AKADEMIA GÓRNICZO-HUTNICZA

IM. STANISŁAWA STASZICA W KRAKOWIE



Wydział Inżynierii Metali i Informatyki Przemysłowej



Praca Magisterska

pt.

„Implementacja metody dynamiki molekularnej dla heterogenicznych platform sprzętowych”

Imię i nazwisko dyplomanta: **Tomasz Nowak**

Kierunek studiów: **Informatyka Stosowana**

Profil dyplomowania:  **Modelowanie i Technologie Informacyjne**

Nr albumu: **232187**

Promotor: **drinż. Łukasz Rauch**

Podpis dyplomanta: Podpis promotora:

Kraków 2014

***Oświadczam, świadomy(-a) odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszy projekt inżynierski wykonałem(-am) osobiście i samodzielnie i że nie korzystałem(-am) ze źródeł innych niż wymienione w pracy.***

Kraków, dnia ………….… Podpisdyplomanta…………….

**Spis treści**

[1. Wstęp 4](#_Toc399442322)

[2. Dynamika molekularna 5](#_Toc399442323)

[2.1. Przegląd artykułów 6](#_Toc399442324)

[2.2. Przegląd algorytmów 6](#_Toc399442325)

[3. Realizacja oprogramowania 8](#_Toc399442326)

[3.1. GPU 8](#_Toc399442327)

[3.2. Nvidia CUDA 8](#_Toc399442328)

[3.3. Pthreads 9](#_Toc399442329)

[3.4. OpenGL 10](#_Toc399442330)

[3.5. Projekt aplikacji 10](#_Toc399442331)

[3.6. Interfejs użytkownika s 13](#_Toc399442332)

[3.7. Implementacja 15](#_Toc399442333)

[3.8. Optymalizacja 21](#_Toc399442334)

[4. Testy wydajności 22](#_Toc399442335)

[4.1. Testy sekwencyjne 22](#_Toc399442336)

[4.2. Testy zrównoleglenia 22](#_Toc399442337)

[4.3. Testy optymalizacji 22](#_Toc399442338)

[4.4. Test wielu GPU 22](#_Toc399442339)

[5. Model algorytm - urządzenie 23](#_Toc399442340)

[5.1. XXX 23](#_Toc399442341)

[6. Podsumowanie 24](#_Toc399442342)

[7. Bibliografia 25](#_Toc399442343)

# Wstęp

Rosnąca moc obliczeniowa współczesnego sprzętu dostarcza nowe możliwości rozwiązywania coraz bardziej skomplikowanych zadań. Zagadnieniem bardzo wymagającym obliczeniowo jest dynamika molekularna. Jest jednocześnie ciekawym obszarem badań, który łączy w sobie wiele dziedzin nauki. Zastosowanie dynamiki molekularnej można znaleźć w badaniach nad własnościami fizycznymi materiałów. Symulacje z wykorzystaniem nanostruktur potrafią usprawnić badania, ograniczając czasochłonne wykonywanie doświadczeń z wykorzystaniem rzeczywistych próbek. Skrajnie różnym, ale bazującym na zbliżonej symulacji, doświadczeniem jest analiza wytrzymałości wiązań chemicznych. Domena tego typu symulacji jest dużo mniejsza, ale opiera się wciąż na działaniu sił zewnętrznych na nanostrukturę. Można szukać podobnych zagadnień również w astronomii, biologii, fizyce lub najprościej rzecz ujmując wszędzie, ponieważ wszystko jest zbudowane z atomów. Dlatego właśnie dynamika molekularna opisująca interakcję sił zewnętrznych z atomami, może zostać wykorzystana do dowolnych badań w skali nano. Aby w pełni korzystać z możliwości jakie daje ta metoda niezbędny jest odpowiedni sprzęt umożliwiający przeprowadzenie symulacji. Wykorzystanie akceleratorów graficznych GPU jest odpowiednim rozwiązaniem, ze względu na łatwą dekompozycję danych oraz możliwość efektywnego wykorzystania oferowanej przez sprzęt masowej wielowątkowości. Odpowiednia implementacja zagadnienia umożliwi wielokrotne wykorzystanie powstałej aplikacji do rozwiązywania problemów z różnych dziedzin przy niewielkim nakładzie pracy włożonym w modyfikację i przystosowanie do aktualnie analizowanej sytuacji.

Celem pracy jest stworzenie jak najbardziej uniwersalnej i otwartej implementacji algorytmów dynamiki molekularnej przy wykorzystaniu akceleratorów GPU. Implementacja zostanie wykorzystana do analizy problemów odkształceń materiałów przy różnorodnej definicji działających sił zewnętrznych np. temperatury, prasy, symulacji defektu struktury.

# Dynamika molekularna

Dynamika molekularna jest komputerową symulacją fizycznych ruchów atomów wewnątrz struktury. Atomy mogą wchodzić w interakcję między sobą w kolejnych krokach czasowych, dając wrażenie poruszania się. Siły pomiędzy atomami oraz potencjał energetyczny definiowane są przez mechanikę molekularną pól siłowych.[9]

//ROWNANIE F=m\*a i z gradientem

N – ilość atomów w strukturze

mi – masa atomu

a – przyspieszenie

Fi – działająca siła

V(r1 …. rN) – funkcja pozycji atomów; reprezentowana jest przez potencjał energetyczny

W praktyce potencjał jest zapisywany jako suma interakcji pomiędzy parami atomów [8]:

//ROWNANIE V(r1….rN) = SUMu1…………u2….

Najczęściej używanym praktycznie potencjałem jest potencjał Lenarda – Jonesa (LJ). Jest najczęściej wykorzystywany w modelowaniu interakcji pomiędzy atomami. Definiuje go poniższe równanie [8]:

//ROWNANIE LJ PDF Ti.odio.twiki

r – dystans pomiędzy atomami będącymi w interakcji

DELTA – średnica

EURO – minimum lokalne potencjału

EURO i DELTA są stałymi wyznaczanymi tak, aby definiowały właściwości fizyczne materiału [8].

Najbardziej czasochłonną częścią algorytmu jest obliczanie sił interakcji pomiędzy atomami. Zajmuje to ok. 90% czasu całej symulacji. Analizując przebieg symulacji widać, że obliczenie siły występuje pomiędzy każdym atomem struktury z każdym z pozostałych. Daje to złożoność obliczeniową rzędu O(N2). [8]

## Przegląd artykułów

W trakcie analizowania problemu dynamiki molekularnej, można spotkać się z wieloma sposobami optymalizacji algorytmu tak, aby znacząco zmniejszyć złożoność obliczeniową przy jednoczesnym zachowaniu jak największej dokładności.

Jednym ze sposobów jest zastosowanie tzw. promienia odcięcia. Metoda ta redukuje ilość obliczeń tylko do tych niezbędnych, ponieważ interakcja z bardziej oddalonymi atomami jest na tyle znikoma, że może zostać pominięta.[8]

Problem zdefiniowany globalnie tj. przechowywanie całej struktury w pamięci urządzenia nie jest optymalny i ciężki do równoległego wykonywania na więcej niż i układzie GPU jednocześnie. Aby usprawnić obliczenia i przygotować dane wejściowe na platformę równoległą można zastosować listę sąsiadów. Dla każdego atomu, na podstawie promienia odcięcia definiuje się listę atomów, które spełniają dany warunek i tylko one trzymane są w pamięci podczas obliczeń interakcji z wybranym atomem. Daje to możliwość podzielenia danych niezależnie od wielkości całej struktury i rozesłania mniejszych list na kilka niezależnych urządzeń. Rozwiązuje również problem z pamięcią, gdy struktura jest na tyle duża, że nie zmieści się w pamięci pojedynczego urządzenia. Lista sąsiadów jest aktualizowana co pewien określony przedział czasowy. [8]

Tworzenie listy sąsiadów jest skuteczne podczas symulacji bazujących na gęsto upakowanych atomach (ciała stałe, gęste ciecze) ponieważ atomy są ułożone bardzo blisko siebie i nie mają możliwości szybkiego przemieszczania się. Atomy gazów, mają dużo większe odległości i są niejednorodnie rozłożone w strukturze. Wymagałoby to zastosowanie dużo większego promienia odcięcia i znacznie częstszych aktualizacji list sąsiadów, aby wynik symulacji był wciąż miarodajny i poprawny.

## Przegląd algorytmów

Na przestrzeni lat powstało wiele implementacji dynamiki molekularnej. Algorytmy różniły się od siebie ze względu na zależności związane z architekturą sprzętu na którym powstawały lub zależnościami samej aplikacji. Z punktu widzenia dekompozycji danych można podzielić je następująco [8]:

* Dekompozycja atomów

Każdy z procesorów ma przydzieloną podzbiór danych wielkości N/P (N jest liczbą atomów; P jest liczbą procesorów) na początku symulacji. Każdy z procesorów musi przechowywać identyczną kopię informacji o atomach (metoda replikacji danych). Metoda ta znalazła zastosowanie wśród architektur z pamięcią współdzieloną. [8]

* Dekompozycja siły

Podzbiór danych pętli sił jest przydzielany każdemu procesorowi. Redukuje to kosztowną komunikację oraz koszt dostępu do pamięci do (PIERWIASTEK Z P) w porównaniu z dekompozycją atomów. Problemem jest odpowiednie zbalansowanie rozkładu danych pomiędzy procesory. [8]

* Dekompozycja przestrzeni

Polega na dekompozycji geometrycznej fizycznej struktury atomów. Każdy procesor oblicza siły działające na atomy tylko na fragmencie całej struktury. W trakcie symulacji procesory wymieniają ze sobą tylko atomy które przekroczyły granice swojej części struktury i przemieściły się do sąsiedniej. Metoda ta wspiera wielkoskalowe symulacje. Osiąga optymalną skalowalność O(N/P). [8]

Biorąc pod uwagę architekturę jaka będzie wykorzystywana do wykonywania symulacji, najlepszą metodą dekompozycji danych jest dekompozycja atomów. Główną zaletą tego podejścia jest bardzo dobre zrównoważenie obciążenia pomiędzy urządzeniami oraz łatwa do osiągnięcia skalowalność. Model współdzielonej pamięci GPU pomiędzy procesorami umożliwia osiągnięcie wysokiej wydajności w trakcie obliczeń.

# Realizacja oprogramowania

## GPU

Karty graficzne stają się coraz bardziej zaawansowane, przez co zaczynają być wykorzystywane również do ogólnych zastosowań. Superkomputery rozwiązują wiele skomplikowanych zadań obliczeniowych na liczbach zmiennoprzecinkowych. Procesory GPU dzięki swojej równoległej architekturze, są idealnym wsparciem w obliczeniach. Aby mogły zostać zastosowane w tym celu, potrzebny jest zestaw instrukcji który umożliwi pisanie aplikacji w sposób zrozumiały zarówno dla urządzenia jak i programisty. [7]

Nvidia jest wiodącym producentem procesorów GPU. Jest również twórcą CUDA - biblioteki i środowiska umożliwiającego wykorzystanie kart graficznych do ogólnych zastosowań.

## Nvidia CUDA

Nvidia CUDA jest platformą służącą do obliczeń równoległych oraz modelem programistycznym umożliwiającym uzyskać znacząco lepszą wydajność obliczeniowa wykorzystując moc układu graficznego GPU.

Środowisko daje programiście bezpośredni dostęp do zestawu wirtualnych instrukcji i pamięci na równoległych jednostkach procesorów. Można dzięki temu wykorzystywać GPU do ogólnych zastosowań nie tylko do przetwarzania grafiki. Biblioteki CUDA wspierają popularne języki programowania: C, C++, Fortran.

Aplikacje wspierające akceleratory graficzne GPU oraz biblioteki CUDA wykorzystywane są między innymi w:

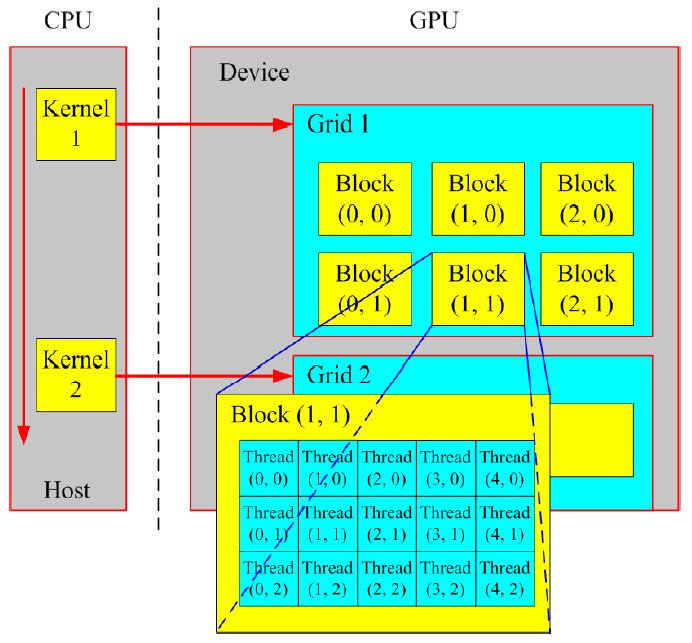
* astronomii,
* biologii,
* chemii,
* fizyce,
* finansach,
* procesach przemysłowych.

Model programowania CUDA definiowany jest przez kilka podstawowych założeń prawdziwych dla programów pisanych pod karty Nvidii:

* kod równoległy tzw. *kernel* jest uruchamiany na urządzeniu (karcie graficznej)   
  i wykonywany przez wiele wątków,
* wątki są grupowane w bloki wątków,
* kod równoległy jest pisany dla pojedynczego wątku
  + każdy wątek wykonuje unikalną część kodu
* bloki są grupowane w siatki

Na potrzeby implementacji zagadnienia będącego tematem pracy wykorzystane zostały [4]:

* CUDA Toolkit 5.5
* CUDA Samples – biblioteki wspomagające oraz funkcje pomocnicze
* CUDA Tools – narzędzia do profilowania i badania pamięci



Rys 3.1. Model programowania CUDA

## Pthreads

POSIX Threads często nazywany również Pthreads jest standardem POSIX dla wątków. Standard ten definiuje API[[1]](#footnote-2) do tworzenia i manipulowania wątkami. Implementacja ta jest dostępna dla wielu systemów operacyjnych opartych na systemie UNIX takich jak: FreeBSD, NetBSD, OpenBSD, GNU/Linux, Mac OS X, Solaris. Istnieje również implementacja przeznaczona dla systemu Microsoft Windows zawarta w WinAPI[[2]](#footnote-3). [6]

Pthreads definiuje zestaw typy, funkcje, stałe dla języka C. Implementacja znajduje się w pliku *pthreads.h* biblioteki zarządzającej wątkami. Zawiera ponad 100 funkcji, które można podzielić na grupy: [6]

* zarządzanie wątkami,
* muteksy,
* zmienne warunkowe,
* synchronizacja

Podstawowym celem Pthreads jest przyspieszenie realizacji zadań. Porównując koszty tworzenia i zarządzania procesem, wątek może być utworzony przy minimalnym udziale zasobów systemu.

W powyższym projekcie technologia Pthreads została wykorzystana w celu zarządzania akceleratorami obliczeń, jakimi są karty graficzne. Bierze udział w procesie podziału danych i przygotowaniu ich do przesłania na zewnętrzne urządzenia. Pomaga również w końcowym procesie zarządzania wynikami obliczeń dostarczonymi do maszyny gospodarza (eng. host).

## OpenGL

OpenGL jest wieloplatformową technologią używaną przez różnorodne języki programowania umożliwiającą przetwarzanie obrazu dwu- i trójwymiarowego. Wykorzystywany jest do interakcji z procesorami graficznymi GPU aby uzyskać sprzętową akcelerację przetwarzanego obrazu. OpenGL jest stale rozwijaną technologią, która wspiera pojawiające się na rynku najnowsze karty graficzne wiodących producentów.

## Projekt aplikacji

Jednym z celów projektu było stworzenie aplikacji, która będzie nie tylko implementacją mechaniki dynamiki molekularnej, ale również otwartym oprogramowaniem łatwym w użytkowaniu i rozbudowie. Istotną cechą programu jest możliwość wizualizacji wyników, jak również śledzenia zmian zachodzących w strukturze w czasie rzeczywistym. W przypadku uruchamiania testów na klastrach obliczeniowych jest możliwość uruchomienia aplikacji wyłącznie w trybie tekstowym.

Aplikacja wykorzystuje funkcje pomocnicze oraz obsługę błędów pochodzące z biblioteki CUDA. Każdy z plików będący częścią tej biblioteki jest oznaczony na początku specjalnym tekstem. Wspomnienie źródła pochodzenia takiego kodu umożliwia jego wykorzystanie w dowolnym celu edukacyjnym nie łamiąc jednocześnie praw jego stosowania. Przykładowe oznaczenie:

/\*\*

\* Copyright 1993-2013 NVIDIA Corporation. All rights reserved.

\*

\* Please refer to the NVIDIA end user license agreement (EULA) associated

\* with this source code for terms and conditions that govern your use of

\* this software. Any use, reproduction, disclosure, or distribution of

\* this software and related documentation outside the terms of the EULA

\* is strictly prohibited.

\*

\*/

////////////////////////////////////////////////////////////////////////////////

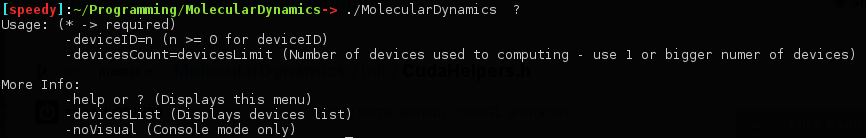
// These are CUDA Helper functions for initialization and error checking

Projekt zakładał stworzenie spójnej aplikacji składającej się z jak najbardziej niezależnych od siebie elementów współpracujących ze sobą. Takie założenie umożliwiło realizację oprogramowania, składa się z poniższych komponentów:

* MolecularDynamics  
  Jest to główny komponent programu. Struktura wejściowe jest tutaj tworzona i inicjalizowana. Odpowiada również za wczytanie parametrów i uruchomienie symulacji.
* Simulation  
  Jest komponentem zajmujący się przetwarzaniem danych wejściowych oraz sterowaniem przebiegu symulacji na GPU. Przetwarzane są tutaj wszelkie parametry dodatkowe podawane przez użytkownika w trakcie uruchamiania programu. Mają one wpływ na typ symulacji oraz urządzenia, które zostaną wykorzystane do jej przeprowadzenia.
* Atom  
  Jest podstawowym elementem wykorzystywanym w dynamice molekularnej. Przechowuje szczegółowe informacje o położeniu aktualnym oraz początkowym, typie czy sile działającej na pojedynczy atom w strukturze molekularnej.
* Structure  
  Jest kombinacją setek, tysięcy lub nawet milionów atomów, tworzących całość jaką jest struktura molekularna. Przechowuje informacje o typie zastosowanego potencjału, sposobie przyłożenia siły zewnętrznej oraz wartości tej siły. Jest głównym parametrem wejściowym do symulacji.
* GpuHandler  
  Jest komponentem ułatwiającym zarządzanie urządzeniami GPU. Przechowuje identyfikatory dostępnych kart graficznych oraz ich ilość. Umożliwia ich inicjalizowanie i przygotowanie do obliczeń.
* GpuDisplay  
  Jest częścią aplikacji zarządzającą wizualizacją danych. Zajmuje się inicjalizacją środowiska OpenGL, umożliwia konwersję danych wykorzystywanych w obliczeniach do postaci umożliwiającej ich wyświetlenie oraz obsługą myszki i klawiatury. Zapewnia możliwość interakcji ze strukturą w czasie rzeczywistym oraz podgląd każdej części struktury w dowolnym powiększeniu.
* GpuKernel  
  Jest to komponent zajmujący się wykonywaniem obliczeń przy pomocy GPU. Zarządza pamięcią oraz danymi przesyłanymi na urządzenie. Uruchamia funkcje przeznaczone na urządzenie (eng. kernel). Umożliwia wykonywanie obliczeń w trybie wizualizacji lub trybie tekstowym.
* CudaHelpers  
  Jest miejscem do przechowywania elementów aplikacji wykorzystywanych przez wszystkie komponenty wchodzące w interakcję z GPU. Przechowuje dostępne kernele oraz funkcje zajmujące się wielowątkowym przetwarzaniem na kilku urządzeniach GPU.
* Log  
  Jest użytecznym komponentem, umożliwiającym tworzenie logów na standardowym wyjściu jak i w pliku tekstowym. Ułatwia śledzenie błędów gdy aplikacja niespodziewanie wyrzuci błąd lub przestanie działać.

## Interfejs użytkownika s

Program zaprojektowany jest tak, aby był jak najprostszy w obsłudze. Jest uruchamiany z linii poleceń z kilkoma, najczęściej używanymi i najistotniejszymi, wymaganymi parametrami. Pozostałe ustawienia symulacji wczytywane są z plików konfiguracyjnych. Takie podejście nie wymaga ponownej kompilacji przy zmianie rozwiązywanego problemu lub przy modyfikacji struktury atomów. Dostępne opcje uruchomienia mogą zostać wyświetlone po wpisaniu parametrów ‘help’ lub ‘?’.



Rys 3.2. Opcje interfejsu linii poleceń podczas uruchamiania programu.

W celu poprawnego uruchomienia programu wymagane jest podanie jednego z dwóch parametrów: ‘deviceID’ lub ‘devicesCount’. Pozwoli to wykonać obliczenia na konkretnie wybranym urządzeniu GPU lub na losowo wybranym z dostępnych. Aby wskazać konkretną kartę należy podać jej ID. Aby określić ID urządzenia, które ma być użyte można skorzystać z opcji wyświetlenia wszystkich dostępnych urządzeń Nvidii. Służy do tego parametr ‘devicecsList’.



Rys 3.3. Lista dostępnych urządzeń GPU

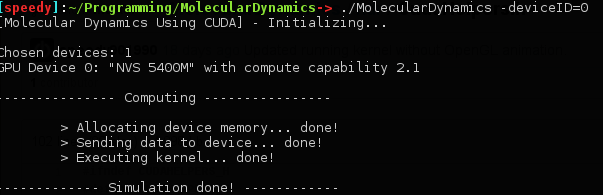
W przypadku pominięcia wyboru urządzenia program zwróci błąd. Jest to celowy zabieg, aby zwiększyć kontrolę i świadomość z jakich urządzeń oprogramowanie będzie korzystało podczas symulacji.

Pozostała część konfiguracji wczytywana jest z katalogu *config*. Znajdują się tam dwa pliki: *simulation.cfg* oraz *structure.cfg.* Pierwszy z nich zawiera informacje o typie użytego potencjału. Drugi natomiast przechowuje konfigurację struktury wejściowej: jej rozmiar oraz typ i wartość przyłożonej siły.



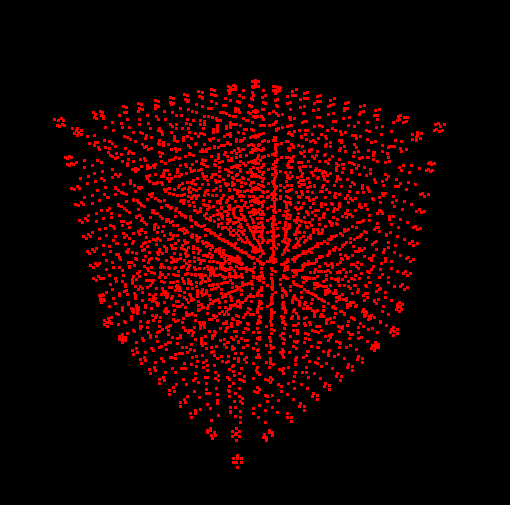
Rys 3.4. Plik konfiguracyjny programu – structure.cfg.

Po dokonaniu poprawnej konfiguracji program może zostać uruchomiony w dwóch trybach: z wizualizacją lub bez. Niezależnie od wybranej opcji wynik wyświetlany przez konsolę będzie taki sam. Pozwala to śledzić aktualny status wykonywania symulacji oraz zaobserwować moment zakończenia. Dane wyświetlane są szczegółowo przez co widać na którym urządzeniu wykonywane są obliczenia, kiedy dane wysyłane są na GPU, a kiedy uruchomiony zostanie kernel.



Rys 3.5. Przykład poprawnie uruchomionej i wykonanej symulacji.

W przypadku uruchomienia symulacji z trybem wizualizacji, udostępniony zostaje interfejs do sterowania wyświetlaniem struktury w czasie rzeczywistym. Aby obrócić strukturę należy przytrzymać lewy przycisk myszy oraz przesunąć urządzenie wskazujące w kierunku w którym struktura ma zostać obrócona. Pod prawym przyciskiem myszy zaimplementowana została funkcja przybliżania i oddalania obrazu. Pozwala to na przybliżenie dowolnej części struktury i obserwację bezpośredniego ruch pojedynczych atomów. Klawisz *Escape* służy do zakończenia symulacji i zwolnienia pamięci zaalokowanej w czasie działania programu.



Rys 3.6. Wizualizacja struktury w czasie rzeczywistym.

## Implementacja

Proces implementacji został podzielony na dwa etapy. Pierwszym było przygotowanie danych do obliczeń i zapewnienie ich przepływu przez wszystkie komponenty wyszczególnione w rozdziale 3.5. Na tym etapie celem było uzyskanie niezależności i szybkiego dostępu. Dane przekazywane były bez kopiowania ich co zaoszczędziło sporo zasobów systemowych. Jedynym elementem gdzie wymagane było kopiowanie, jest moment przesyłania ich do urządzenia GPU. Ten element również został ograniczony do minimum, gdyż do pamięci karty graficznej trafiają tylko dane bezpośrednio wykorzystywane w danej iteracji. Wszystkie z wymienionych operacji wykonywane były po stronie ‘gospodarza’, czyli komputera PC lub klastra. Drugim etapem implementacji było napisanie zarówno funkcji bezpośrednio wykonywanych przez GPU. Zadania te dzieliły się na część obliczeniową oraz graficzną – wizualizację. Celem było zapewnienie płynnego przepływu danych pomiędzy komputerem PC, a urządzeniami GPU.

**Symulacja**

**GpuHandler**

**Wyniki**

**Wizualizacja**

**Brak wizualizacji**

**Kernel**

**Kernel**

Rys 3.7. Schemat przepływu danych w programie

Wyzwaniem jakiemu należało sprostać w pierwszej kolejności było połączenie wizualizacji danych z jednoczesnym wykonywaniem obliczeń, a to wszystko przy użyciu jednej karty graficznej. Problem tkwił w sposobie wykorzystania GPU w obydwu przypadkach. Przetwarzanie obrazu jest podstawową funkcjonalnością tego typu urządzeń. Dane przeznaczone do wyświetlenia na ekranie są przechowywane w zupełnie innym formacie niż te, przeznaczone do obliczeń. Aby połączyć te dwie funkcjonalności ze sobą wykorzystany został OpenGL Vertex Buffer Object (VBO). Taki obiekt akceptuje dane w formacie *float4* co znacznie ułatwia implementację. Wektory współrzędnych atomów znajdujące się w trzech osobnych tablicach x, y, z przepisane do zmiennej typu *float4\**  bez trudu zostają przesłane i wyświetlone dzięki technologii OpenGL.

|  |
| --- |
| \_\_global\_\_ void update\_display (float4 \*pos, Structure \* input, float time)  {  int atomsCount = input->atomsCount;  int tmpCount = 0;  float u, v, w;  for (int i=0 ; (i<input->dim.x) && (tmpCount < atomsCount) ; i++) {  for (int j=0 ; (j<input->dim.y) && (tmpCount < atomsCount) ; j++) {  for (int k=0 ; (k<input->dim.z) && (tmpCount < atomsCount); k++) {  u = input->atoms[tmpCount].pos.x \* 0.1f;  w = input->atoms[tmpCount].pos.y \* 0.1f;  v = input->atoms[tmpCount].pos.z \* 0.1f;  **pos[tmpCount] = make\_float4(u, w, v, 1.0f);**  tmpCount++;  }  }  }  } |

Kod 3.1. Kernel konwertujący dane obliczeniowe na dane wyświetlane przez OpenGL.

Taka konwersja musi nastąpić po każdej iteracji obliczeniowej. Jest to niezbędne, aby w czasie rzeczywistym móc obserwować zmiany struktury atomów. Można próbować przyspieszyć wydajność przy wyświetlaniu konwertując dane co kilka iteracji. Zalecane jest to jednak tylko podczas symulacji o małej dynamice lub gdy nie zachodzi potrzeba dogłębnej analizy zmian, a tylko ogólny zarys powstałych zmian.

Aby w płynny i przejrzysty sposób połączyć wyświetlanie z obliczeniami, część obliczeniowa jest częścią głównej pętli OpenGL’a. Zasubskrybowanie na funkcję *display()* sprawia, że przy każdym przebiegu głównej pętli OpenGL wykonuje operacje w niej zawarte. Zazwyczaj są to elementy odpowiedzialne za translację, ustawienia kolorów, kopiowania buforów, zmiany widoku. W tym przypadku wstrzyknięta jest również funkcja odpowiedzialna za wykonywanie obliczeń przy pomocy technologii CUDA. Jej uruchomienie powoduje modyfikację bufora VBO, który niedługo później jest renderowany przez OpenGL’a.

|  |
| --- |
| void GpuDisplay::display() {  sdkStartTimer(&timer);  // run CUDA kernel to generate vertex positions  **runCuda(&cuda\_vbo\_resource);**  glClear(GL\_COLOR\_BUFFER\_BIT | GL\_DEPTH\_BUFFER\_BIT);  // set view matrix  glMatrixMode(GL\_MODELVIEW);  glLoadIdentity();  //glTranslatef(translate\_x, translate\_y, translate\_z);  glTranslatef(0, 0, translate\_z);  glRotatef(rotate\_x, 1.0, 0.0, 0.0);  glRotatef(rotate\_y, 0.0, 1.0, 0.0);  // render from the vbo  glBindBuffer(GL\_ARRAY\_BUFFER, vbo);  glVertexPointer(4, GL\_FLOAT, 0, 0);  glEnableClientState(GL\_VERTEX\_ARRAY);  glColor3f(1.0, 0.0, 0.0);  glDrawArrays(GL\_POINTS, 0, mesh\_width \* mesh\_height \* mesh\_depth);  glDisableClientState(GL\_VERTEX\_ARRAY);  glutSwapBuffers();  g\_fAnim += 0.01f;  sdkStopTimer(&timer);  computeFPS();  } |

Kod 3.2. Funkcja display odpowiedzialna za uruchomienie obliczeń i wyświetlenie rezultatów.

Po zapewnieniu poprawnego przetwarzania i konwertowania danych można przejść do przeanalizowania zaimplementowanego algorytmu dynamiki molekularnej. Pierwsza, podstawowa wersja zakłada prostą implementację zachowując ideę i poprawne odwzorowanie fizyki procesu. W kolejnych rozdziałach pokazane będą próby optymalizacji tego rozwiązania oraz inne podejście do rozwiązywanego problemu.

|  |
| --- |
| \_\_global\_\_ void MD\_LJ\_kernel(Structure \*input, Structure \*output, float time) {  int tid = threadIdx.x + blockIdx.x \* blockDim.x;  int atomIndexStart = tid;  int atomIndexEnd = input->atomsCount;  float force[3] = {0.0f, 0.0f, 0.0f};    // COMPUTING  float dX = 0;  float dY = 0;  float dZ = 0;  float x = 0, y = 0, z = 0;  float distance = 0;  float potential = 0;  float deltaTimeSquare = pow(0.05f, 2);  for (int i=atomIndexStart ; i<atomIndexEnd ; i += blockDim.x \* gridDim.x) {  force[0] = 0.0f;  force[1] = 0.0f;  force[2] = 0.0f;  for (int j=0 ; j<input->atomsCount ; j++) {  if (i == j)  continue;    dX = input->atoms[j].pos.x - input->atoms[i].pos.x;  dY = input->atoms[j].pos.y - input->atoms[i].pos.y;  dZ = input->atoms[j].pos.z - input->atoms[i].pos.z;  distance = sqrtf(pow(dX, 2) + pow(dY, 2) + pow(dZ, 2));    if (distance >= 2.5)  continue;    potential = 4 \* (pow((1.0f/distance), 12) - pow((1.0f/distance), 6) );  if (potential > 50)  continue;  force[0] += -(dX / distance) \* potential;  force[1] += -(dY / distance) \* potential;  force[2] += -(dZ / distance) \* potential;  }  output->atoms[i].pos.x = input->atoms[i].pos.x + 0.5 \* force[0] \* deltaTimeSquare;  output->atoms[i].pos.y = input->atoms[i].pos.y + 0.5 \* force[1] \* deltaTimeSquare;  output->atoms[i].pos.z = input->atoms[i].pos.z + 0.5 \* force[2] \* deltaTimeSquare;  }  } |

Kod 3.3. Główny kernel symulacji implementujący dynamikę molekularną.

W tym podejściu implementacji kernela, poszczególne atomy zostały przydzielone osobnym wątkom na GPU. Indeks atomu, a zarazem wątku, jest obliczana ze wzoru: *tid = threadIdx.x + blockIdx.x \* blockDim.x*. Wykorzystując wbudowane w CUDA zmienne można jednoznacznie określić szukany indeks tak, aby był unikalny. Jest to istotne, aby kilkakrotnie nie modyfikować tych samych danych i nie doprowadzać do sytuacji gdzie mógłby pojawić się wyścig między wątkami. Pozycja atomu jest modyfikowana na podstawie wpływu otaczających go sąsiednich molekuł. Do obliczeń brane są tylko te sąsiadujące atomy, których odległość jest mniejsza bądź równa określonej odległości – w tym przypadku 2.5 jednostki. Jest to zastosowanie tzw. promienia odcięcia. Jeśli odległość mieści się w ograniczeniach, obliczana jest siła oddziaływania na podstawie potencjału energetycznego oraz samej odległości. Na tym etapie mamy do czynienia ze statyką molekularną. Ostatnim krokiem, aby można było mówić o dynamice, jest uwzględnienie zewnętrznej siły działającej na strukturę. W przypadku tej funkcji symulowana jest obróbka termiczna. Taki typ symulacji charakteryzuje się tym, że zewnętrzna siła działa na strukturę z zewnątrz dookoła i równomiernie. Pomijany jest w tym miejscu nierównomierny rozkład temperatury w piecu. W całej symulacji masa atomu została ustalona na wartość 1, przez co może ona zostać pominięta w obliczeniach. Jest to istotny zabieg, ponieważ GPU ma w ten sposób kilka operacji zmiennoprzecinkowych do wykonania mniej. W skali tysięcy atomów uzyskuje się większą wydajność. W kolejnych rozdziałach zostaną przedstawione sposoby optymalizacji oraz testy wykonane przy użyciu opisywanego oprogramowania. Istotną częścią funkcji jest zastosowanie warunków brzegowych. Jeden z nich został przytoczony powyżej i jest nim promień odcięcia. Drugi natomiast to kontrola obliczanego potencjału energetycznego. Potencjał LJ charakteryzuje się tym, że w miarę zbliżania się do 0 (oś x) jego wartości drastycznie rosną (rys 3.8). Aby zapobiec niekontrolowanemu zachowaniu atomów przy tak dużych wartościach potencjału został on ograniczony do maksymalnej wartości tak, aby analizowana struktura nie uległa zniszczeniu nawet przy bardzo bliskim kontakcie atomów.

RYS LJ EXCEL

Rys 3.8. Potencjał Lenard-Jones dla parametrów użytych w symulacji.

## Optymalizacja

Pierwszym etapem optymalizacji była redukcja przejść pętli w całej symulacji. Dzięki temu znacznie udało się zredukować liczbę wykonywanych operacji. Redukcja dotyczyła połączenia dwóch etapów: przepisania danych ze struktury wyjściowej na wejściową z przygotowaniem danych do wyświetlenia. (Kod

|  |
| --- |
| \_\_global\_\_ void update\_structure\_and\_display(float4 \*pos, Structure \*input, Structure \*output) {  int tid = threadIdx.x + blockIdx.x \* blockDim.x;  float u, v, w;  input->atomsCount = output->atomsCount;  for (int i=tid ; i<input->atomsCount ; i+=blockDim.x \* gridDim.x) {  input->atoms[i].pos.x = output->atoms[i].pos.x;  input->atoms[i].pos.y = output->atoms[i].pos.y;  input->atoms[i].pos.z = output->atoms[i].pos.z;  input->atoms[i].force = output->atoms[i].force;  input->atoms[i].acceleration = output->atoms[i].acceleration;  input->atoms[i].status = output->atoms[i].status;  input->atoms[i].fixed = output->atoms[i].fixed;  u = input->atoms[i].pos.x \* 0.1f;  w = input->atoms[i].pos.y \* 0.1f;  v = input->atoms[i].pos.z \* 0.1f;  pos[i] = make\_float4(u, w, v, 1.0f);  }  } |

Kod 3.4. Optymalizacja aktualizacji danych struktury oraz wyświetlania.

Analizując zmieniony kod 3.4. można zauważyć, że 3 pętle for stosowane do uaktualniania struktury zostały zastąpione pojedynczą. Dodatkowo pętla ta została zrównoleglona podobnie jak w kodzie 3.3. Pomimo, że funkcja update\_structure\_and\_display nie zawiera wielu operacji zmiennoprzecinkowych, to powyższe zabiegi znacznie zwiększyły wydajność programu co zostanie zaprezentowane w rozdziale 4 podczas analizie wydajności. Świadczy to o tym, że czas dostępu do pamięci przy dużej ilości operacji może być istotnym parametrem optymalizacji. Do tej pory wszystkie dane były przechowywane w pamięci RAM karty graficznej. Warto więc zastanowić się nad wykorzystaniem pamięci SHARED, czyli bardzo szybkiej, lecz jednocześnie małej, pamięci współdzielonej lub nawet rejestrów GPU.

**//TODO---------------------------------------------------Further optimization**

xx

# Testy wydajności

Testy wydajności będą wykonywane przy stałej konfiguracji, aby wyniki były miarodajne i porównywalne pomiędzy testami. Po uzyskaniu najszybszego i najbardziej wydajnego algorytmu wykonane zostaną testy na klastrze obliczeniowym, gdzie zbadany będzie również temat analizy skalowalności oraz modelu algorytm-urządzenie. Powyższe zagadnienia zawarte będą w rozdziale 5.

Analiza i optymalizacja będzie wykonywane przy wykorzystaniu poniższej konfiguracji sprzętowej:

|  |  |
| --- | --- |
| GPU | **Nvidia Quadro NVS 5400M** |
| CPU | Intel(R) Core(TM) i3-3110M CPU @ 2.40GHz |
| RAM | DDR3 8GB 1333MHz |
| HDD | 500GB 7200 obrotów |

Tab 4.1. Konfiguracja sprzętowa – optymalizacja i analiza wydajności

Do wykonywania testów, budowania kodu i konfiguracji środowiska wykorzystane zostało poniższe oprogramowanie:

|  |  |
| --- | --- |
| System operacyjny | Debian 3.2.57-3+deb7u2 x86\_64 GNU/Linux kernel #3.2.0-4-amd64 |
| CUDA | Nvidia CUDA 5.5 + Nvidia CUDA Toolkit |
| Sterowniki GPU | Nvidia Driver Version 319.60 |
| OpenGL | 1.4 |
| g++ | g++ (Debian 4.7.2-5) 4.7.2 |

Tab 4.2. Konfiguracja oprogramowania – optymalizacja i analiza wydajności

Konfiguracja wejściowa programu wykonana na podstawie pliku *structure.cfg*:

|  |  |
| --- | --- |
| dimX | 15 |
| dimY | 15 |
| dimZ | 15 |
| force | 1 |
| forceType | ALL\_AROUND |

Tab 4.2. Konfiguracja oprogramowania – optymalizacja i analiza wydajności

Każda symulacja zostanie wykonana 100 krotnie. Zostanie wyciągnięty średni wynik spośród wszystkich testów. Pozwoli to na uzyskanie optymalnych wyników.

Rozdział 4 prezentuje jedynie testy wydajności. Nie będą poruszane tutaj sprawy modelowania defektów i graficzne przedstawienie rozwiązywanych problemów. Odkształcenia i ich modelowanie szerzej opisane zostało w osobnym rozdziale.

## Pomiar czasu

W celu uzyskania jak najdokładniejszego pomiaru czasu wykorzystany został wbudowany w biblioteki CUDA mechanizm. Opiera się on na zdarzeniach, które przechwytują dokładny czas na starcie i końcu okresu pomiarowego. Wbudowana funkcja na podstawie wyników subskrypcji tych dwóch zdarzeń określa ile upłynęło pomiędzy nimi czasu.

|  |
| --- |
| cudaEvent\_t start;  handleTimerError(**cudaEventCreate**(&start), START\_CREATE);  cudaEvent\_t stop;  handleTimerError(**cudaEventCreate**(&stop), STOP\_CREATE);    handleTimerError(**cudaEventRecord**(start, NULL), START\_RECORD);  /\*  \*\* OBLICZENIA  \*/  cudaDeviceSynchronize();  handleTimerError(**cudaEventRecord**(stop, NULL), STOP\_RECORD);  handleTimerError(**cudaEventSynchronize**(stop), SYNCHRONIZE);  handleTimerError(**cudaEventElapsedTime**(&msecTotal, start, stop), ELAPSED\_TIME); |

Kod 4.1. Pomiar czasu symulacji.

Powyższy przykład przedstawia szablon jaki jest stosowany do pomiaru czasu. Zawiera wszystkie niezbędne funkcje do przechwytywania, synchronizacji i obliczania czasu wraz z obsługą błędów. Po zebraniu danych w taki sposób można przeprowadzić szacowanie uzyskanej wydajności. Wymaga to, oprócz zmierzonego czasu, również ręcznego obliczenia ilości operacji wykonanych w algorytmie. Dzięki temu możliwe będzie wyliczanie wielkości GFLOPS czyli ilości operacji na sekundę.

|  |
| --- |
| float msecPerSimulation = msecTotal / nIter;  double flopsPerSimulation = 55.0 \* structure->atomsCount \* structure->atomsCount + 9.0 \* structure->atomsCount + 2 \* (structure->atomsCount + 256 - 1 )/ 256;  double gigaFlops = (flopsPerSimulation \* 1.0e-9f) / (msecPerSimulation / 1000.0f);  printf("\n\t\tPerformance= %.2f GFlop/s, Time= %.3f msec, Size= %.0f Ops, WorkgroupSize= %u threads/block\n",  gigaFlops,  msecPerSimulation,  flopsPerSimulation,  block.x \* block.y);  ///////////////// Wynik przykładowego wykonania powyższego kodu////////////////////  Performance= 14.68 GFlop/s, Time= 42.675 msec, Size= 626514778 Ops, WorkgroupSize= 256 threads/block |

Kod 4.2. Obliczanie wydajności uzyskanej przez algorytm.

Kod 4.2 jest przykładem obliczenia wydajności. Istotnym elementem jest wartość zaznaczona grubą czcionką. Jest to bardzo istotny parametr – średni czas dostępu do pamięci (w tym wypadku RAM) na GPU. Jest to wartość dokładna, określona przez deweloperów Nvidii, która znajduje się między innymi w przykładach dołączonych do CUDA Toolkit. Wynik wykonania kodu przedstawia wartość GFLOPS, czas wykonania pojedynczego przejścia algorytmu, ilość operacji jakie zostały wykonane w trakcie obliczeń oraz ilość wątków pracujących w ramach jednego bloku.

## Testy sekwencyjne

Testy sekwencyjnego przebiegu algorytmu przy wykorzystaniu procesorów GPU zawsze są dyskusyjne. Na potrzeby analizowanych problemów ‘testy sekwencyjne’ będą uważane za punkt odniesienia, do obliczania przyspieszenia czy skalowalności. Nie jest tu zakładane, że operacja zostanie wykonana przez dokładnie jeden wątek na jednym multiprocesorze. Stosując nomenklaturę CUDA, testy sekwencyjne są uważane za wywołanie kernela z następującymi parametrami:

|  |
| --- |
| Kernel\_GPU<<< 1, 1 >>>(…); |

Kod 4.3. Parametry kerenla dla testów sekwencyjnych.

Takie wykonanie funkcji GPU sprawi, że zostanie ona wykonana przez 1 blok z 1 wątkiem w ramach jednego bloku. W czasie działania programu może jednak dojść do przełączania kontekstu pomiędzy wątkami oraz pomiędzy multiprocesorami. Niestety nie ma możliwości kontrolowania takich sytuacji.

## Testy zrównoleglenia

TODO

## Testy optymalizacji

TODO

## Test wielu GPU

TODO

# Model algorytm - urządzenie

## XXX

# Podsumowanie

# Bibliografia

[1] Tomasz Nowak, *Modelowanie defektów strukturalnych w skali nano z wykorzystaniem akceleratorów obliczeń GPGPU*

[2] Daniel Bachniak, *Modelowanie defektów strukturalnych w skali nano z wykorzystaniem metody statyki molekularnej w heterogenicznych architekturach sprzętowych*

[3] https://developer.nvidia.com (05.04.2014)

[4] http://www.nvidia.com/ (05.04.2014)

[5]https://computing.llnl.gov/ (05.04.2014)

[6] http://en.wikipedia.org/wiki/POSIX\_Threads (05.04.2014)

[7]http://wakespace.lib.wfu.edu/bitstream/handle/10339/38561/Proctor\_wfu\_0248M\_10427.pdf (10.08.2014)

[8] http://ti.odio.twiki.di.uniroma1.it/pub/CI/WebHome/ParMolDyn.pdf (28.08.2014

[9] http://en.wikipedia.org/wiki/Molecular\_dynamics

1. API – eng. application programming interface [↑](#footnote-ref-2)
2. Windows API – zestaw wbudowanych interfejsów programistycznych dostępnych dla systemów Microsoft Windows [↑](#footnote-ref-3)