom pair interactions. Like with the Hertzian model, the interaction between two spheres will be used. In the case of VdW forces the total potential will be developed from points. The potential of VdW has the form:

$$V(r) = \frac{-c}{r^6} \tag{2.42}$$

where r is the radius between the two points. Using the geometry set up in fig. 2.4, the potential of a point of the cantilever with the sample is:

$$V(r) = -2\pi\rho_{sample} c \int\int \frac{x}{(\sqrt{x^2+y^2})^6} dxdy \tag{2.43}$$

$$\Rightarrow V(d) = \frac{-\pi c \rho_{sample}}{6d^3} \tag{2.44}$$

where d is the shortest distance from the point to the sample. Integrating over the volume of the sphere representing the tip:

$$V(d) = \frac{-\pi c \rho_{sample} \rho_{tip}}{6d^3} \int_0^{-2R_{tip}} \frac{\pi(2R_{tip} - y)y}{(d+y)^3} dy \tag{2.45}$$

$$\Rightarrow V(d) = \frac{HR_{tip}}{6d} \tag{2.46}$$

where H is the hamaker's constant. Finishing of the formulation, the DMT model is as follows:

$$f_{dmt} = \begin{cases} \frac{4E^* R_{tip}^{1/2}(h+a_0-u)^{3/2}}{3} - \frac{HR_{tip}}{6a_0^2}, & \text{if } u < h + a_0 \\ -\frac{HR_{tip}}{6(h-u)^2}, & \text{if } u \geq h + a_0 \end{cases} \tag{2.47}$$

where H is the Hamaker constant and a_0 is the intermolecular distance.

When imaging in a humid ambient environment the affects of capillary forces need to be considered. This is due to a layer of water which forms on the surface of samples. Using the model from Kober et al.(45):

$$f_{cap} = \begin{cases} \frac{2\Delta E}{D_0^2}((h-u) - D_0), & \text{if } h + a_0 < u < D_0 \\ \frac{2\Delta E}{D_0^2}(a_0 - D_0), & \text{if } u < h + a_0 < D_0 \\ 0, & \text{if } u > D_0 \end{cases} \tag{2.48}$$

where ΔE is energy dissipation and D_0 is the height at which the water neck breaks from the cantilever tip.

For the samples which demonstrate viscoelastic behavior, the tip-sample interaction can be modeled using the Kelvin-Voigt (KV) model(46; 47). Following the idea of a spring and damper in parallel KV is modeled as:

$$f_{KV} = \begin{cases} 0, & u > h + a_0 \\ 0, & u \leq h + a_0 \text{ and } \dot{u} > 0 \\ -\eta_s \dot{u} \sqrt{R_{tip}(h-u)}, & u \leq h + a_0 \text{ and } \dot{u} < 0 \end{cases} \tag{2.49}$$

where η_s is the viscous damping coefficient of the sample.

2.3 Inverse problem

The inverse problem is defined as follows:

IP: *Given the cantilever properties, (E, I, μ), the measured cantilever tip deflection $U(x,t)$, and*

the given cantilever base displacement $y_{ex}t$, calculate the parametrized tip-sample interaction $f_{ex}(x,t)$.

One approach to solving the inverse problem is to convert it into an unconstrained optimization problem through the minimization of a chosen cost functional $\mathcal{J}$ that minimizes the difference between $U(x,t)$ and $u(x,t)$. An appropriate choice of the cost functional $\mathcal{J}$ acts as a metric that quantifies the mismatch between a guess value of the tip-sample interaction and the true tip-sample interaction. The choice of the cost functional plays a significant part in the accuracy and efficiency of the inversion process. The proper choice of the cost functional ensures reasonable speed of calculation and a smooth phase space. Extensive computational experiments suggested the use of the following cost functional:

$$\mathcal{J}_{L2}^2(F) = \int_0^{t_{max}} [U(x,t) - u(x,t)]^2 \, dt \tag{2.50}$$

where U is the experimentally measured tip deflection. u is the calculated tip deflection for given the tip-sample interaction F (i.e., by solving the forward problem **FP**).

The unconstrained optimization problem is posed as follows:
Given cantilever properties (E, I, μ), cantilever base movement $y_{ex}(t)$, and the experimental cantilever tip deflection $(U(x,t))$, find the parametrized tip-sample interaction F^* such that $\mathcal{J}_{L2}(F^*) \leq \mathcal{J}_{L2}(F)$ for any F, where $\mathcal{J}_{L2}$ is defined in Eqn. 2.50.

CHAPTER 3. High-speed GPU framework development

This section details the computational framework for solving the forward and inverse problems formulated in the previous chapter. A key challenge is the necessity of very fast force-inversion for real time dAFM imaging of soft samples to be possible. Posing the direct problem as a set of ODE's and the force inversion as an unconstrained optimization problem over these ODEs allows to leverage the computational advantages provided by GPUs. The rational for using GPU's is guided by the following reasons: (1) ability to construct a large set of forward problems in parallel; (2) faster analysis given faster memory accesses compared to CPUs; (3) GPU compute architecture is well suited for problems with minimal parallel dependencies; and (4) GPU compute architecture is well suited for problems where the computation-to-memory-access ratio is larger than one. A brief description of GPU hardware and CUDA software concepts utilized in the developed framework follows.

3.1 Introduction

Utilizing GPU's for computation is different than on CPU's. GPU's require a large number of threads of execution that are processed in parallel to be efficient. In contrast, CPU's are generally more efficient with few threads.

Memory: GPUs (running CUDA) have very large computation capability compared to the speed at which they can access memory. GPU's hide this memory latency by performing computation and memory grabs simultaneously. While sets of threads (called warps) are waiting for their data from memory, other warps get computed. The availability of hierarchies of memory allows significant room for designing algorithms to optimize memory access, thus enhancing speed. We briefly describe the memory modes that are used in the current formulation; *global,*

shared, texture, and constant.

1. Global memory is the main memory storage on GPU and is the slowest to access requiring hundreds of clock cycles. Global memory is retrieved in groups of bytes for warps based on the requirements of the threads. Warps can only grab memory that is in sequential order.[1]

2. Shared memory is a very fast, small block of memory (16 kb on compute capability 1.3 and below, up to 48 kb for compute capability 2.0) which is accessible only within each block of threads.

3. Texture memory is a cached global memory.

4. Constant memory can only be assigned by the CPU and is a cached read-only memory.

Given the finite memory resources and speed, memory management is critical as most GPU algorithms are limited by their memory throughput (48).

Computation: Through the use of CUDA architecture and programming tools, the management and control of GPU computation and data parallelism is possible[2]. In CUDA, threads are organized into blocks which are executed on the same streaming multi-processor (SMP). Each GPU only has a finite number of SMPs and as a result can only computer a finite number of blocks at the same time. SMPs execute threads in groups (or warps). Warps are chosen to be processed based on the availability of the requested memory resources. Thus, optimally choosing threads and threads per block can significantly enhance memory access and performance.

Communication: Any data dependency between threads requires special considerations. Shared memory is the best method of dealing with any data dependencies. This means that inter-thread communication is best handled within each block. Communication between blocks can occur through a global sync between all GPU and CPU threads but is very inefficient.

[1] So if thread 0 needs memory from array position 0 and thread 1 requires array position 1000000, the warp will request two accesses to global memory (costing several hundred clock cycles twice). Alternatively, if the memory in array position 1000000 was in position 1 instead, only one memory access would be required.

[2] CUDA provides a compiler and basic functions to perform computational tasks. The CUDA tool-kit also includes a best practices guide which describes the advantages and limitations of GPU computing and how to get the best performance.

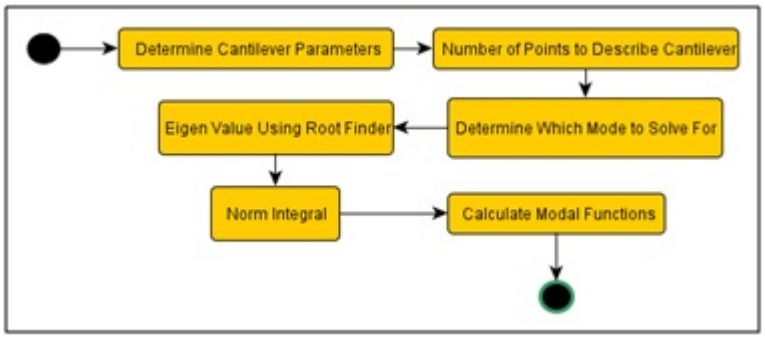

Figure 3.1 Flowchart of serial forward problem solver initialization process

Through the use of good parallel programming practices, the shared memory can be used efficiently to communicate between threads.

We utilize CUDA programming to implement both the forward and inverse problems on GPU's. *Our approach takes advantage of the GPU compute structure by designing an algorithm which minimizes data dependency between threads, maximizes the number of computations per global memory access, and minimizes CPU/GPU data communication.* The optimization problem approach allows for minimal data dependency through the solving of many forward problems, that are independently solved on multiple threads on the GPU.

3.2 Forward Problem

Fast calculations of solutions to Eqn. 2.31 are achieved through a CUDA based high-order ODE solver. The three key stages in solving the forward problem (FP) are: initialization (see fig. 3.1), including memory set-up and modal function calculation; calculation of modal coefficients via high-order ODE solvers; and using modal functions and modal coefficients to calculate displacements.

Computational issues A Newton root solver and Simpson integration modules are used to solve for the modal resonance frequency (ω_i) and normalization factor (to make the modal functions orthogonal), respectively. Calculating the modal functions using the hyperbolic trigonometric form generally presented in texts (34) causes over-flow errors for higher modes. We recast the calculation to the following equivalent exponential form to enable accurate cal-

culation without overflow:

$$\Phi_i(x) = \frac{2\sin(\beta_i x) - 2G\cos(\beta_i x) + (G-1)\exp^{\beta_i x} + (G+1)\exp^{-\beta_i x}}{2}, \tag{3.1}$$

$$G = \frac{2\exp^{-\beta_i L}\sin(\beta_i L) - \exp^{-2\beta_i L} + 1}{2\exp^{-\beta_i L}\cos(\beta_i L) - \exp^{-2\beta_i L} + 1}, \tag{3.2}$$

where β_i is the i^{th} solution of $cos(\beta L)cosh(\beta L) + 1 = 0$ and $\omega_i = \beta_i^2\sqrt{\frac{EI}{\mu L^4}}$.

Note that ω_i^2 grows very quickly with increasing mode order i, making Eqn. 2.31 a stiff ODE. We utilize an explicit 4^{th} order Runge-Kutta (RK) or an implicit first order Euler method to solve the ODEs.[3] Explicit schemes were tested because they are efficient and the RK fourth order method converged in the range of time steps typical for AFM experiments. Implicit first order Euler was tested and found to give accurate solutions with less computational overhead that explicit RK fourth order when the implicit system was solved by hand and hard coded.

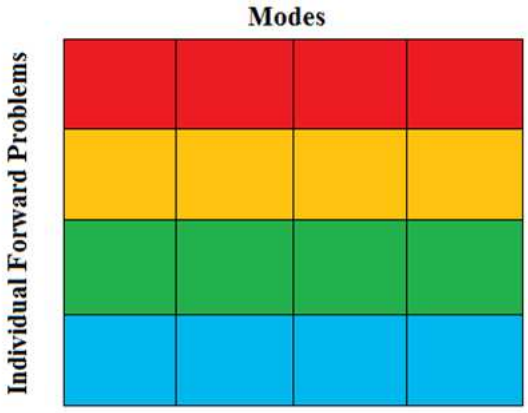

Figure 3.2 Memory Allocation and Parallelization: A schematic of the forward problem parallel algorithm. Squares represents an individual thread. Each row represents an individual forward problem. Each mode of each forward problem is given a thread. Thus, each block will solve several forward problem solutions simultaneously.

Memory allocation Solutions of the forward problem are obtained by calculating the modal functions (Φ_i) and the modal coefficients (η_i). The modal functions remain invariant and

[3]First order explicit Euler method was tested but failed to converge to a solution with practical time-step size of greater than 10^{-8}. Second or third order methods have not been tested and could be a possible method of reducing calculation time if they converge.

thus only need to be calculated once. We calculate and store the modal functions serially as part of the initialization process. Two parallel strategies are possible to assist in solving Eq. 2.31 for $\eta_i(t)$: *parallel prefix for force inner products and direct parallelism across modes.* Moreover, assuming that the force is variable separable (into spatial and temporal components), the spatial component of the force can be determined during initialization. This allows converting the computation of integrals involved on the RHS of Eqn. 2.31 to a one time calculation. Pre-calculating force integrals leaves implementing a framework with parallelism across modes. Every modal coefficient ODE solve is handled by a unique thread. Thus, two dimensional blocks of threads are set-up as (m, n_{fpb}), where m is the number of modes and n_{fpb} is the number of forward problem solutions per block, as shown in Fig. 3.2. For example, on compute capability 1.3 GPUs, using eight modes, the current parallel framework can run up to 64 problems per block.

We next analyze the memory complexity of the framework. This elucidates the rational for deploying the various data structures in the available memory hierarchies. The major memory needs are as follows:

- The cantilever parameters, E, I, μ. Since the cantilever parameters are assigned as part of the initialization process and require little memory, they are a good choice for constant memory.

- Parameters of the tip-sample interaction. This requires 4 bytes for each parameter, thus requiring $8n_{fps}$ bytes for n_{fps} simultaneous forward problem solutions using the visco-elastic model Eqn. 2.34. Interaction parameters are left in the global memory since they have to be optimized in the inverse problem[4].

- Modal coefficients η_i require $4mn_{fps}$ bytes, where m is the number of modes used. The output deflection points require $n_{dp}n_{fps}$ bytes, where n_{dp} is the number of deflection points computed. Usually deflections at three points on the AFM cantilever are measured ($n_{dp} = 3$). The modal coefficients and deflection points are stored in shared memory because of the constant updating during the forward solve.

[4]The access speed of the interaction force parameters could potentially be improved by utilizing texture memory but has not been implemented in this work.

3.3 Forward problem runtime analysis

Complexity analysis For the forward problem, the main calculations affecting runtime complexity are the force integral calculations and the ODE solve. The runtime complexity for the serial forward problem is $\mathcal{O}([n_t n_x + n_t]m)$, where m is the number of modes, n_t is the number of time steps and n_x is the number of spatial points used to compute the force integral. However, assuming variable separation of the forces reduces serial runtime complexity to $\mathcal{O}(mn_t)$. By deploying across m threads on a GPU, the parallel runtime complexity is $\mathcal{O}(n_t)$.

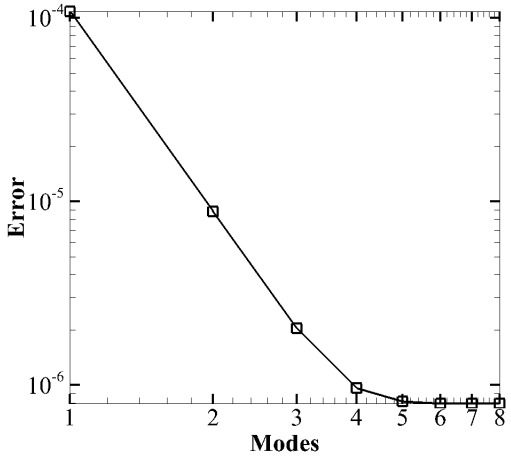

Figure 3.3 Convergence compared to 8 modes for simple sinusoidal base movement

3.3.1 Speedup characteristics: Comparing CPU vs GPU implementation

In addition to the GPU based implementation discussed in the preceding section, we also implemented a CPU based version of the ODE solver for comparison. The validity of the results of both implementations are ensured by comparing with analytical solutions that are obtained using a sinusoidal base movement and constant forces (34)(see Appendix A).

We discuss runtime trends and accuracy details in this subsection. Each forward problem is run for 1000 time steps (unless otherwise stated) using time-step, $\Delta t = 10^{-7}$. A maximum of eight modes are used. Eight modes can satisfactorily track the cantilever evolution (with an error of 3.310^{-6}). Error is defined as $\frac{|V_t - V_c|_{L2}}{|V_t|_{L2}}$, where V_t is the true value and V_c is computed. In testing convergence, the L2 error of various modes are compared to the solution obtained

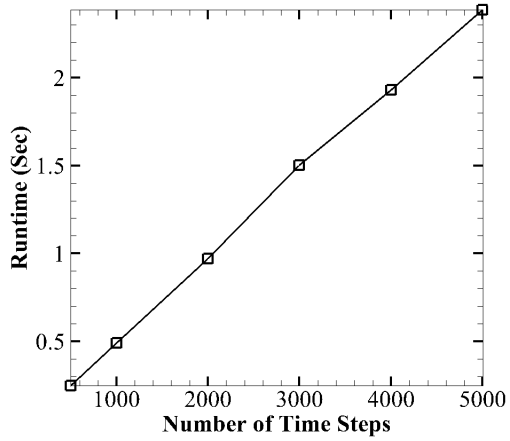

Figure 3.4 Number of time points vs. runtime in seconds for CPU framework

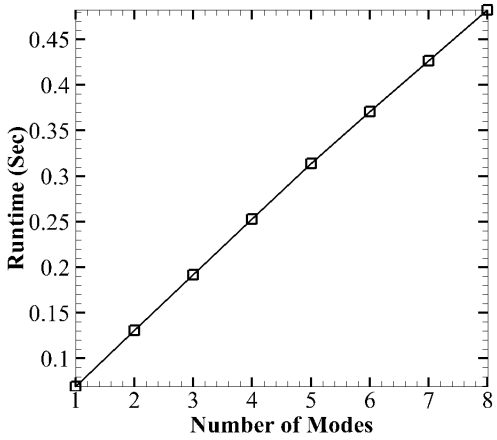

Figure 3.5 Number of modes vs. runtime in seconds for CPU based framework

using eight modes. Fig. 3.3 shows the plot of error versus number of modes. For this simple case, four modes are sufficient for resolving a sinusoidal deflection.

While considering runtime complexity three parameters are most dominant; number of modes, number of points used to describe modal functions, and number of time points. We analyze all three of them independently, first for the CPU based implementation and subsequently for the GPU based implementation. Runtime as a function of the number of modes is shown in Fig. 3.5 with the number of time points fixed at 1000 (note that 8 modes take 0.482 seconds). Fig. 3.4 plots runtime as a function of the number of time points. When seeking real-time inversions, runtime must be of the order of few hundred milliseconds. These runtime

analyses show that calculation times using CPUs are too slow to solve the inverse problem in real time.

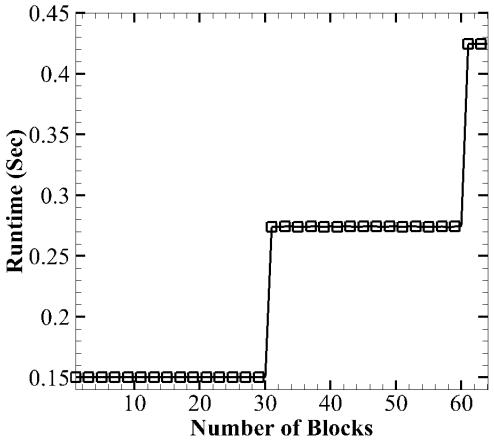

Figure 3.6 Number of blocks vs. runtime in seconds

In contrast to the CPU results, the GPU based results are very promising. With the goal of solving several hundreds of forward problems to solve one inverse problem, we utilize two metrics to illustrate the capabilities of the GPU based framework: (a) runtime of individual forward solves; (b) the number of forward solves which can be calculated in parallel. Both metrics depend on multiple factors: number of time steps (n_t), number of modes (m), number of solutions per block (n_{fpb}), and total number of forward problem solutions (n_{fps}). A Nvidia Quadro FX 5800 GPU is used which limits the number of blocks that can run in parallel to 30 (2 blocks per streaming multiprocessor(SMP), 15 SMP's)[5]. Beyond this, with all SMPs filled, the blocks have to wait for an open SMP. This can be clearly seen in Fig. 3.6 where the runtime jumps after 30 and 60 blocks. Fixing the number of blocks ($\frac{n_{fps}}{n_{fpb}}$) to 30 and modes (m) to 8, runtime is analyzed by varying the number of time steps and number of solutions per block. Fig. 3.7 show the effect of n_{fpb} on runtime. As the number of solutions per block increases, the runtime increases in a non-linear way. Using 32 solutions per block most efficiency utilize GPU resources giving forward solves with 100 millisecond runtime. Furthermore, as n_{fpb} increases beyond 32, the solutions per second gain is small while 32 solutions only requires half of the shared memory. Fig. 3.8 shows that increasing n_t causes a linear increase in runtime (as

[5]For more details on the system used for testing and implementation, see appendix C.

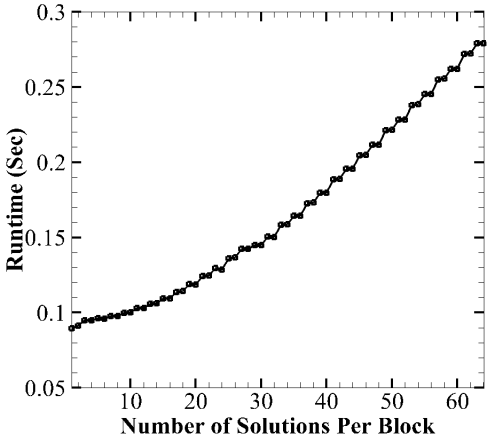

Figure 3.7 Number of solutions per block vs. runtime in seconds

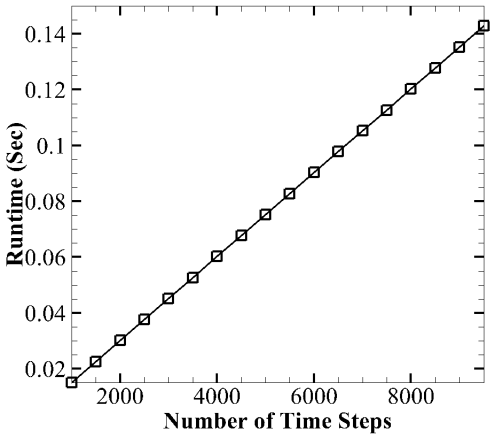

Figure 3.8 Number of time points vs. runtime in seconds

predicted by the complexity analysis in the previous section).

By appropriately choosing the number of solutions per block to ensure proper memory allocation, the runtime for a single forward solve on the GPU took 0.0151 seconds. *This is a speed-up of 32 over the CPU framework.* More importantly, the GPU based framework can execute several forwards solves simultaneously. **960 forward solves are computed in 0.0151 seconds, translating to an effective speed-up of 30000**. Effective speed-up is the direct comparison between our CPU (single-core) performance and GPU performance.

3.4 Inverse problem

The choice of the optimization algorithm is driven by the following constraints:

1. The GPU based forward solver implementation is able to compute several forward solution in parallel.

2. The parametrization of the cost function $\mathcal{J}$ may be high dimensional.

3. Furthermore, the landscape of $\mathcal{J}$ may be non-smooth, necessitating a gradient free method.

4. The existence of multiple local minima that have to be discarded.

A gradient-free, global search algorithm that satisfies these constraints is the particle swarm optimization (PSO) (49). PSO finds the global minima by starting with a large population of candidate solutions (or particles), and moving these particles around in the search-space according to certain rules over the particle position and velocity. Each particle is influenced by its local best known position and the best known positions in the search-space, which are updated at every iteration as better positions are found by other particles. Generally, particle locations and velocities are chosen using a uniform distribution in the search space (50; 51). The update of velocity uses the following equation:

$$v_i = wv_{i-1} + c_1 r_1 (b_l - x_{i-1}) + c_2 r_2 (b_g - x_{i-1}), \tag{3.3}$$

where v_i is the velocity at iteration i, w, c_1, and c_2 are weighting factors, r_1 and r_2 are random numbers, b_l and b_g are the local and global bests respectfully, and x_i is the position of the particle at iteration i. Recent theoretical results suggest that an appropriate choice of the parameters guarantee convergence (50; 51). We utilize an optimized GPU based implementation of the PSO algorithm (52). In an attempt to save communication costs, the framework developed by (52) uses a ring topology (ie. particles only communicate best information to their neighbours in memory), resulting in a delay in all particles knowing the global best information. To further demonstrate how PSO works an example is presented. For this example we will use PSO on the function $f(x) = x^2$ with two particles looking for a minima. These two particles

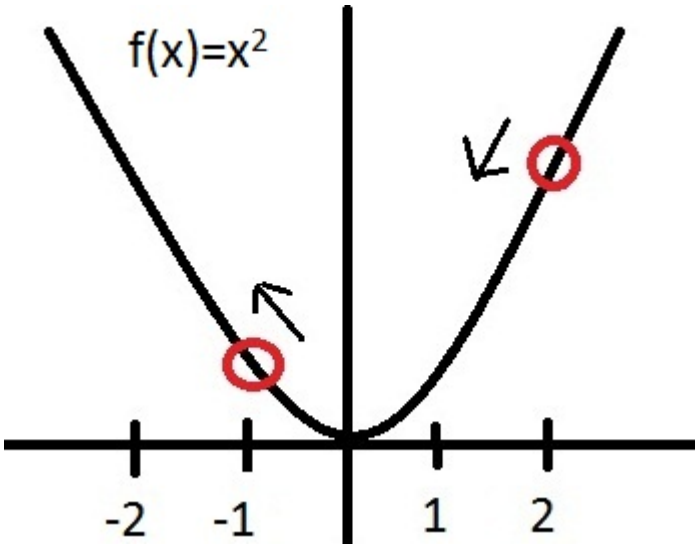

Figure 3.9 PSO example initial set-up. One particle is "randomly" positioned at $x = -1$ with a negative velocity, the other particle is "randomly" placed at $x = 2$ with a negative velocity.

would normally start at a random location but for simplicity let us assume the particles are "randomly" placed at $x = -1$ and $x = 2$. Still further let us assume that the random initial velocities are -1 and -1 respectively. This initial set-up of the particles locations and velocities is shown in fig. 3.9. For eqn. 3.3 we will use parameter values as follows for simplicity, $w = 0.33$, $c_1 = 0.33$, and $c_2 = 0.33$. A typical set of generations of PSO using the details just specified goes as follows:

Particle 1	Particle 2	Global Best
$x = -1, y = 1$	$x = 2, y = 4$	$x = -1$
$x = -2, y = 4$	$x = 1, y = 1$	$x = -1, 1$
$x = -2.07326, y = 4.29841$	$x = 0.66667, y = 0.4489$	$x = 0.66667$
$x = -1.88529, y = 3.55433$	$x = 0.55777, y = 0.31111$	$x = 0.55777$

The process of successive generations will continue until a specified goal value for cost function (or functional) value is met.

CHAPTER 4. Results

4.1 Introduction

This section discusses a hierarchy of increasingly complex inverse problems. We start with the simpler problem of extracting the base vibration characteristics given tip deflections. This problem also explores the choice of the search space parameters and their effect on runtime. The next two subsections deal with real time inversions of elastic and visco-elastic tip-sample interactions, followed by inversions of more complex tip-sample interactions.

4.2 Forced vibration

'Experimental' tip deflection data was computed by forcing the base to vibrate to a simple sinusoidal driven signal:

$$y(t) = a\ sin(2\pi f t), \tag{4.1}$$

where a is the amplitude and f is the frequency of cantilever base vibration. The 'experimental' tip deflection was obtained by setting $a = 2nm$ and $f = 25600\ Hz$. This 'experimental' tip deflection – subsequently used to drive the inverse problem – is obtained using the CPU based serial framework, thus resolving the issue of inverse crime (53).

We analyze the performance of the inversion framework by starting with a one dimensional search space for the amplitude of base vibrations and fixing $f = 25600\ Hz$. We provide physically meaningful bounds on the amplitude, $[0\ nm, 100\ nm]$ and set $n_t = 1000$. The number of time steps was chosen to provide a sufficient number of data points per oscillation period given the typical time-step used. Fig. 4.1 shows that the mean and variance of the number of generation for convergence shrinks as the number of particles used in the PSO

increases. Faster convergence is expected since the density of the particles in the search space is increasing. Given that the framework produces a generation about every 16 milliseconds, *the mean convergence time for the 1D case is just over 48 milliseconds for 512 particles.*

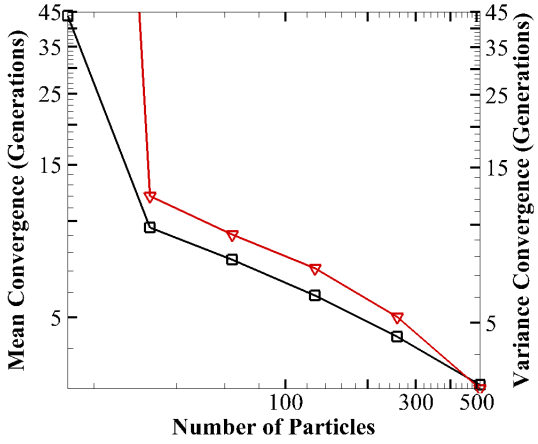

Figure 4.1 Number of particles effect on mean(□) and variance(∇) of generations to convergence with 1D search space

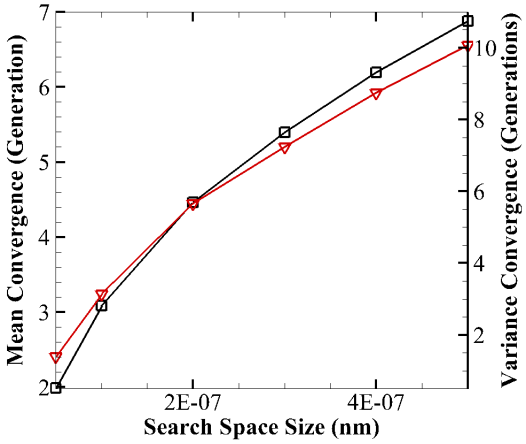

Figure 4.2 Search Space effect on mean(□) and variance(∇) of generations to convergence with 1D search space

We next investigate the affect of search space size on the performance of the inversion framework, as shown in Fig. 4.2. In general, as the particle density in the search space decreases, the mean number of generations required for convergence increases. For the case of the largest search space ($[0\ nm, 500\ nm]$) the mean number of generation of 6.8767 corresponds

to a runtime of around 112 milliseconds. *This suggests that a tight bound on the unknown force parameters can significantly decrease inversion times.*

The affect of convergence cutoff on convergence is shown in Fig. 4.3. The general trend of increasing accuracy requirements raises the mean number of generations required for convergence. To achieve an error of 0.001 requires an average of 7.99 generations (taking 128 milliseconds) verses the 48 milliseconds required for achieving a relative error of 0.01. The effect of n_t on mean generations was also tested. The number of time steps was varied from 1,000 to 10,000 and did not show any effect on the result.

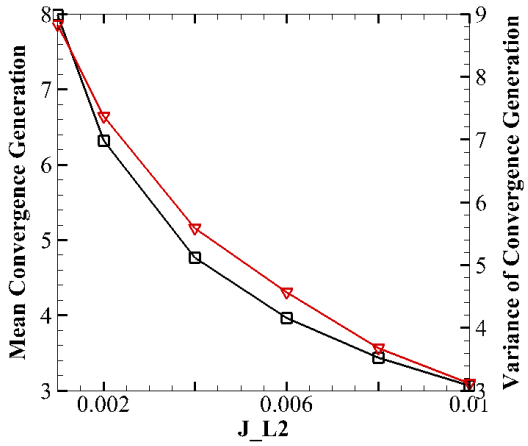

Figure 4.3 $\mathcal{J}_{L2}$ target effect on mean(□) and variance(∇) of generations to convergence with 1D search space

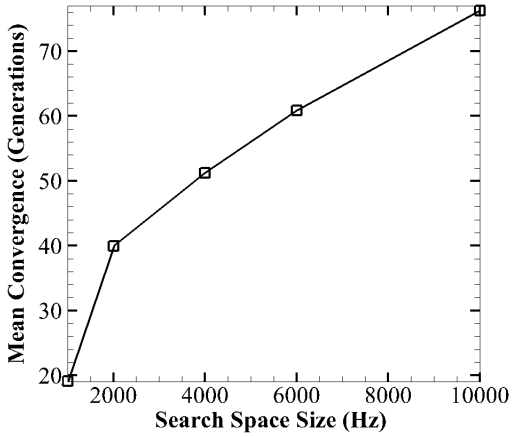

Figure 4.4 Search Space effect on mean(□) of generations to convergence with 2D search space

We next tested the inversion framework by using a 2D search space (making both a and f unknown). Fig. 4.4 shows a similar trend to the 1D case where the mean number of generations required for convergence increases as the search space increases. With a search space of 1000 Hz, 19.0458 generation on average is required for convergence resulting in a compute time of about 320 milliseconds.

To explore frequency dependent properties, the AFM can be driven by a chirp base vibration. This linearly varies the frequency of base oscillation with time:

$$y(t) = a\ sin(2\pi[f + gt]t), \tag{4.2}$$

where g is defined as the frequency gain parameter. We utilize the inversion framework to extract the parameters of this chirp signal. Matching a chirp base movement requires exploring a 3D search space for (a, f, g). The 'experimental' tip deflection was obtained by setting $a = 2\ nm$, $f = 10,000\ Hz$ and $g = 20,000,000\ \frac{Hz}{Sec}$. Using a search space of $[0\ nm, 100\ nm] \times [7,500\ Hz, 12,500\ Hz] \times [19,000,000\ \frac{Hz}{Sec}, 21,000,000\ \frac{Hz}{Sec}]$, the framework extracts correct values of (a, f, g) in 14.9 generations corresponding to a runtime of about 240 milliseconds.

4.3 Spring tip-sample interaction

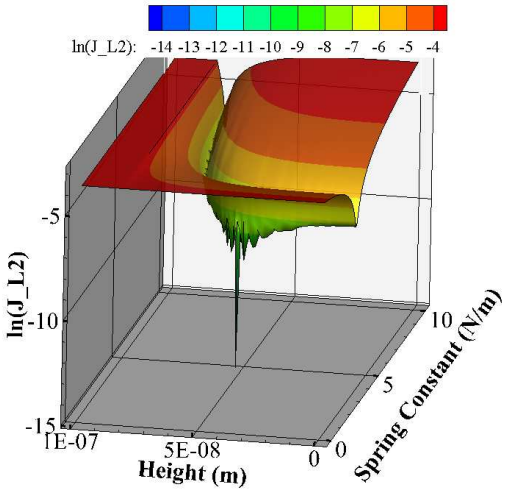

Figure 4.5 $\mathcal{J}_{L2}$ Space of spring and cantilever-sample separation

A simple yet extensively used model for tip-sample interaction is one that assumes an elastic response of the soft sample. Extracting spatial variation in elasticity is important for a variety

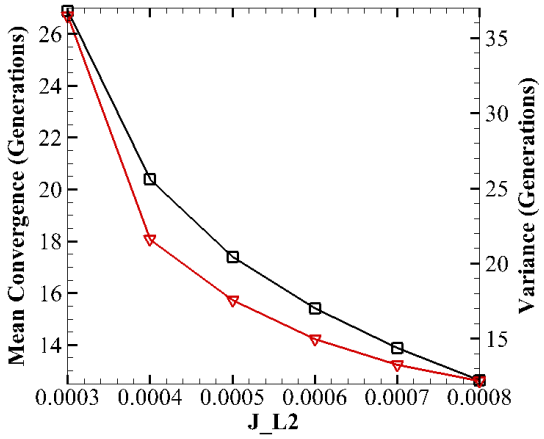

Figure 4.6 $\mathcal{J}_{L2}$ requirement effect on mean(□) of generations to convergence with spring tip-sample search space

of applications in addition to non-destructive scanning of the sample (e.g., tumors have higher stiffness that normal cells; as collagen dehydrates there is a change in elasticity (5; 15)).

We utilize a Hooke's law based parametrization for the elastic response of the sample. The modelling of the tip-sample interaction as a spring is in-line with most AFM experimental force calculation models. This model has the following form:

$$f(t) = -k(u - h) \tag{4.3}$$

where $u - h$ is the distance the cantilever tip has pressed into the sample and k is the tip-sample spring constant. With the base movement parameters known, only two tip-sample force parameters are unknown, the cantilever-sample separation and sample spring constant. Cantilever-sample separation is defined as distance from the cantilever's neutral axis to the sample. The 'experimental' tip deflection was created using the following parameters: $h = 50\ nm$, and $k = 4N/m$, where h is the cantilever-sample separation and k is the sample spring constant. These parameters were chosen to mimic results found in (16).

Calculating the force parameters for a spring sample is computationally complex due to the non-convex nature of the phase space as shown in Fig. 4.5, and results in more calculation time being required shown in Fig. 4.6. Fig. 4.5 shows that several combinations of k and h produce similar values, creating a symmetrical valley near the basin of the global minima. The resulting

valley of similar combinations is expected because both k and h only affect the amplitude of vibration and not the phase. Using a cutoff convergence threshold of $\mathcal{J}_{L2} = 0.0005$ results in a 5% error in calculating k and h resulting in a mean convergence of about 18 generations, resulting in an average runtime of 288 milliseconds.

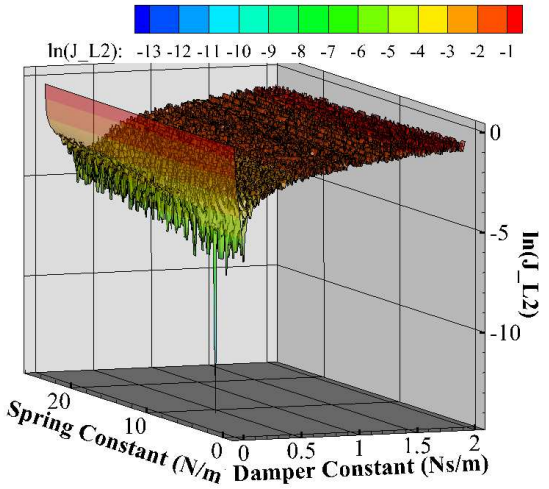

Figure 4.7 $\mathcal{J}_{L2}$ Space of spring and damper

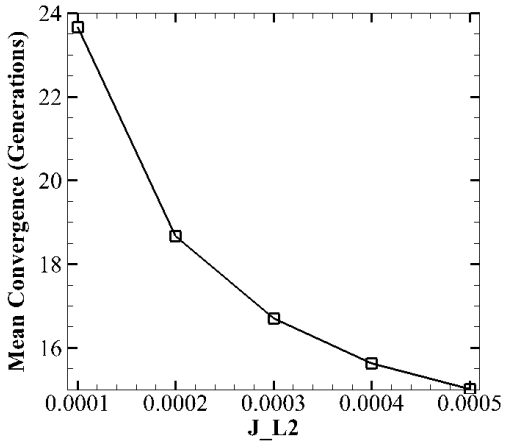

Figure 4.8 $\mathcal{J}_{L2}$ requirement effect on mean(□) of generations to convergence with spring/damper tip-sample search space

4.4 Spring/damper tip-sample interaction

We increase the complexity of the tip-sample interaction parametrization by next assuming a visco-elastic response of the sample. Understanding the visco-elastic variations is particularly important in understanding aspects of polymer physics(46). A Kelvin-Voigt parametrization is extensively used to model visco-elastic tip-sample interactions (16). This is essentially a spring damper system:

$$f(x,t) = -k(u-h) - \zeta_s \dot{u} \tag{4.4}$$

where ζ_s is the viscous damping coefficient. We assume that the sample separation for the AFM cantilever neutral axis is known. This inversion consists of estimating two material properties. The 'experimental' tip deflections were obtained by using the following parameters: $h = 50\ nm$, $k = 4N/m$, and $\zeta_s = 0.1\frac{Ns}{m}$. Using a search space of $[0\ N/m, 25\ N/m] \times [0\ \frac{Ns}{m}, 2\ \frac{Ns}{m}]$ the convergence profile with increasing number of generations of the PSO scheme is shown in Fig. 4.8.

The phase space for inverting the visco-elastic tip-sample interaction is highly corrugated and has multiple local minima as shown in Fig. 4.7. The shape of search space clearly demonstrates the need for gradient free optimization methods. The large number of local minima is due to the phase change caused by the damping force. The average runtime for inversion (with 16 generations) was 256 milliseconds.

4.5 Hertzian contact model

For the first tip-sample model that is applied in practice we present the Hertzian contact model (40). Hertz contact provides a realistic model for tip-sample interactions which do not include adhesion (41). The model is:

$$f(t) = \begin{cases} \frac{4E^* R_{tip}^{1/2}(h-u)^{3/2}}{3}, & \text{if } u \leq h \\ 0, & \text{if } u > h \end{cases} \tag{4.5}$$

where $E^* = (\frac{1-\nu_{tip}^2}{E_{tip}} + \frac{1-\nu_{sample}^2}{E_{sample}})^{-1}$, ν is Poisson's ratio, R_{tip} is the cantilever tip radius, and h is the location of the top of the sample. Our approach for this inversion is to estimate

the equivalent elasticity E^* . To create the 'experimental' deflections, UHMWPE and PDMS are used as sample models to simulate real data. The important data for this model are UHMWPE: $\nu = 0.4, E = 0.104$ GPa (54); PDMS: $\nu = 0.5, E = 750$ kPa (55; 56); Other data: $h = 81.5\ nm$, $a = 2\ nm$, and $f = 25000\ Hz$. Using a search space of $[0, 100E^*]$ runtime tests for UHMWPE and PDMS were preformed for various accuracy criteria. A quick note on the determination of maximum error of calculated material properties. To determine maximum error, a cost functional data set is produced similar to previous plots (see fig. 4.7) and a level set of parameter values is evaluated using our standard for error calculation, the maximum error is the parameter values who produce the largest error in that level set.

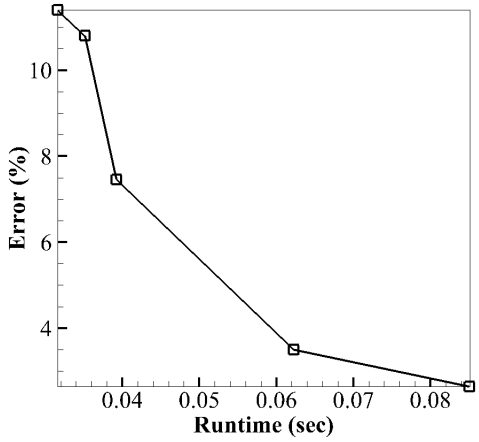

Figure 4.9 Runtime vs accuracy in calculating E^* for UHMWPE

UHMWPE As shown in fig. 4.9, equivalent elasticity for UHMWPE presents no difficulty for the computational framework to calculate accurate solutions very quickly. For calculating the most accurate reconstruction of E^* with an error of less than 2.7%, mean calculation time does not exceed 90 milliseconds.

PDMS Figure 4.10 shows the E^* results for PDMS. Like with UHMWPE, E^* is able to be calculated quickly and accurately with the lowest error maximum error for this test of 1% and a mean runtime of less than 80 milliseconds.

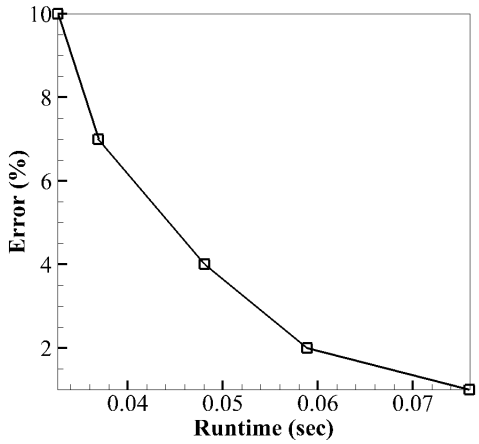

Figure 4.10 Runtime vs accuracy in calculating E^* for PDMS

4.6 DMT model

Adding a layer of complexity to the Hertz contact model we present work using the DMT contact model. DMT adds adhesion to the Hertzian contact through the addition of Van der Waal(VdW) forces(42) which applies to low adhesion situations and matches with experiments(43; 44). The DMT model is as follows:

$$f(t) = \begin{cases} \frac{4E^* R_{tip}^{1/2}(h+a_0-u)^{3/2}}{3} - \frac{HR_{tip}}{6a_0^2}, & \text{if } u < h + a_0 \\ -\frac{HR_{tip}}{6(h-u)^2}, & \text{if } u \geq h + a_0 \end{cases} \tag{4.6}$$

where H is the Hamaker constant and a_0 is the intermolecular distance. This particular inversion has presented some difficulties and as a result will need further explanation of the details of the process that was used. As a result of these difficulties, only data for PDMS will be shared. When approaching the DMT inversion there are three parameters which need to be estimated, equivalent elasticity E^*, Hamaker constant H, and intermolecular radius a_0. The important data for creating the 'experimental' data PDMS: $H = 4.4e-20$, $a_0 = 0.165nm$; Other data: $h = 5.2\ nm$, $a = 2\ nm$, $f = 20000\ Hz$, and number of time points to analyze $n_t = 30000$. Note the changes in forcing frequency and the number of time points analyzed. The frequency was changed to highlight the VdW affects. As a result the span of time in-which the force affects need to be considered increased for this particular scenario, resulting in increased calculation times. To minimize this increase in calculation time the following procedure for search was

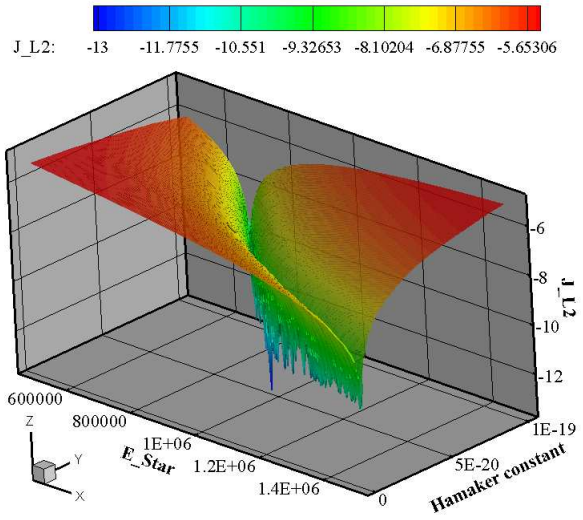

Figure 4.11 Hamaker's constant and equivalent elasticity search space

followed:

1. Determine a_0.

   Set H constant to zero.

   Take a reasonable guess for E^* and set search space accordingly.

   Set cost functional threshold accordingly for accuracy.

2. Determine H and E^* simultaneously.

   Set a_0 search space to above calculated value.

   Set cost functional threshold accordingly for accuracy.

Firstly approaching step one with a guess equivalent elasticity of $E^* = 9999946.5$ and search space of $[0.1a_0, 10a_0]$. With a maximum error of 1% the mean inversion time was 9.85 seconds. Given the linear runtime scaling, if the number of time points analyzed could be reduced to 1000 like with previous models, the runtime would reduce to approximately 0.328 seconds. Addressing the calculation of step two, the search space for H and E^* were set-up in a similar manner to a_0. Figure 4.11 shows the search space for various H and E^* values with fixed a_0. This search space presents a curved valley similar to the spring/height search space presented in fig. 4.5. The runtime results were tested for a maximum error of 1% and yielded a mean

calculation time of 22.22 seconds which would be approximately 0.741 seconds if the analysis could be reduced to 1000 time steps. In total, the mean calculation time for the DMT model for PDMS is 32.07 seconds.

CHAPTER 5. Conclusion and future work

5.1 Conclusion

The atomic force microscope (AFM) is a versatile, high-resolution scanning tool used to characterize topography and material properties of a large variety of specimens. Its applicability to characterize soft specimens like tissue and gels is currently constrained by the inability to appropriately control tip-sample interaction forces. A major bottleneck to control tip-sample interaction is the ability to extract these tip-sample forces in real time from the deflection signal. This paper illustrates a first approach to a near real-time framework for tip-sample force inversion. We utilize the hardware advantages and parallel capabilities of GPUs to develop a fast inversion strategy. A fast, parallel forward solver is developed that shows a 30000 fold speed-up over a comparable CPU implementation, resulting in milli-second calculation times. Posing the inverse problem as an unconstrained optimization problem allows us to integrate a GPU based gradient-free global Particle Swarm Optimization framework with the forward solver. We illustrated the framework on three classes of tip-sample interaction inversions. Each of these inversions is performed in sub-second timings showing potential for on-line integration with the AFM.

5.2 Future work

To improve this computational framework, efforts need to be made to this proof of concept into application. Some examples of extensions of this work to meet application needs include:

1. Integrating with AFM hardware for deployment

2. Investigating additional complex tip-sample parametrization that account for other tip-

sample interactions

3. Investigate other cost functionals

   One example of a cost functional to investigate would be to account for both deflection amplitude and phase changes. Current implementation only accounts for amplitude changes.

4. Perform nano-composition mapping of soft tissue

5. Further develop strategies to minimize the number of time points required for calculation and analysis.

APPENDIX A. Extra information

A.1 Dynamic cantilever model data

L2 error	0.00333531
Length	525 microns
Elasticity	$1.76 * 10^{11}$ Pa
Width	35 microns
Height	5 microns
Inertia	3.64583 kg m^2
Mass per unit length	$4.0775 * 10^{-7}$ $\frac{kg}{m}$
Start Time	0 Sec
End Time	0.004 Sec
Time step	$4 * 10^{-8}$ Sec
Number of modes	8
Number of modal function points	501
Amplitude	5 nm
Frequency	25500 Hz
First natural frequency	25468.9 Hz
Steady State Deflection Amplitude	3.2 microns

A.2 Modal analysis

The deflection of the cantilever then is represented as:

$$u(x,t) = \sum_{i=1}^{N} \eta_i(t)\Phi_i(x), \tag{A.1}$$

where $\Phi_i(x)$ represents a set of orthogonal modal shape functions which are computed by solving the homogeneous eigenvalue problem (34) and $\eta_i(t)$ are corresponding modal coefficients. In order to determine the modal coefficients, we define an inner product as:

$$\langle f, g \rangle = \int_L fg dx, \tag{A.2}$$

where f and g are functions.

A.3 Hardware

Testing and implementation of the proposed framework occurred on a GPU mini-cluster consisting of two Dell Precision T7500 workstations. Each node is equipped with 12 GB of DDR3 RAM, 500 GB of 10K RPM hard drives and 2 Intel Xeon 2 GHz quad-core CPUs. The main computational power comes from 4 GPUs donated to us by NVIDIA: each node is accelerated with 2 NVIDIA QUADRO FX 5800. One such card provides 240 cores, 4 GB of RAM with 102 GB/s bandwidth and is CUDA compatible (with 1.3 compute capability). The nodes are connected via dedicated Gbit Ethernet.

BIBLIOGRAPHY

[1] J. Loos, "The art of SPM: Scanning probe microscopy in materials science," *Advanced Materials*, vol. 17, no. 15, pp. 1821-1833, 2005.

[2] L. G. M. Beekmans, M. A. Hempenius, and G. J. Vancso, "Morphological development of melt crystallized poly(propylene oxide) by in situ AFM: formation of banded spherulites," *European Polymer Journal*, vol. 40, no. 5, pp. 893-903, 2004. Selected papers from the 3rd International Conference on Scanning Probe Microscopy of Polymers.

[3] J. Hahm and S. J. Sibener, "Time-resolved atomic force microscopy imaging studies of asymmetric ps-b-pmma ultrathin films: Dislocation and disclination transformations, defect mobility, and evolution of nanoscale morphology," *The Journal of Chemical Physics*, vol. 114, no. 10, pp. 4730-4740, 2001.

[4] S. Cho, A. Quinn, M. Stromer, S. Dash, J. Cho, D. Taatjes, and B. Jena, "Structure and dynamics of the fusion pore in live cells," *Cell Biol Int.*, vol. 26, pp. 35-42., 2002.

[5] H. Lin, D. O. Clegg, and R. Lal, "Imaging real-time proteolysis of single collagen i molecules with an atomic force microscope," *Biochemistry*, vol. 38, pp. 9956-9963, 1999.

[6] P. K. Hansma, J. P. Cleveland, M. Radmache, D. A. Walters, P. E. Hillner, M. Bezanilla, M. Fritz, D. Vie, H. G. Hansma, C. B. Prater, J. Massie, L. Fukunaga, J. Gurley, and V. Elings, "Tapping mode atomic force microscopy in liquids," *Applied Physics Letters*, vol. 64, pp. 1738-1740, 1994.

[7] P. Parot, Y. F. Dufre, P. Hinterdorfer, C. L. Grimellec, D. Navajas, J.- L. Pellequer, and S. Scheuring, "Past, present and future of atomic force microscopy in life sciences and medicine," *Journal of Molecular Recognition*, vol. 20, pp. 418-431, 2007.

[8] J. K. H. Hörber1 and M. J. Miles, "Scanning probe evolution in biology," *Science*, vol. 302, pg. 1002-1005, 2003.

[9] P. P. Lehenkari, G. T. Charras, A. Nykanen, and M. A. Horton, "Adapting atomic force microscopy for cell biology," *Ultramicroscopy*, vol. 82, pg. 289-295, 2000.

[10] JP Cleveland, B. Anczykowski, AE Schmid, and VB Elings, "Energy dissipation in tapping-mode atomic force microscopy," *Applied Physics Letters*, vol. 72, pg. 2613, 1998.

[11] R. Garcia and A. San Paulo, "Attractive and repulsive tip-sample interaction regimes in tapping-mode atomic force microscopy," *Physical Review B*, vol. 60, no. 7 pg. 4961-4967, issn. 1550-235X, 1999.

[12] J. Yoreo, "Life at the Nanoscale," *Lawrence Livermore National Laboratory*, Website, 2011. https://www.llnl.gov/str/May04/DeYoreo.html.

[13] S. Hu and A. Raman, "Chaos in atomic force microscopy," *Physical Review Letters*, vol. 96, no. 3, p. 0361072, 2006.

[14] C. A. V. Eysden and J. E. Sader, "Frequency response of cantilever beams immersed in viscous fluids with applications to the atomic force microscope: Arbitrary mode order," *Journal of Applied Physics*, vol. 101, pp. 044908-044918, 2007.

[15] F. E. Feninat, T. Ellis, E. Sacher, and I. Stangel, "A tapping mode AFM study of collapse and denaturation in dentinal collagen," *Dental Materials*, vol. 17, pp. 284-288, 2001.

[16] S. Hu and A. Raman, "Inverting amplitude and phase to reconstruct tip-sample interaction forces in tapping mode atomic force microscopy," *Nanotechnology*, vol. 19, no. 37, p. 375704, 2008.

[17] W.-J. Chang, C.-M. Lin, J.-F. Lee, and S.-L. Lin, "Determination of damping force between atomic force microscope tips and sample using an inverse methodology," *Physics Letters A*, vol. 343, no. 1-3, pp. 79-84, 2005.

[18] J. Sader, "Frequency response of cantilever beams immersed in viscous fluids with applications to the atomic force microscope," *Journal of Applied Physics*, vol. 84, no. 1, pg. 64-76, 1998.

[19] J. W. M. Chon, P. Mulvaney, and J. Sader, "Experimental validation of theoretical models for the frequency response of atomic force microscope cantilever beams immersed in fluids," *Journal of Applied Physics*, vol. 87, no. 8, pg. 3978-3988, 2000.

[20] M. Clark, J. Sader, J. Cleveland, and M. Paul, "Spectral properties of microcantilevers in viscousfluid," *Physical Review E*, vol. 81, no. 4, p. 046306, 2010.

[21] D. Kiracofe and A. Raman, "Microcantilever dynamics in liquid environment dynamic atomic force microscopy when using higher-order cantilever eigenmodes," *Journal of Applied Physics*, vol. 108, no. 3, pp. 034320034320, 2010.

[22] O. Payton, L. Picco, A. Champneys, M. Homer, M. Miles, and A. Raman, "Experimental observation of contact mode cantilever dynamics with nanosecond resolution," *Review of Scientific Instruments*, vol. 82, p. 043704, 2011.

[23] M. Salapaka, D. Chen, and J. Cleveland, "Linearity of amplitude and phase in tapping-mode atomic force microscopy," *Physical Review B*, vol. 61, no. 2, p. 1106, 2000.

[24] A. Sebastian, A. Gannepalli, and M. Salapaka, "A review of the systems approach to the analysis of dynamic-mode atomic force microscopy," *IEEE Trans. on Control Systems Technology*, vol. 15, no. 5, pp. 952959, 2007.

[25] D. Abramovitch, S. Andersson, L. Pao, and G. Schitter, "A tutorial on the control of atomic force microscope," in *Proceedings of American Control Conference*, (New York City, NY), pp. 34993502, July 2007.

[26] S. Devasia, E. Eleftheriou, and S. O. R. Moheimani, "A survey of control issues in nanopositioning," *IEEE Trans. on Control Systems Technology*, vol. 15, pp. 802823, 2007.

[27] G. M. Clayton, S. Tien, K. K. Leang, Q. Zou, and S. Devasia, "A review of feedforward control approaches in nano-positioning for high-speed spm," *ASME Journal of Dynamic Systems, Measurement and Control*, vol. 131, pp. 0611011 to 06110119, 2009.

[28] Z. Xu, K. Kim, Q. Zou, and P. Shrotriya, "Broadband measurement of rate-dependent viscoelasticity at nanoscale using scanning probe microscope: Poly(dimethylsiloxane) example," *Applied Physics Letters*, vol. 93, no. 13, p. 133103, 2008.

[29] K. Kim, Q. Zou, and C. Su, "A new approach to scan-trajectory design and track: AFM force measurement example," *ASME Journal of Dynamic Systems*, Measurement and Control, 2008. in print.

[30] Z. Xu and Q. Zou, "A model-based approach to compensate for the dynamics convolution effect on nanomechanical property measurement," *Journal of Applied Physics*, 2010. accepted.

[31] P. Xie, Z. Xu, and Q. Zou, "Compensation for the dynamics effect on nanoscale broadband viscoelasticity measurements using adaptive filtering approach," *IEEE Trans. on Instrumentation and Measurement*, 2010. in print.

[32] T. Rodriguez and R. Garcia, "Tip motion in amplitude modulation (tapping-mode) atomic-force microscopy: Comparison between continuous and point-mass models," *Applied Physics Letters*, vol. 80, num. 9, 2002.

[33] C.A.V. Eysden and J.E. Sader, "Resonant frequencies of a rectangular cantilever beam immersed in a fluid," *Journal of Applied Physics*, vol. 100, Dec 2006.

[34] L. Meirovitch, *Fundamentals of Vibrations*. 1221 Avenue of the Americas, New York, NY, 10020: McGraw-Hill, 2001.

[35] D. Parks and L. Anand, "Mechanics and Materials II," Website, 2011. http://ocw.mit.edu/courses/mechanical-engineering/2-002-mechanics-and-materials-ii-spring-2004.

[36] V. Morris, A. Kirby, and A. Gunning, *Atomic Force Microscopy for Biologists*, 57 Shelton Street, Covent Garden, London, WC2H 9HE, United Kingdom: Imperial College Press, 2010.

[37] J. Israelachvili, *Intermolecular & Surface Forces*, 24-28 Oval Road, London, NW1 7DX, United Kingdom: Academic Press Limited, 1992.

[38] K. Johnson, *Contact Mechanics*, 37 East 57th Street, New York, New York, 10022, USA: Cambridge University Press, 1985.

[39] R. Reifenberger and A. Raman, "Fundamentals of Atomic Force Microscopy," Website, 2010. http://nanohub.org/resources/9598

[40] Hertz, H. R., 1882, "Ueber die Beruehrung elastischer Koerper," (On Contact Between Elastic Bodies), *Gesammelte Werke* (Collected Works), Vol. 1, Leipzig, Germany, 1895.

[41] U. Schwarz, O. Zworner, P. Koster, and R. Wiesendanger, "Quantitative analysis of the frictional properties of solid materials at low loads. I. Carbon compounds," *Physical Review B*, vol. 56, no. 11, 15 Sept. 1997.

[42] B. Derjaguin, V. Muller, and Y. Toporov, "Effect of Contact Deformations on the Adhesion of Particles," *Journal of Colloid and Interface Science*, vol. 53, no. 2, Nov. 1975.

[43] H. Butt, B. Cappella, and M. Kappl, "Force measurements with the atomic force microscope: Technique, interpretation and applications," *Surface Science Reports*, vol. 59, pg. 1-152, 2005.

[44] J. Cho, D. Lee, J. Lim, K. Cho, J. Je, and J. Yi, "Evaluation of the adhesion properties of inorganic materials with high surface energies," *Langmuir*, vol. 20, no. 23, pg. 10174-10178, 2004.

[45] M. Kober, E. Sahagun, M. Fuss, F. Briones, M. Luna, and J. Saenz, "Adhesion hysteresis in dynamic atomic force microscopy," *Physica Status Solidi Rapid Research Letters*, vol. 2, no. 3, pg. 138-140, 2008.

[46] R. Garcia, C.J.Gomez, N.F.Martinez, S. Patil, C. Dietz, and R. Magerie, "Identification of nanoscale dissipation Processes by dynamic atomic force Microscopy," *Physics Review Letters*, vol. 97, issue 1, 2006.

[47] J. Melcher, C. Carrasco, X. Xu, J. Carrascosa, J. Gomez-Herrero, P. Jose de Pablo, A. and Raman, "Origins of phase contrast in the atomic force microscope in liquids," *Proceedings of the National Academy of Sciences* 106, 13655, 2009.

[48] "CUDA toolkit reference manual," Website, 2010. http://developer.download.nvidia.com/compute/cuda/3_2/toolkit/docs/CUDA_Toolkit_Reference_Manual.pdf.

[49] J. Kennedy and R. Eberhart, "Particle swarm optimization," *IEEE International Conference on Neural Networks Proceedings*, vol. 4, pg. 1942 - 1948, 1995.

[50] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information Processing Letters*, 85, pp. 317-325, 2003.

[51] M. Jiang, Y.P. Luo, and S.Y. Yang, "Stochastic convergence analysis and parameter selection of the standard particle swarm optimization algorithm," *Information Processing Letters*, 102, pp. 8-16, 2007.

[52] L. Mussi, F. Daolio, and S. Cagnoni, "CUDA-PSO," Version 1.0, Universita' degli Studi di Parma, Italy, 2011, http://www.ce.unipr.it/people/mussi/projects/CUDA-PSO-v1.0-html_documentation/main.html.

[53] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, 233 Spring Street, New York, NY, 10013, USA: Springer Science and Business Media, Inc., 2005.

[54] S. Kurtz, "The UHMWPE Handbook:Ultra-High Molecular Weight Polyethylene in Total Joint Replacement ," 525 B Street, Suite 1900, San Diego, CA, 92101, USA: Elsevier Academic Press, 2004.

[55] J. Mark, "Polymer Data Handbook," 198 Madison Avenue, New York, New York, 1001, USA: Oxford University Press, Inc., 1999.

[56] J. Lötters, W. Olthuis, P. Veltink, and P. Bergveld, "The mechanical properties of the rubber elastic polymer polydimethylsiloxane for sensor applications," *Journal of Micromechanics and Microengineering*, no. 7, pg 145-147, 1997.