AKADEMIA GÓRNICZO-HUTNICZA

IM. STANISŁAWA STASZICA W KRAKOWIE



Wydział Inżynierii Metali i Informatyki Przemysłowej



PROJEKT INŻYNIERSKI

pt.

„Realizacja frontalnego solwera MES z wykorzystaniem technologii OpenCL”

Imię i nazwisko dyplomanta: **Paweł Wal**

Kierunek studiów: **Informatyka Stosowana**

Nr albumu: **240202**

Opiekun: dr inż. Łukasz Rauch

Podpis dyplomanta: Podpis opiekuna:

Kraków 2014

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

Kraków, dnia ………….… Podpis dyplomanta…………….

Spis treści

[1 WSTĘP 3](#_Toc375576077)

[2 WPROWADZENIE TEORETYCZNE 5](#_Toc375576078)

[2.1 METODY ROZWIĄZYWANIA UKŁADÓW RÓWNAŃ LINIOWYCH 5](#_Toc375576079)

[2.2 METODA ELIMINACJI GAUSSA 6](#_Toc375576080)

[2.2.1 ELIMINACJA W PRZÓD 6](#_Toc375576081)

[2.2.2 PODSTAWIANIE WSTECZ 8](#_Toc375576082)

[2.3 MACIERZE 10](#_Toc375576083)

[2.3.1 CHARAKTERYSTYKA MACIERZY W METODZIE ELEMENTÓW SKOŃCZONYCH 11](#_Toc375576084)

[2.3.2 METODY PRZECHOWYWANIA MACIERZY RZADKICH 13](#_Toc375576085)

[2.4 ARCHITEKTURA OPENCL 14](#_Toc375576086)

[2.4.1 PARALELIZM DANYCH 15](#_Toc375576087)

[3 IMPLEMENTACJA SOLWERA 19](#_Toc375576088)

[3.1 ANALIZA ELEMENTÓW SOLWERA 19](#_Toc375576089)

[3.1.1 MOŻLIWOŚCI ZRÓWNOLEGLENIA METODY GAUSSA 19](#_Toc375576090)

[4 BADANIA WYDAJNOŚCI 19](#_Toc375576091)

[5 WNIOSKI 19](#_Toc375576092)

[6 BIBLIOGRAFIA 19](#_Toc375576093)

# WSTĘP

Ponad czterdzieści lat temu, w roku 1970, Bruce Irons opublikował swoją koncepcję programu rozwiązującego układy równań w postaci macierzy rzadkich metodą frontalną. Jego główną motywacją do opracowania tej techniki była nie tylko wydajność, ale również zużycie pamięci[5]. Irons pracował z komputerem ICT 1905, który w najlepszym razie mógł mieć 96 kilobajtów pamięci operacyjnej w postaci 32,768 słów po 24 bity. Nic więc w tym dziwnego, że poszukiwał rozwiązania jak najbardziej optymalnego pamięciowo.

Dziś co prawda pamięć jest znacznie mniej ograniczonym zasobem – przeciętny komputer biurowy dysponuje na ogół kilkoma gigabajtami, większość kart które wykorzystuje się do obliczeń nie będzie posiadała mniej niż gigabajt. Jak jednakże mówi przysłowie – apetyt rośnie w miarę jedzenia. Wraz ze wzrostem pamięci i mocy obliczeniowej wzrastają również rozmiary problemów – w tym tych z zakresu metody elementów skończonych - na które zasadzają się badacze.

Wraz z biegiem lat technologia obliczeniowa ewoluowała, kulminując obecnie w wielordzeniowych procesorach i koprocesorach, oraz – z samej swojej natury przystosowanych do operacji zmiennoprzecinkowych – kartach graficznych wykorzystywanych w zagadnieniach GPGPU. Szczególnie te ostatnie stanowią interesującą materię. Składają się z jednego lub więcej procesorów strumieniujących, z których każdy dysponuje na ogół dużą ilością (powyżej stu) rdzeni obliczeniowych, mogących wykonywać obliczenia już nie tylko równolegle, ale w sposób masowo równoległy, niedostępny dla urządzeń opartych na klasycznych procesorach. Dla kodu obliczeniowego uruchamianego na takich urządzeniach istnieje jednak szereg ograniczeń[4]: transfery z pamięci dostępu ogólnego (pamięci hosta) do pamięci urządzenia są kosztowne czasowo gdyż wymuszają okresy w których ani CPU, ani GPU nie wykonują obliczeń, operacje atomowe na pamięci globalnej są powolne, problem stanowi również nadmierna dywergencja wątków.

Niniejsza praca jest poświęcona implementacji solwera podążającego za ideą solwera frontalnego zaprezentowanego w 1970 roku przez Ironsa, rozumianej jako rozłożenie złożonego problemu na serię częściowo tylko od siebie zależnych, lżejszych pamięciowo i znacznie prostszych obliczeniowo pod-problemów. Celem przyświecającym tej pracy jest również stworzenie oprogramowania wykorzystującego w najlepszy dostępny sposób możliwości masowej równoległości oferowane przez nowoczesne, wysokowydajne urządzenia obliczeniowe.

Mimo iż oprogramowanie podąża za myślą Ironsa, postulowana przez niego metoda rozwiązywania układów równań liniowych nie została dosłownie zastosowana. Zasada matematyczna na której opiera się stworzone oprogramowanie jest wyprowadzona ze zmodyfikowanej metody eliminacji Gaussa. Modyfikacje mają na celu wykorzystanie mocnych stron zastosowanych urządzeń obliczeniowych, jednocześnie omijając ich ograniczenia i potencjalne słabości.

Do realizacji samego oprogramowania została wykorzystana architektura i zestaw bibliotek OpenCL. Każdy z wiodących producentów sprzętu obliczeniowego ma zazwyczaj swoją własną architekturę którą obsługują jego urządzenia – jest to na przykład oprogramowanie CUDA dla urządzeń firmy NVIDIA czy Stream od firmy ATI. W odróżnieniu od nich OpenCL jest otwartym standardem, którego implementacje dla konkretnych urządzeń należą wprawdzie do ich producentów, lecz jego zastosowanie jest możliwe na urządzeniach NVIDIA, ATI, Intel, a nawet na niektórych procesorach w architekturze ARM pod kontrolą systemu Android[12]. Standard OpenCL dostarcza warstwę abstrakcji, dzięki której urządzenie na którym wykonywany jest kod nie ma znaczenia dla twórcy oprogramowania, tak długo jak implementacja OpenCL dla tej platformy jest z nim zgodna.

Stworzony kod został przetestowany pod kątem wydajności dla szeregu macierzy o różnych rozmiarach i charakterystykach. [TODO: wymienić urządzenia, GT550Ti, GT540M, Tesla M2090, Intel(R) Xeon(R) CPU X5650 @ 2.67GHz, xeon phi?]

# WPROWADZENIE TEORETYCZNE

## METODY ROZWIĄZYWANIA UKŁADÓW RÓWNAŃ LINIOWYCH

Koniunkcję pewnej liczby równań liniowych w skład których wchodzi ten sam zestaw zmiennych nazywa się układem równań liniowych. Układ takowy można zapisać w postaci równania macierzowego:

Bądź jego ogólnej reprezentacji macierzowej:

Ze względu na wygodę, w przykładach umieszczanych w literaturze często stosowana jest również reprezentacja w postaci macierzy rozszerzonej:

Ze względu na czytelność, powyższa forma będzie stosowana w pozostałej części tego dokumentu.

Zostało zaproponowanych wiele metod ich rozwiązywania. Oprogramowanie komputerowe służące do ich rozwiązywania nazywane jest solwerami.

Metody służące do rozwiązywania układów równań liniowych można podzielić na dwie główne kategorie[7]:

1. Metody iteracyjne
2. Metody bezpośrednie

Do metod bezpośrednich zaliczana jest między innymi metoda eliminacji Gaussa oraz metoda faktoryzacji (dekompozycji) LU.

## METODA ELIMINACJI GAUSSA

Metoda eliminacji Gaussa składa się z dwóch zasadniczych faz, to jest fazy eliminacji w przód oraz fazy eliminacji wstecz. Uważna analiza obu tych faz pozwala wysnuć kluczowe w modyfikacji algorytmu dla oprogramowania równoległego wnioski.

### ELIMINACJA W PRZÓD

Pierwsza z nich opiera się na liniowej operacji algebraicznej[9], czyli na wiedzy, iż każde równanie w układzie równań liniowych może zostać zastąpione równaniem powstałym z połączenia tego równania z dowolnym innym występującym w tym układzie.

Przedstawiając układ równań w postaci macierzowej, daje to podstawy do stosowania jednej z macierzowych operacji elementarnych, czyli dodawania lub odejmowania od siebie wierszy bądź ich wielokrotności.

Celem fazy eliminacji w przód jest doprowadzenie układu równań w postaci macierzowej do górnej macierzy trójkątnej (znanej również jako macierz schodkowa). Dokonuje się tego eliminując - przy pomocy operacji elementarnych – pewną liczbę niewiadomych z poszczególnych równań, reprezentowanych przez wiersze w macierzy [1,9].

Prześledźmy przykład fazy eliminacji w przód za [1]. Na rysunku 1 dany jest układ równań w postaci macierzy rozszerzonej:

Rysunek 1. Układ równań liniowych w postaci macierzowej

Jest to układ równań wygenerowany przez prosty problem metody elementów skończonych. Na pierwszy rzut oka widać że jest symetryczna oraz pasmowa – mimo że pasmo to jest bardzo szerokie. Są to dwie przydatne właściwości macierzowych postaci równań opisujących problemy metody elementów skończonych.

Pierwszym krokiem eliminacji w przód jest wyeliminowanie zmiennych oraz przy pomocy pierwszego wiersza macierzy pomnożonego przez odpowiednią liczbę. Po przeprowadzeniu tej operacji, macierz wygląda jak na rysunku 2.

Rysunek 2. Układ równań liniowych w postaci macierzowej po pierwszym kroku eliminacji w przód

Następnym krokiem jest postąpienie analogicznie w stosunku do zmiennych oraz , co zostało uwidocznione na rysunku 3.

Rysunek 3. Układ równań liniowych w postaci macierzowej po drugim kroku eliminacji w przód

Ostatnim krokiem w przypadku tej przykładowej macierzy jest zredukowanie wyrazu , jak zostało to uwidocznione na rysunku 4.

Rysunek 4. Układ równań liniowych w postaci macierzowej po zakończeniu eliminacji w przód

Analizując przebieg tego przykładu można wysnuć dwa pomocne wnioski. Dla każdego równania, czyli wiersza macierzy, można – przy założeniu wykonywania operacji w sposób sekwencyjny – wyznaczyć ilość operacji elementarnych które są konieczne do doprowadzenia go do pożądanej postaci. W powyższym przykładzie dla wiersza 1 było to zero operacji elementarnych, wiersz numer 2 wymagał jednej operacji elementarnej, zaś wiersze 3 i 4 – po dwie. Wychodząc z powyższego wniosku postawić można kolejny - iż dla danej macierzy można szybko wyznaczyć maksymalną ilość operacji elementarnych konieczną do sprowadzenia jej do macierzy schodkowej przypadającej na wiersz. Jak łatwo zauważyć, ta ilość operacji jest związana bezpośrednio z szerokością pasma macierzy: .

Drugi wniosek wyprowadzić można z faktu, iż eliminacja w przód sprowadza macierz do postaci macierzy schodkowej. Cechą charakterystyczną tej macierzy jest to, iż pierwszy wyraz niezerowy w danym wierszu musi znajdować się na diagonali macierzy; bezpośrednio oznacza to, że żadne dwa wiersze nie mogą mieć identycznego pierwszego wyrazu niezerowego.

### PODSTAWIANIE WSTECZ

Drugą z faz rozwiązania układu równań liniowych przy pomocy metody Gaussa jest faza podstawiania wstecz (ang. back substitution). Etap ten przeprowadzany jest na macierzy sprowadzonej do górnej macierzy trójkątnej, czyli macierzy w postaci schodkowej utworzonej w fazie eliminacji w przód. Rysunek 5 przedstawia macierz w formie schodkowej gotową do przeprowadzenia fazy podstawiania wstecz.

Rysunek 5. Układ równań liniowych w postaci macierzy schodkowej

Kontynuując analizę tego przykładu za [9], zauważyć można iż wartość wyrazu może zostać wyliczona bez przeprowadzania żadnych dodatkowych operacji, podczas gdy pozostałe wyrazy wymagają operacji odpowiednio więcej. Wyraz zostaje więc wyliczony poprzez obustronne dzielenie jak uwidoczniono na rysunku 6.

Rysunek 6. Macierz schodkowa po przeprowadzeniu pierwszego kroku podstawienia wstecz

Wiedząc iż , możliwe jest teraz wyliczenie za pomocą równania danego trzecim wierszem macierzy: . Po wykonaniu odpowiednich operacji, uzyskujemy wynikową macierz przedstawioną na rysunku 7.

Rysunek 7. Macierz schodkowa po przeprowadzeniu drugiego kroku podstawienia wstecz

W drodze analogicznych działań dla dwóch pozostałych poszukiwanych wyrazów, oraz – danych równaniami opartymi odpowiednio na drugim i pierwszym wierszu macierzy w formie zaprezentowanej na rysunku 7 – uzyskiwana jest ostatecznie macierz przedstawiona na rysunku 8.

Rysunek 8. Finalna postać macierzy – rozwiązany układ równań liniowych

Macierz w tej postaci zawiera wyłącznie rozwiązanie układu równań liniowych, czyli cel zastosowania metody eliminacji Gaussa.

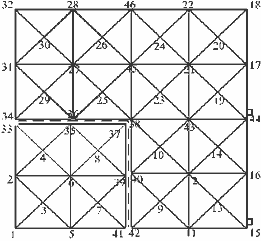
Na podstawie powyższego przykładu można wysnuć kolejny istotny wniosek. W przypadku fazy podstawiania wstecz, każdy kolejny krok – wyliczenie kolejnego wyrazu – wymaga wyliczenia wszystkich wyrazów którymi jest on dany, czyli inaczej wszystkich wyrazów od których jest zależny. Inaczej mówiąc, wyliczenie wyrazu wymaga – co można przyjąć jako ogólną regułę – wyliczenia wyrazów .

Drugim istotnym spostrzeżeniem jest fakt, iż kolejność rozwiązywania równań z dołu do góry, typowa dla górnej macierzy trójkątnej, jest pewną abstrakcją. W istocie nieistotnym jest, którym wierszem macierzy dany jest który wyraz rozwiązywanego układu równań, tak długo jak istnieje metoda identyfikacji właściwego dla danego wyrazu wiersza.

## MACIERZE

Proponowane rozwiązanie ma służyć przede wszystkim jako solwer dla układów równań liniowych przedstawionych w postaci macierzowej wygenerowanych przez oprogramowanie rozwiązujące problemy z zakresu metody elementów skończonych. Wychodząc z tego założenia można wyciągnąć pewne wnioski co do charakteru rozpatrywanych macierzy.

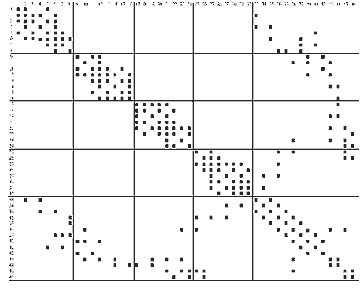
### CHARAKTERYSTYKA MACIERZY W METODZIE ELEMENTÓW SKOŃCZONYCH



Rysunek 8. Siatka elementów skończonych[[1]](#footnote-1)

Analizując siatkę elementów skończonych przedstawioną na rysunku 8, łatwo zauważyć iż każdy element, a nawet cała siatka jako taka, może być postrzegana jako rodzaj grafu nieskierowanego. Globalna macierz sztywności w metodzie elementu skończonego jest konstruowana na podstawie umieszczenia w niej w odpowiednich punktach – wynikających z numeracji węzłów w elementach oraz numeracji samych elementów – wartości z lokalnych macierzy sztywności wyliczonych dla poszczególnych elementów[8].

Wychodząc z powyższych założeń można powiedzieć, że macierz sztywności tworzona w metodzie elementów skończonych jest w swojej strukturze podobna do macierzy sąsiedztwa dla grafu nieskierowanego, które to macierze są inherentnie symetryczne. Intuicyjnie „jeśli wierzchołek 1 jest sąsiadem wierzchołka 2 to wierzchołek 2 jest sąsiadem wierzchołka 1”. Identyczną logikę – zamieniając jedynie słowo „wierzchołek” na „węzeł” – można zastosować w przypadku siatki elementów skończonych. Dzięki temu uzyskujemy pewność iż wygenerowana macierz będzie symetryczna.



Rysunek 9. Macierz sztywności dla metody elementów skończonych[[2]](#footnote-2)

Na rysunku 9 została uwidoczniona macierz sztywności dla przedstawionej powyżej siatki elementów skończonych. Na tym przykładzie dobrze uwidoczniona jest druga z interesujących cech macierzy którymi posługuje się metoda elementów skończonych: macierze te są rzadkie. W praktyce oznacza to iż wiele z ich elementów to zera.

W pracy[10] Reginald Tewarson postulował że macierz rzadka to taka, która przy wymiarach ma około niezerowych elementów, w praktyce – dwa do dziesięciu elementów niezerowych przypadających na każdy wiersz dla dużych . Są to wprawdzie definicje skonstruowane w zupełnie innych czasach – praca Tewarsona została opublikowana w 1973 roku – i niezbyt precyzyjne (nie jest na przykład sprecyzowane co autor miał na myśli przez duże ), lecz pozwalają na wytworzenie definicji intuicyjnej: macierz jest rzadka jeśli więcej niż połowa jej elementów to zera. Praktycznie każda macierz wygenerowana przez metodę elementów skończonych spełnia to założenie.

Specyficzną formą macierzy rzadkich są macierze pasmowe, czyli takie macierze rzadkie w których elementy niezerowe są skupione w paśmie ułożonym na przekątnej, którego środek wyznacza główna przekątna macierzy i zero lub więcej przekątnych po obu jej stronach. Macierze pasmowe bardzo dobrze nadają się do zastosowania metody Gaussa i jej pochodnych, gdyż z samej ich definicji wynika, iż niewiele operacji jest koniecznych by sprowadzić je do postaci macierzy schodkowej – wymagana jest jedynie eliminacja elementów znajdujących się pod główną przekątną macierzy.

W metodzie elementu skończonego uzyskanie macierzy pasmowych wymaga odpowiedniego ponumerowania elementów i węzłów. Niektóre solwery[3] integrują się z rozwiązaniem samego problemu metody elementów skończonych do tego stopnia że same zmieniają numerację węzłów i elementów by uzyskać korzystniejszą dla ich metody działania strukturę macierzy. Ujmuje to jednak takim rozwiązaniom uniwersalności i utrudnia ich implementację w rozwiązaniach zewnętrznych.

### METODY PRZECHOWYWANIA MACIERZY RZADKICH

Duże macierze rzadkie na ogół przechowuje się w pamięci komputera w postaci skompresowanej[10]. Pozwala to przechować większą macierz niż pozwalałaby na to – przy klasycznych metodach składowania – pamięć zastosowanego urządzenia, lub – odwrotnie – przechować daną macierz kosztem mniejszej ilości pamięci.

W kontekście solwera działającego przeważnie na kartach graficznych zużycie pamięci jest istotne z co najmniej dwóch powodów. Po pierwsze, dzięki kompresji można umieścić w pamięci urządzenia obliczeniowego większą macierz na raz. Po drugie, szybkość całkowitego działania solwera będzie związana między innymi z czasem transferu danych na urządzenie za pośrednictwem szyny danych PCI-Express[4]. Z tych dwóch względów konieczne jest rozważenie najpopularniejszych schematów przechowywania macierzy rzadkich.

Jednym z najprostszych schematów jest metoda koordynatowa (Coordinate Format), w której elementy niezerowe macierzy rzadkiej przechowywane są w formie tripletów , gdzie oraz stanowią odpowiednio wiersz i kolumnę w której znajduje się wartość . O ile najprostsza, metoda ta zużywa miejsc w pamięci (dla uproszczenia kalkulacji pominięta jest kwestia stricte implementacyjna, tj. użytych typów danych), gdzie to ilość niezerowych elementów w macierzy.

Wariacją na temat metody koordynatowej i jednym z popularniejszych schematów przechowywania macierzy rzadkich, jest metoda Compressed Sparse Row Format[[3]](#footnote-3) (CSR). W tej metodzie przechowywane w pamięci są trzy wektory.

Pierwszy z nich, o długości , zawiera niezerowe wartości z macierzy ułożone w kolejności od lewej do prawej oraz od góry do dołu macierzy. Drugi wektor zawiera indeksy kolumn w których konkretne wartości w pierwszym wektorze znajdowały się w macierzy (we właściwych sobie wierszach). Trzeci wektor zaś zawiera informację o tym od których indeksów zaczynają się poszczególne wiersze w dwóch poprzednich wektorach.

Przykład metody CSR przechowywania macierzy został przedstawiony na rysunku 10.

AA: 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0

JA: 1 4 1 2 4 1 3 4 5 3 4 5

IA: 1 3 6 10 12 13

Rysunek 10. Przykład macierzy zapisanej w formacie CSR

W tym przypadku ilość zużytej pamięci wynosi – dwa pierwsze wektory muszą mieć długość równą ilości niezerowych wartości w macierzy, trzeci zaś w typowym dla metody elementów skończonych przypadku – gdzie w macierzy nie ma wierszy w których nie ma żadnych danych – ma długość równą wymiarowi macierzy.

Ponieważ z reguły niezerowych elementów w macierzy wygenerowanej w metodzie elementów skończonych będzie więcej niż wynosi wymiar macierzy , metoda ta wygrywa kompresją z metodą koordynatową ( jest mniejsze niż ). Istnieją również wariacje na temat tej metody, najoczywistszym z których jest format Compressed Sparse Column (CSC), różniący się tym, iż drugi wektor przechowuje indeksy wierszy, a trzeci początki kolumn w pozostałych wektorach.

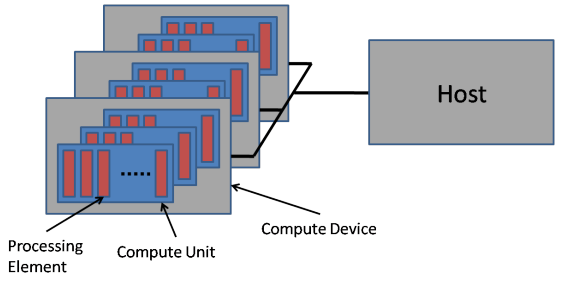
Podczas śledzenia zachowania metody Gaussa oczywistym stało się, iż ze względu na niemożność dynamicznej alokacji pamięci na urządzeniu oraz fakt iż popularne metody kompresji macierzy nie przewidują alokacji pamięci dla elementów zerowych które mogą ulec zmianie, dla solwera będzie konieczne zaprojektowanie uwzględniającej to metody kompresji macierzy stanowiącej kompromis pomiędzy wydajnością pamięciową, transferu a obliczeniową.

## ARCHITEKTURA OPENCL

Jak wspomniano we wstępie do niniejszej pracy, OpenCL (lub Open Computing Language) jest stworzonym przez Khronos Group otwartym standardem tworzenia oprogramowania dla nowoczesnych urządzeń obliczeniowych. Architektura ta posiada kilka specyficznych cech, których zrozumienie jest konieczne dla jej skutecznego wykorzystania.

### SKŁADNIKI ŚRODOWISKA OPENCL

Idea funkcjonowania platformy OpenCL została przedstawiona na rysunku 11. Środowisko składa się z hosta – „gospodarza”, który zleca wykonanie konkretnych obliczeń jednemu lub więcej urządzeniom obliczeniowym. Każde z nich posiada wiele jednostek obliczeniowych, każda z których z kolei posiada więcej niż jeden element przetwarzający.



Rysunek 11. Składniki środowiska dla architektury OpenCL[6]

### SKŁADNIKI ARCHITEKTURY OPENCL

Sama architektura składa się z trzech zasadniczych części[6]: specyfikacji języka, API platformy i API czasu wykonania.

Specyfikacja języka OpenCL definiuje składnię programów i kerneli które zostaną uruchomione z użyciem pozostałych dwóch części architektury. Jest oparty na standardzie ISO C99, lecz ze zmodyfikowanymi, dodanymi lub usuniętymi słowami kluczowymi, a także bez niektórych funkcjonalności.

### PARALELIZM DANYCH

OpenCL – podobnie jak karty graficzne, które są główną grupą urządzeń na których wykorzystuje się tą technologię – działa na zasadzie paralelizmu danych[13]. W praktyce oznacza to iż w przeciwieństwie do paralelizmu zadań, który zakłada wykonanie różnych zadań w tym samym czasie, OpenCL zakłada wielokrotne wykonanie identycznego (lub podobnego) zadania na elementach pewnego zbioru danych.

W tabeli 1 zostało przedstawione porównanie standardowego kodu skalarnego w języku C, oraz kodu paralelnego względem danych w języku OpenCL.

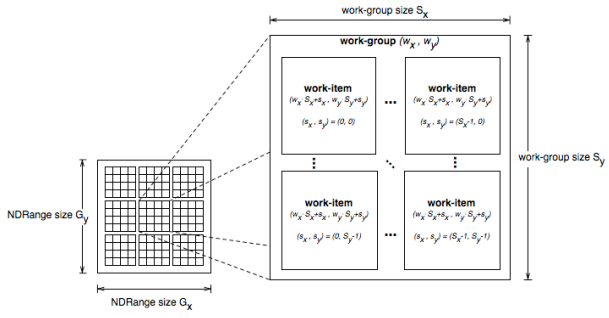
Tabela 1. Porównanie kodu paralelnego względem danych z funkcją skalarną (za [6]).

|  |  |
| --- | --- |
| Skalarna funkcja w języku C | Funkcja paralelna względem danych |
| void sqrt(int n, const float \*a, float \*res)  {  int i;  for(i=0; i<n; i++)  res[i] = a[i]\*a[i];  } | kernel void dp\_square  (global const float \*a, global float \*res)  {  int id = get\_global\_id(0);  res[id] = a[id] \* a[id];  } |

Paralelizm danych jest w OpenCL realizowany za pośrednictwem programów, składających się z jednego lub więcej kerneli. Program może też zawierać dodatkowe funkcje wykorzystywane przez kernele oraz stałe dane. Podczas wykonania danego kernela wiele jego kopii jest równolegle wykonywanych na elementach przetwarzających urządzenia obliczeniowego. Wykonania te zwane są work-itemami[13]. W dalszej części niniejszej pracy będą wykorzystywane terminy „zadanie” by określić work-item i „grupa robocza” by określić work-group.

### ZADANIA I GRUPY ROBOCZE

Zadania są powiązane ze sobą w grupy robocze, jak uwidoczniono na rysunku 12.



Rysunek 12. Zadania i grupy robocze w OpenCL [13].

Przedstawiona po lewej stronie powierzchnia – NDRange – to metoda organizacji pamięci i rozplanowania zadań w architekturze OpenCL. NDRange to N-wymiarowa przestrzeń (OpenCL pozwala na wykorzystanie jedno-, dwu- lub trójwymiarowej przestrzeni). Przestrzeń ta jest dzielona na grupy robocze.

Każda z grup roboczych dysponuje indeksami lokującymi ją w konkretnym punkcie NDRange. Każde zadanie – stanowiące odrębną instancję kernela – dysponuje unikalnym numerem identyfikacyjnym globalnym, oraz unikalnym wewnątrz grupy roboczej identyfikatorem lokalnym.

Z punktu widzenia paralelizmu danych, NDRange jest swoistym mapowaniem konkretnych instancji kernela na konkretne elementy obszaru pamięci który ma zostać przetworzony. Do dodania do siebie dwóch macierzy N x N można wykorzystać kernele uruchomione przy dwuwymiarowym NDRange z Gx = Gy = N. Do wyliczenia iloczynu skalarnego z kolei należałoby wykorzystać kernele uruchomione na jednowymiarowym NDRange.

Podział zadań na grupy robocze służy też do podziału barier na dwa rodzaje. W kernelach OpenCL można wykorzystywać bariery lokalne, które synchronizują tylko wątki wewnątrz grupy roboczej, bądź bariery globalne, które zsynchronizują wszystkie wątki działające na NDRange.

### ZARZĄDZANIE PAMIĘCIĄ

Jednym z istotnych założeń architektury OpenCL o którym należy pamiętać projektując dla niej oprogramowanie jest fakt, iż pojedyncza grupa robocza będzie wykonywana symultanicznie na jednej jednostce obliczeniowej (a wielu elementach przetwarzających). Ta wiedza przydatna jest podczas analizy modelu pamięci OpenCL.

W OpenCL wyróżniamy cztery osobne przestrzenie w pamięci[13]. Są to:

1. **Pamięć stała.** W tej kategorii pamięci przechowywane są wartości niezmienne podczas całego wykonania kernela. Może ona zostać zainicjalizowana przez „gospodarza”.
2. **Pamięć globalna.** Do tej pamięci mają dostęp (zapis i odczyt) wszystkie wątki we wszystkich grupach roboczych. Jeśli urządzenie na to zezwala, operacje na tej pamięci mogą być cache’owane. Jest to zazwyczaj największy dostępny obszar pamięci na urządzeniu obliczeniowym, ale najczęściej oferuje również najwyższy czas dostępu.
3. **Pamięć prywatna.** Do tego rodzaju pamięci ma dostęp tylko pojedyncze zadanie – to, które ją zaalokowało.
4. **Pamięć lokalna.** Pamięć współdzielona przez wszystkie wątki w danej grupie roboczej, niedostępna poza grupą roboczą. Jeśli jest to wspierane przez konkretne urządzenie, może być mapowana na obszar pamięci cache przy konkretnej jednostce obliczeniowej, dzięki czemu dostęp do takiej pamięci jest znacznie szybszy dla danej jednostki i uruchomionych na niej zadań. Jeśli nie ma takiej możliwości, jako pamięć lokalna dla danej grupy roboczej zostanie zamapowany pewien obszar pamięci globalnej, co oczywiście nie pozwoli na uzyskanie takiego przyspieszenia jak prawdziwa pamięć lokalna.

Tabela 2 pokazuje jakie urządzenia i w jaki sposób mogą alokować poszczególne rodzaje pamięci.

Tabela 2. Dostęp do pamięci oraz możliwości jej alokacji w OpenCL (za [13])

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Globalna | Stała | Lokalna | Prywatna |
| Gospodarz | Dynamiczna alokacja  Odczyt i zapis | Dynamiczna alokacja  Odczyt i zapis | Dynamiczna alokacja  Brak dostępu | Brak alokacji  Brak dostępu |
| Kernel | Brak alokacji  Odczyt i zapis | Statyczna alokacja  Odczyt | Statyczna alokacja  Odczyt i zapis | Statyczna alokacja  Odczyt i zapis |

# IMPLEMENTACJA SOLWERA

## RÓWNOLEGŁA METODA GAUSSA

W obliczu wysnutych w rozdziale 2.2 wniosków można pokusić się o analizę potencjalnych możliwości zrównoleglenia metody eliminacji Gaussa.

### ELIMINACJA W PRZÓD

Kluczowe wnioski jeśli chodzi o zrównoleglenie fazy eliminacji w przód zostały wyciągnięte w rozdziale 2.2.1. Cała macierz musi zostać sprowadzona do postaci macierzy schodkowej, co oznacza iż dla każdego wiersza konieczne jest wykonanie pewnej ilości eliminacji wyrazów.

Aby eliminacja pierwszego niezerowego wyrazu w danym wierszu była efektywna i nie spowodowała niepożądanych efektów, eliminację należy przeprowadzić przy pomocy wiersza o identycznej charakterystyce, tj. tym samym pierwszym niezerowym wyrazie. Co więcej, eliminacje takie muszą trwać, dopóki dla żadnego wiersza w macierzy nie będzie takiego innego wiersza, który miałby ten sam pierwszy niezerowy wyraz (zgodnie z faktem iż w macierzy schodkowej pierwszy niezerowy wyraz w wierszu jest unikalną charakterystyką każdego wiersza).

Można zatem wyznaczyć pewną operację która musi zostać wykonana na wszystkich elementach pewnego zbioru danych – czyli operacja eliminacji w przód może zostać zrównoleglona w zgodzie z paradygmatem paralelizmu względem danych.

### ZAPEWNIENIE FORMY SCHODKOWEJ MACIERZY PO WYKONANIU ELIMINACJI

Powstaje wszakże inny problem – ponieważ OpenCL nie gwarantuje kolejności wykonania poszczególnych grup roboczych, może się zdarzyć iż eliminacje – prowadzące skądinąd macierz do formy, którą można sprowadzić do macierzy schodkowej przy pomocy operacji zamiany wierszy – nie dadzą macierzy schodkowej.

Optymalne dla rozwiązania było również uniknięcie całkowicie operacji zamiany wierszy, gdyż ma znacznie większy narzut pamięciowy i obliczeniowy niż operacje dodawania i mnożenia wiersza przez wyraz (jedyne dwie operacje konieczne do przeprowadzenia eliminacji w przód).

Stąd do rozwiązania został wprowadzony dodatkowy wektor, obok pamięci przechowującej macierz oraz wektor prawej strony. Wektor ten stanowi swoiste mapowanie rzeczywistej pozycji wiersza w macierzy do pozycji, którą zajmowałby w macierzy schodkowej (zgodnie z założeniem że wiersz o pierwszym niezerowym wyrazie będzie zajmował -tą pozycję w macierzy).

Mapowanie takie rozwiązuje jeszcze jeden problem, mianowicie szybkiego zweryfikowania czy dany wiersz posiada unikalny pierwszy niezerowy wyraz, a jeżeli nie – to względem jakiego wiersza należy go wyeliminować; informacja ta jest zapisana w tymże dodatkowym wektorze i oszczędza kosztownych przeszukań dużych obszarów pamięci.

# BADANIA WYDAJNOŚCI

Work in progress…

# WNIOSKI

Work in progress…

# BIBLIOGRAFIA

[1] Bathe, K. J.: Finite element procedures in engineering analysis. *Englewood Cliffs: Prentice-Hall*. 1982.

[2] Bernstein, A.J.: Analysis of Programs for Parallel Processing. *IEEE Transactions on Electronic Computers.* 1966 Vol. EC-15, no. 5, pp. 757-763

[3] Butrylo, B. [et al.]: A Survey of Parallel Solvers for the Finite Element Method in Computational Electromagnetics. *International Journal for Computation and Mathematics in Electrical and Electronic Engineering.* 2004 Vol. 23, no. 2, pp. 531-546.

[4] Cook, S.: CUDA Programming. A Developer’s Guide to Parallel Computing with GPUs. *Waltham: Elsevier.* 2013.

[5] Irons, B.: A Frontal Solution Program For Infinite Element Analysis. *International Journal for Numerical Methods in Engineering.* 1970 Vol. 2, pp. 5-32

[6] Introduction to OpenCL Programming – Training Guide. *AMD*. 2010

[7] Jamil, N.: A Comparison of Direct and Indirect Solvers for Linear Systems of Equations. *International Journal of Emerging Sciences.* 2012 Vol. 2, no. 2, pp. 310-321

[8] Milenin, A.: Podstawy MES. Zagadnienia termomechaniczne. *AGH*. 2010

[9] Rońda J., Oliver G.J., Introduction to numerical methods with Matlab procedures. *AGH*. 2010

[10] Tewarson, R.P.: Sparse matrices. *New York: Academic Press, Inc.* 1973

[11] ICT 1900 Series Central Processors 1904, 1905, *ICT Press release* (ICT), 1964 p. 4. [dostęp: 2013-11-02], Dostępny w Internecie: <<http://bitsavers.trailing-edge.com/pdf/ict_icl/1900/brochures/1904_Central_Processor_Sep64.pdf>>

[12] Conformant Products [online]. *Khronos Group*. [dostęp: 2013-12-11], Dostępny w Internecie: <<http://www.khronos.org/conformance/adopters/conformant-products#opencl>>

[13] The OpenCL Specification [online]. *Khronos Group.* [dostęp: 2013-12-15], Dostępny w Internecie: <<http://www.khronos.org/registry/cl/specs/opencl-1.0.48.pdf>>

1. Rysunek został zaczerpnięty z <http://icis.pcz.czest.pl/~roman/mat_dyd/prz_rown/mac_rzadkie/4_2.html> [dostęp 23-12-2013]. [↑](#footnote-ref-1)
2. Rysunek został zaczerpnięty z <http://icis.pcz.czest.pl/~roman/mat_dyd/prz_rown/mac_rzadkie/4_2.html> [dostęp 23-12-2013]. [↑](#footnote-ref-2)
3. Opis za <https://www.icm.edu.pl/kdm/Numeryka:_Macierzy#Compressed_Sparse_Row_Format> [↑](#footnote-ref-3)