AKADEMIA GÓRNICZO-HUTNICZA

IM. STANISŁAWA STASZICA W KRAKOWIE



Wydział Inżynierii Metali i Informatyki Przemysłowej



PROJEKT INŻYNIERSKI

pt.„Opracowanie modelu urządzenie-program dla wybranych solwerów układów równań liniowych wykonywanych na heterogenicznych architekturach sprzętowych”

Imię i nazwisko dyplomanta: **Mateusz Szewczyk**

Kierunek studiów: **Informatyka Stosowana**

Profil dyplomowania:  **Modelowanie i Technologie Informacyjne**

Nr albumu: **240197**

Opiekun: dr inż. Łukasz Rauch

Podpis dyplomanta: Podpis opiekuna:

Kraków 2013

***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 ………….… Podpis dyplomanta…………….

Spis treści

[1. Wstęp 4](#_Toc375657565)

[2. Heterogeniczne platformy obliczeniowe. 5](#_Toc375657566)

[2.1. Procesory graficzne GPU 5](#_Toc375657567)

[2.2. Porównanie wydajności oraz energochłonności procesorów CPU z GPU. 7](#_Toc375657568)

[2.3. OpenCL 10](#_Toc375657569)

[3. Przegląd solwerów układów równań liniowych 11](#_Toc375657570)

[3.1. Solwery iteracyjne 11](#_Toc375657571)

[3.2. Solwery bezpośredni 12](#_Toc375657572)

[4. Solwery wybrane do badań. 14](#_Toc375657573)

[4.1. ViennaCl 14](#_Toc375657574)

[4.2. Solwer Pawła Wala (jak opisać?) 14](#_Toc375657575)

[4.3. Porównanie wybranych solwerów. 14](#_Toc375657576)

[4.3.1. Metoda Gradientu Sprzężonego 14](#_Toc375657577)

[4.3.2. Metoda Gradientów BiSprzężonych 15](#_Toc375657578)

[4.3.3. GMRES 15](#_Toc375657579)

[4.3.4. Metoda LU 15](#_Toc375657580)

[5. Testy solwerów 16](#_Toc375657581)

[6. Opracowanie modelu urządzenie-program 17](#_Toc375657582)

[7. Podsumowanie 18](#_Toc375657583)

[8. Bibliografia 19](#_Toc375657584)

# Wstęp

# Heterogeniczne platformy obliczeniowe.

Heterogeniczne platformy obliczeniowe to systemy, które wykorzystują więcej niż jeden rodzaj procesora. Są to wielordzeniowe systemy, których wydajność wzrasta nie tylko poprzez dodawanie kolejnych rdzeni, lecz także poprzez wprowadzenie wyspecjalizowanego przetwarzania i obsługi poszczególnych zadań. Architektura heterogeniczna składa się z systemu zawierającego kilka rodzajów procesorów. Zazwyczaj jest to procesor (CPU) oraz procesor graficzny (GPU) znajdujące się najczęściej na tej samej matrycy w celu uzyskania najlepszej wydajności. Procesory graficzne poza przetwarzaniem i renderowaniem grafik 3D, mogą także wykonywać skomplikowane obliczenia matematyczne na dużych zbiorach danych, podczas gdy procesor obsługuje system operacyjny oraz wykonuje tradycyjne zadania szeregowe.

Do końca 2010 roku niemalże każdy komputer stacjonarny posiadał wielordzeniowe procesory. Przetwarzanie wielordzeniowe jest jednak problematyczne. Dodatkowe rdzenie oraz pamięć podręczna powodują wzrost rozmiarów oraz wymagają większej ilości energii.

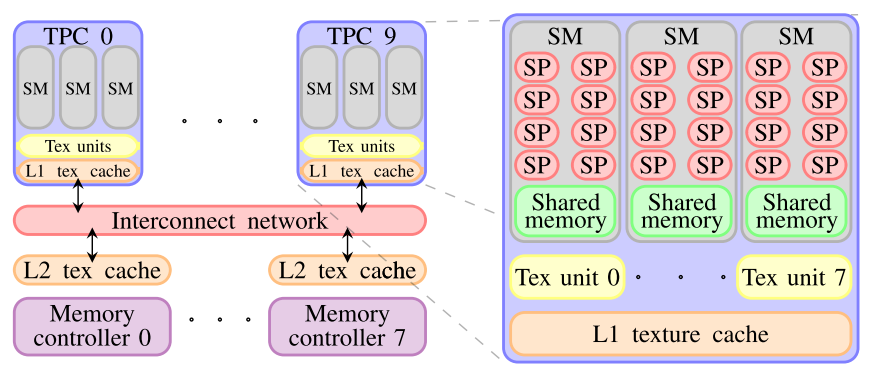
W tym samym czasie jakość i złożoność rozwiązań na procesorach graficznych znacznie wzrosła, co spowodowało, że zaczęto się nimi w większej mierze interesować. Dużą zaletą procesorów graficznych jest możliwość wykonywania przetworzeń wektorowych. Oznacza to że mogą one wykonywać równoległe operacje na bardzo dużych zbiorach danych, zarazem pobierając znacznie mniej energii niż podobny tradycyjny procesor. To jest powodem, dla którego procesory graficzne są w stanie wyświetlać tak realistyczne grafiki w wielu grach. Początkowo były wykorzystywane jedynie do odciążenia procesora oraz zwiększenia wydajności grafiki 3D. Teraz stały się one jeszcze bardziej atrakcyjne ze względu na wykorzystanie ich do bardziej ogólnych celów, takich jak równoległe obliczenia i zadania programistyczne.[[1]](#footnote-1)

## Procesory graficzne GPU

Procesory graficzne oryginalnie były projektowane do wykorzystania w grach komputerowych, gdzie różne obiekty geometryczne są renderowane. Każdy element wyjściowej grafiki składa się z pikseli, a GPU wykorzystuje swój procesor, aby wyliczyć równolegle każdy z kolorów pojedyńczych pixeli. W tej chwili teoretycznym szczytem możliwości dla procesora graficznego jest bliska trzem teraflopom [7]. Ograniczeniem w przypadku GPU jest jednak to, że GPU zazwyczaj zamontowane jest na szynie PCI express. Druga generacja PRI express pozwala na transfer danych z prędkością 8 Gb/s pomiędzy pamięcią CPU i GPU, gdzie podczas badań osiągalna prędkość wynosci 5.2 Gb/s.

GPU jest symetrycznym procesorem wielordzeniowym, który jest dostępny i sterowany wyłącznie poprzez procesor CPU, przez co stają się one systemem heterogenicznym. Procesor graficzny pracuje asymchronicznie z CPU, umożliwiając jednoczesne wykonanie i przeniesienie pamięci. W tej chwili głównymi producentami na rynku są AMD, Intel oraz AMD.

Przykładem architektury GPU wykorzystywanej w architekturach heterogenicznych może być NVIDIA GT200 programowana zazwyczaj przy pomocy CUDA, która eksponuje model programowania SPMD przy użyciu dużej liczby wątków zorganizowanych w bloki. Wszystkie bloki obsługują ten sam program, a wątki w jednym bloku można synchronizować. Mogą one także komunikować się przy pomocy pamięci współdzielonej. Komunikacja między samymi blokami jest jednak ograniczona do operacji atomowych wykonywanych w pamięci globalnej. Bloki są automatycznie dzialone na osnowy, składające się z 32 nitek. Bloki są zaprojektowane, tak aby prowadziły strumieniowanie wieloprocesorowe (SM) w czasie wykonywania, a każda osnowa jest wykonywana w modelu SIMD. Jest to realizowane poprzez wykonywanie tych samych instrukcji przez cztery kolejne działania na ośmiu procesorach skalarnych. Rozbieżność w przepływie kodu pomiędzy wątkami obsługiwana jest sprzętowo poprzez automatyczne maskowanie wątków oraz serializację, przez co różnice w osnowach zmniejszają wydajność, a pomiędzy poszczególnymi osnowami nie ma wpływu na szybkość.

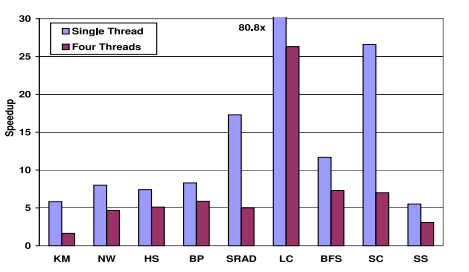


*Rys. 1 Architektura NVIDIA GT200 GPU. Skróty mają następujące znaczenie: TPC – klastr do przetwarzania tekstur; SM – multiprocesor strumieniowy; Tex unit – jednostka tekstury, Tex cache – pamięć podręczna tekstur; SP – procesor skalarny. Źródło [7]*

## 

## Porównanie wydajności oraz energochłonności procesorów CPU z GPU.

W 2009 roku zespół naukowców wykonał badania dotyczące wydajności procesorów graficznych [3]. Do badań użyto karty nVidia GeForce GTX 280 GPU z zegarem ustawionym na 1.3 Ghz oraz czterordzeniowego procesora Inter Core 2 Extreme CPU z zegarem ustawionym na 3.2 Ghz. Badania te wykazały, że w porównaniu z jednowątkową implementacją te same obliczenia wykonane na procesorze graficznym są szybsze od 5.5 do 80.8 razy, natomiast w przypadku czterowątkowej implementacji sięgają od 1.6 do 26.3 razy.

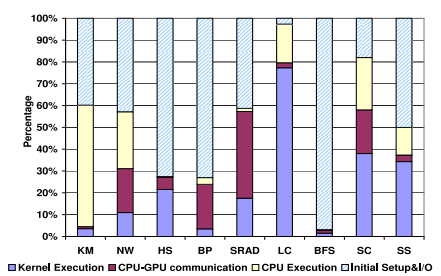


Rys. 2 Przyśpieszenie procesora graficznego w porównaniu z odpowiednikami obliczeń da jedno (Single Thread) i czterowątkowej (Four Threads) implementacji dla tradycyjnego procesora.

Powyższa grafika reprezentuje wyniki badań. Przeprowadzona została ona z wykorzystaniem różnych algorytmów. Skróty oznaczają kolejno:

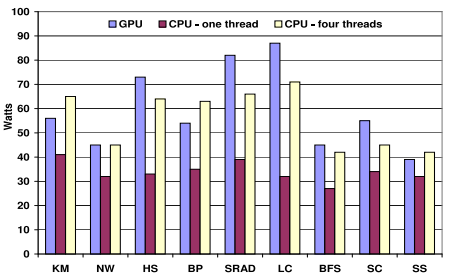
* KM – algorytm centroidów, który jest wykorzystywany w analizie skupień[[2]](#footnote-2).
* NW – algorytm Needlemana Wunscha umożliwiający znalezienie optymalnego globalnego dopasowania dwóch sekwencji[[3]](#footnote-3).
* HS – HotSpot czyli dokładny i szybki model termiczny znajdujący zastosowanie w architekturze[[4]](#footnote-4) .
* BP – propagacja wsteczny czyli jeden z podstawowych algorytmów uczenia nadzorowanego dla sieci nauronowych[[5]](#footnote-5).
* SRAD – algorytm oparty na dyfuzji równań różniczkowych cząstkowych [3]
* LC – wykrywa i śledzi białe krwinki naczyń krwionośnych w wideo mikroskopii wykorzystując GICOV [3]. [3]
* BFS – algorytm przeszukiwania wszerz.
* SC – Stream Cluster - algorytm rozwiązujący problem grupowania w internecie [3].
* SS - Similarity Score – algorytm stosowany w cel grupowania dokumentów WWW do obliczenia podobieństwa między nimi [3].

Badania wykazały wyższość rozwiązania wykorzystującego procesor graficzny. W każdym przypadku przyśpieszenie osiągnięte poprzez wykorzystanie GPU było znaczne. Powyższy wykres został opracowany nie biorąc pod uwagę operacji wejścia/wyjścia, a także początkowych ustawień, które ograniczają znacznie wydajność rozwiązań opartych na procesorach graficznych. Ilustruje to poniższy wykres.



Rys. 3 Procentowy podział czasu na kolejne zadania.

Teoretyczną granicę wydajności dla zrównoleglonych obliczeń wyznacza prawo Amdahla. W rzeczywistości granica ta jest niższa od teoretycznego maximum. Spowodowane jest to czasem wymaganym przez procesor do wykonania odpowiednich rozkazów, a także komunikacji pomiędzy CPU i GPU. Warto zwrócić uwagę, że w przypadku algorytmu LC gdzie komunikacja ta została ograniczona do minimum wydajność była największa.



Rys. 4 Energochłonność GPU, CPU dla jednowątkowego oraz czterowątkowego rozwiązania

Ze względu na dużą ilość dostępnych rozwiązań związanych z wysoką wydajnością obliczeń, ważnym zagadnieniem jest również energochłonnośc poszczególnych rozwiązań. W wyniku przeprowadzonych badań wynika, że jednordzeniowe obliczenia pobrały mniej mocy niż GPU w każdym przypadku. Porównując jednak GPU z czterowątkowymi obliczeniami proporcje te się zmieniają i CPU pobiera w niektórych przypadkach więcej mocy niż procesor graficzny. Ostateczne wyniki przeprowadzonych badań wykazały jednak, że biorąc pod uwagę stosunek wydajności do energożerności zawsze lepiej wypada GPU.

## OpenCL

OpenCL jest pierwszym w pełni darmowym językiem programowania, którego celem jest wykonywanie obliczeń zrównoleglonych na architekturach heterogenicznych, którego popularność szybko rośnie . OpenCL pozwala programistom skupić się na pisaniu aplikacji, a nie architekturze, za pośrednictwem jednej, przenośnej bazy kodu źródłowego. Podczas korzystania z OpenCL programista może korzystać z jednolitych narzędzi i języka, które będą działać na zrównoleglonych procesorach. API jest niskopoziomowe natomiast składnia przypomina w dużej mierze język C. [[6]](#footnote-6) Kod napisany w OpenCL może być uruchomiony na wielu platformach, nie wymagając przy tym zmian w kodzie.

# Przegląd solwerów układów równań liniowych

Solwery wykorzystywane do rozwiązania układów równań liniowych można podzielić na iteracyjne i bezpośrednie.

## Solwery iteracyjne

Solwery iteracyjne były używane w dużym zakresie we wczesnych latach 60’tych ubiegłego wieku, jednak z czasem zaprzestano wykorzystywanie ich do obliczeń. Było to spowodowane odkryciem, że liczba operacji, które były przeprowadzane przez metody iteracyjne przekraczała w znacznym stopniu limity wyznaczone w założeniach teoretycznych. Zastąpione zostały one o wiele bardziej efektywnymi solwerami bezpośrednimi, opartymi o trójkątny rozkład macierzy. Wzrastająca moc obliczeniowa oraz możliwość zrównoleglania obliczeń sprawiła jednak ponowne ożywienie w świecie metod iteracyjnych. Możliwości nowych architektur spowodowały, że solwery iteracyjne były wydajniejsze niż bezpośrednie metody dla bardzo dużych problemów tj. większych niż 100 000 elementów [1]. Kolejną zaletą było o wiele mniejsze zużycie pamięci, ponieważ wymaga on mniej więcej takiej samej ilości pamięci co oryginalne dane [5]. Jest to spowodowane tym, że współczynnik macierzy pozostawiany jest w oryginalnej formie i traktowany jako liniowy operator do wyliczania produktu wynikowego macierzy i wektora. Dodatkowo w przypadku solwera iteracyjnego użytkownik może wykorzystać zrównoleglanie obliczeń .

Iteracyjne metody rozwiązywania układów równań liniowych polegają na krokowym wyznaczaniu rozwiązania. Z każdą kolejną iteracją wynik jest poprawiany, aż do momentu gdy wynik zbliży się wystarczająco do założonego przybliżenia. W związku z tym końcowy wynik solwera iteracyjnego jest zawsze obarczony pewnego stopnia błędem. Teoretycznie używając solwera iteracyjnego można uzyskać idealny wynik przy nieskończonej ilości kroków, jednak w rzeczywistych obliczeniach ilość iteracji jest zawsze wartością skończoną. Złe warunki początkowe mogą jednak spowodować duży błąd w ostatecznych wynikach[[7]](#footnote-7). Kiedy solver iteracyjny rozwiązuje źle uwarunkowany problem wyniki mogą być wolne lub nie mieć zbieżności. Głównymi operacjami metod iteracyjnych są mnożenia macierzy i wektorów oraz rozwiązywanie liniowych równań algebraicznych dotyczących uwarunkowań wstępnych [4]. Do metod iteracyjnych rozwiązywania układów równań należąmiędzy innymi:

- Metoda gradientu sprzężonego - Jest ona jedną z najlepiej znanych metod iteracyjnych wykorzystywaną do rozwiązywania liniowych układów równań z rzadką, symetryczną, dodatnio określoną macierzą[[8]](#footnote-8). W każdym kolejnym kroku wykonywane jest mnożenie macierzy i wektora. Dla symetrycznej dodatnio określonej macierzy o rozmiarze N x N zbieżność można osiągnąć w N-iteracjach [5].

- Metoda gradientu Bi-Sprzężonego BiCG – rozwija metodę gradientów sprzężonych. Umożliwia ona rozwiązanie dla nie- pojedyńczej, niesymetrycznej macierzy.

- Ustabilizowana metoda gradientu Bi-Sprzężonego - służy do rozwiązywania liniowych

układów równań z niesymetrycznych równań liniowych. Jest stabilniejsza od Metody

Gradientu Bi-Sprzężonego.

- Uogólniona Metoda Najmniejszego Residuum GMRES – jest to dosyć popularna metoda do rozwiązywania nieosobliwych, niesymetrycznych macierzy. Badania dowiodły, że w metodzie GMRES stopa zbieżności może poprawiać się wraz z każdą kolejną iteracją [5][6].

## Solwery bezpośredni

Podział solwerów bezpośrednich:

- Bezpośrednie klasyczne

- Frontalne –podejście do rozwiązywania układów liniowych szeroko stosowane w analizie elementów skończonych.

Decydującym czynnikiem przy wyborze solwer bezpośredniego może być chęć użycia metody, która znacząco zmniejszy ilość niezerowych przejść podczas faktoryzacji macierzy. Dodatkowym atutem jest także możliwość wykorzystania pamięci dyskowej podczas dużych obliczeń na komputerach z małą ilością pamięci RAM i zachowanie przy tym wysokiej wydajności i prędkości podczas przetwarzania danych przechowywanych w RAM. Bezpośrednie metody pozwalają również na wykrywanie geometrycznej niestabilności modelu. Do tej pory najbardziej rozpowszechnionym solwerem bezpośrednim był solwer wielofrontalny. Posiada on wszystkie wymienione powyżej zalety, jednak posiada także nadmierną ilość transferów pamięć – pamięć oraz pamięć – dysk – pamięć. Na wielordzeniowym komputerze w architekturze SMP stanowi to problem uniemożliwiający osiągnięcie maksymalnych osiągnięć i przyśpieszenia z wykorzystaniem dodatkowych procesorów. Powszechnym problemem z bezpośrednimi solwerami jest kwadratowa zależność liczby operacji w wymiarze problemu, przekraczanie ilości dostępnej pamięci RAM oraz zbyt duży czas faktoryzacji macierzy [4]. Do solwerów bezpośrednich zalicza się:

- Metodę eliminacji Gaussa – bezpośredni solwer rozwiązujący układy równań pierwszego stopnia.

- Metoda eliminacji Gaussa- Jordana –

- Dekompozycja LU – metoda ta sprowadza się do stworzenia dwóch macierzy trójkątnych dolnej oraz górnej (z ang. Lower – dolna Upper – górna) oraz rozwiązania odpowiedniego równania.

# Solwery wybrane do badań.

## ViennaCl

Do przeprowadzonych badań użyto solwerów wchodzących w skład otwartej biblioteki ViennaCL. Jest to biblioteka umożliwiająca użytkownikowi wykonywać obliczenia naukowe w języku C++, zapewniając przy tym wsparcie dla CUDA, OpenCL oraz OpenMP. Zapewnia prosty dostęp do obliczeń wielowątkowych na architekturach takich jak procesory graficzne. ViennaCL wykonuje głównie obliczenia algebry liniowej, a także rozwiązuje równania za pomocą iteracyjnych i bezpośrednich metod[[9]](#footnote-9). Główne cechy biblioteki ViennaCL to:

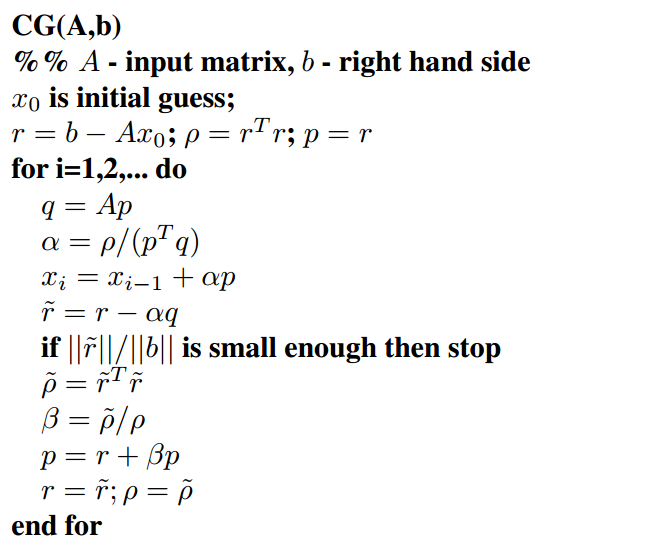
* Wygodne API napisane w języku C++ dla rzadkich i gęstych operacji algebry liniowej.
* Wsparcie dla BLAS czyli wysokiej jakości procedur służących do przeprowadzania podstawowych operacji na macierzach i wektorach[[10]](#footnote-10). ViennaCL za pomocą BLAS wspiera takie operacje jak: wektor\*wektor, macierz\*wektor oraz macierz\*macierz.
* Wsparcie dla obliczeń w technologii CUDA, OpenCL i OpenMP
* Trzy solwery iteracyjne: Gradientów Sprzężonych, Metodę Gradientów BI-Sprzężonych oraz GMRES.
* Solver bezpośredni: Metoda LU

## Solwer Pawła Wala (jak opisać?)

## Porównanie wybranych solwerów.

### Metoda Gradientu Sprzężonego

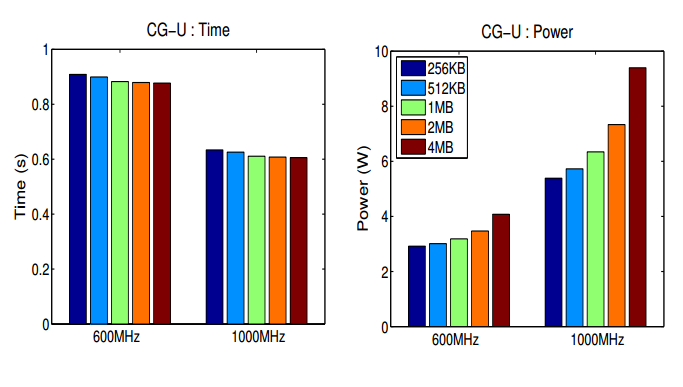
Jest to jedna z najpopularniejszych metod iteracyjnych ponieważ jest dobrze przystosowana do rozwiązywania układów macierzy rozszedzonych. W celu poprawy rozwiązań stosuje się różne techniki uwarunkowań wstępnych [8]. Algorytm metody gradientów sprzężonych dla równania **A\*x = b\*** przedstawiony został na rysunku 5.



Rys. 5 Schemat algorytmu dla metody gradientów sprzężonych. A – macierz wejściowa b –wektor wejściowy x0- przypuszczany wynik. Źródło [9]

Powyższy schemat używa standardowych struktur danych do przechowywania macierzy rzadkiej A oraz wektory p, q, r, które przechowywane są w jednowymiarowych tablicach w pamięci. Pojedyńcza iteracja wymaga jednego przemnożenia wektora z macierzą, dwóch wektorów, trzech operacji dodawania wektorów oraz dwóch dzieleń liczb zmiennoprzecinkowych. Pośród tych operacji najbardziej czasochłonne jest przemnożenie wektora i macierzy, któro zabiera 90% całego czasu operacji. Ponieważ macierz A jest macierzą rzadką, liczba zmiennych zmiennoprzecinkowych nie jest duża, co znacznie zmniejsza liczbę operacji na pamięci podczas przemnożenia. Dodatkowo pobieranie z pamięci elementów wektora p zależy od rzadkości macierzy A [9].

Poniższy rysunek przedstawia wykonane badania na maszynie z procesorem PowerPC 440 z częstotliwością ustawioną na 600 MHZ, a podczas drugiego badania 1000 MHZ. Pamięć wykorzystana w maszynie to DDR2. Badania sprawdziły czas rozwiązania macierzy o rozmiarze 35588 x 35588 oraz moc jaka została wykorzystana w trakcie badań dla róznych wielkości pamięci podręcznej.



Rys. 6 Czas obliczeń (po lewej) oraz moc wykorzystana w trakcie obliczeń (po prawej) dla metody gradientów sprzężonych. Źródło [9]

### Metoda Gradientów BiSprzężonych

### GMRES

### Metoda LU

# Testy solwerów

# Opracowanie modelu urządzenie-program

# Podsumowanie

# Bibliografia

1 verschoorJouglard C. E., Coutinho A.L.G.A, A comparison of iterative multi-level finite element Solvers, Computers & Structures, 69, 655-670, 1998.

2 Benqi G., Weiming C., An iterative and parallel solver based on domain decomposition for the h-p version of the finite element method, Journal of Computional and Applied Mathematics, 83, 71-85, 1997.

3 Shuai C., Boyer M., Jiayuan M., Tarjan D., Sheaffer J.W., Sang-Ha L., Skadron K., Rodinia: A Benchmark Suite for Heterogeneous Computing, Proc. of IEEE International Symposium on Workload Characterization, 44-54, 2009.

4 Fialko S. Y., [Iterative methods for solving large-scale problems of structural mechanics using multi-core computers](http://www.sciencedirect.com/science/article/pii/S1644966513000666), Archives of Civil and Mechanical Engineering, 14, 190-203, 2014.

5 Verschoor M, Jalba A. C.,Analysis and performance estimation of the Conjugate Gradient method on multiple GPUs, Parallel Computing, 38, 552-575, 2012.

6 Van der Vorst H.A., Vuik C., The superlinear convergence behaviour of GMRES, Journal of Computional and Applied Mathematics, 48, 327-341, 1993.

7 Brodtkorb A.R., Dyken C., Hagen T.R., Hjelmervik J.M, Storaasli O.O., State of the art in heterogeneous computing, Scientific Programming, 18, 1-33, 2010.

8 Butrylo B., Musy F., Nicolas L., Perrussel R., Scorretti R., Vollaire C., A Survey of Parallel Solvers for the Finite Element Method in Computional Electromagnetics, The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, 23, 531 – 546, 2004.

9 Malkowski K. Lee I, Raghavan P., Irwin M. J., Conjugate Gradient Sparse Solvers: Performance-Power Characteristics, IPDPS,

W opisie solwera - > opisac wydajnosc

Benchmarking -> Solwer – Sprzet – Dane

Wyniki -> Czas, Global i Local Worksize -> Na Nvidii zużycie energii

1. <http://developer.amd.com/resources/heterogeneous-computing/what-is-heterogeneous-computing/> [↑](#footnote-ref-1)
2. [http://pl.wikipedia.org/wiki/Algorytm\_centroidów](http://pl.wikipedia.org/wiki/Algorytm_centroid%C3%B3w) [↑](#footnote-ref-2)
3. <http://pl.wikipedia.org/wiki/Algorytm_Needlemana-Wunscha> [↑](#footnote-ref-3)
4. <http://lava.cs.virginia.edu/HotSpot/> [↑](#footnote-ref-4)
5. <http://pl.wikipedia.org/wiki/Propagacja_wsteczna> [↑](#footnote-ref-5)
6. <http://www.techdarting.com/2013/06/what-is-opencl.html> [↑](#footnote-ref-6)
7. <http://th-www.if.uj.edu.pl/zfs/gora/metnum12/wyklad05.pdf> [↑](#footnote-ref-7)
8. <http://www.icm.edu.pl/kdm/Metoda_gradient%C3%B3w_sprz%C4%99%C5%BConych_CG> [↑](#footnote-ref-8)
9. <http://viennacl.sourceforge.net/viennacl-about.html> [↑](#footnote-ref-9)
10. <http://pl.wikipedia.org/wiki/Basic_Linear_Algebra_Subprograms> [↑](#footnote-ref-10)