Prof. dr hab. inż. Jacek Kitowski
Katedra Informatyki
Akademia Górniczo-Hutnicza
Al. Mickiewicza 30, 30-059 Kraków
kito@agh.edu.pl

Kraków, 27 grudnia 2017

Recenzja rozprawy doktorskiej pana mgra inż. Filipa Krużela pt. „Odwzorowanie procedur całkowania numerycznego w metodzie elementów skończonych na architektury procesorów masowo wielordzeniowych"

Promotor: dr hab. inż. Krzysztof Banaś, prof. AGH

Niniejsza recenzja przygotowana została na zlecenie Sekretarza Rady Naukowej Instytutu Podstawowych Problemów Techniki PAN, Pana Dra hab. inż. Zbigniewa Ranachowskiego, prof. IPPT PAN (pismo z dnia 6 listopada 2017), działającego na podstawie uchwały Rady Naukowej Instytutu Podstawowych Problemów Techniki PAN z dnia 26 października 2017.

I. Ocena wyboru tematu i tez rozprawy

W ostatnim okresie rozwoju nauk obliczeniowych zwraca się szczególną uwagę na zagadnienia kosztu obliczeń, nieuchronnie związane ze wzrostem skali problemów obliczeniowych, wynikającej z rozwoju technologii. Wiadomo od dawna, że koszt obliczeń zależy w największym stopniu od częstotliwości pracy procesora oraz komunikacji wewnątrz maszyny. Celem minimalizacji tego kosztu architekci systemów obliczeniowych przyjmują szereg rozwiązań, z których głównymi wydają się być obniżenie częstotliwości pracy na rzecz zwiększenia liczby elementów przetwarzających oraz zmniejszenie rozmiaru warstwy komunikacyjnej łączącej elementy przetwarzające. W ten sposób powstały koncepcje procesorów masywnie wielordzeniowych, których efektywne wykorzystanie w obliczeniach wielkiej skali stanowi poważne wyzwanie dla twórców algorytmów dla potrzeb nauk obliczeniowych. Jedną z trudności jest liczba poziomów równoległości do wykorzystania we współczesnym procesorze, hierarchia pamięci oraz różnice koncepcyjne pomiędzy poszczególnymi architekturami procesorów masywnie wielordzeniowych. Stosunkowo często

jednostki wielordzeniowe realizują funkcje wspierające obliczenia wykonywane na procesorze głównym – pełniąc rolę koprocesora – co dodatkowo utrudnia projektowanie algorytmów.

W obliczeniach naukowych jedną z najczęściej stosowanych metod jest metoda elementów skończonych (MES), a badania nad efektywnością w tym zakresie sprowadzają się często do optymalizacji zagadnień algebry liniowej, w tym rozwiązania dużego układu równań liniowych. Jednak przy szerokich badaniach nad tym zagadnieniem istnieje możliwość wybrania na tyle efektywnej metody, że pozostałe kroki, w tym całkowanie numeryczne i tworzenie globalnej macierzy sztywności zaczynają odgrywać istotną rolę w efektywności całego algorytmu.

Recenzowana praca p. mgra inż. Filipa Krużela dotyczy zatem badania metod tworzenia globalnej macierzy sztywności przy założeniu wykorzystania systemów masowo wielordzeniowych. Problem badawczy sprowadzony został do zagadnienia kompleksowej analizy algorytmu całkowania numerycznego, w którym uwzględnia się szereg czynników mających wpływ na wydajność. Dzięki takiemu podejściu jej rezultaty mogą przyczynić się do tworzenia efektywnych algorytmów wykorzystujących masywną wielordzeniowość w innym zakresie problemowym. Do istotnego dorobku należy też opracowanie systemu automatycznego tuningu kodu do konkretnej architektury wielordzeniowej. Reprezentowane w pracy synergiczne podejście do problemu (uwzględniające zagadnienia algorytmiczne i implementacyjne wynikające z różnic w architekturze) korzystnie świadczy o przygotowaniu Autora do pracy naukowej.

Zatem, *kierunek badań, wybór problematyki rozprawy oraz jej celów zaproponowany i zrealizowany przez p. mgra inż. Filipa Krużela oceniam zdecydowanie pozytywnie.*

II. Ocena zawartości rozprawy

Przedstawiona do recenzji rozprawa zawiera pełną analizę badanego algorytmu oraz jego implementację na wybrane systemy o masywnej wielordzeniowości. Jest obszerna, liczy 145 stron zasadniczego tekstu, składając się z 6 rozdziałów, spisu 127 pozycji bibliograficznych (uszeregowanych alfabetycznie) oraz spisu rysunków i tabel. Pracę rozszerzają 3 dodatki (o łącznej długości 15 stron) oraz streszczenie w języku polskim i angielskim. Całość zredagowana jest starannie pod względem edytorskim, z dbałością o przejrzystość zapisu i szatę graficzną.

We wprowadzającym Rozdziale 1 postawiono w sposób ogólny problem wykorzystania współczesnych architektur masywnie wielordzeniowych w naukach obliczeniowych i na tym tle przedstawiono stan badań oraz motywację i cele pracy. Dokonano przeglądu zawartości rozprawy i zdefiniowano nowość naukową rozprawy. Rozdział ten jest przejrzysty i dobrze definiuje zakres rozprawy i jej cele, mimo złożoności całego problemu. Świadczy o rozeznaniu Autora we współczesnej problematyce obliczeń, w tym w zakresie metody elementów skończonych.

Rozdział 2 dotyczy badanych architektur, w tym procesorów ogólnego przeznaczenia (głownie o architekturze x86 pochodzącej z firmy Intel), procesorów graficznych (NVidia i ATI), nietypowego procesora IBM PowerXCell8i, koprocesora Xeon Phi (Knights Corner, pochodzącego z firmy Intel) i AMD APU. Przedstawiono cechy szczególne poszczególnych architektur oraz w podsumowaniu rozdziału zwrócono uwagę – i słusznie -- na problem przenośności. Do pewnych niedostatków dyskusji należy brak krytycznego spojrzenia na

trwałość wykorzystywanych rozwiązań sprzętowych oraz ciągłość rozwoju poszczególnych architektur.

Rozdział 3 w całości poświęcony jest zagadnieniom całkowania numerycznego po obszarze, przy wskazaniu mniejszego nakładu dla wyznaczenia całek brzegowych, przyjmując założenie, że są one obliczane z użyciem klasycznych metod MES i rdzeni CPU. Przyjęto, że elementy w obszarze obliczeniowym całkowane są jednokrotnie. Szczegółowo przedstawiono i przedyskutowano kilka wersji algorytmów oraz – ogólnie – wybrane zagadnienia testowe: zagadnienie Poissona oraz uogólniony problem konwekcji-dyfuzji-reakcji. Określono złożoności obliczeniową i pamięciową. Wprowadzono interesujący parametr, jakim jest intensywność arytmetyczna określająca relacje między obliczeniami a dostępem do pamięci, co w przypadku układów masywnie wielordzeniowych o hierarchicznym układzie pamięci okazuje się niezbędne w dalszych rozważaniach w pracy odnośnie ograniczeń badanych systemów. Rozdział ten ponadto świadczy o dużej biegłości Autora w teoretycznej reprezentacji problemu oraz zagadnienia algorytmicznych MES.

W Rozdziale 4 opisano szczegółowo model programowania i wykorzystywane kompilatory oraz pakiety programistyczne. Do realizacji programu wybrany został korzystnie język C (w wersji firmowej oraz gnu CC), dający możliwość łatwego programowania kart graficznych w modelu CUDA oraz OpenCL. Wśród pakietów wspomagających realizację kodu wynikowego wykorzystano typowe środowiska OpenMP oraz Intel Cilk Plus dla potrzeb wielowątkowości oraz niskopoziomowy zestaw instrukcji Intel Intristic. Wybrane narzędzia bardzo dobrze świadczą o rozeznaniu i doświadczeniu Autora we współczesnych narzędziach programistycznych, jednocześnie podkreślając biegłość programistyczną i przyszłe możliwości rozwoju algorytmu wraz z zachodzącymi zmianami technologicznymi w zakresie budowy procesorów. Na szczególną uwagę zasługuje szczegółowość dyskusji dotyczącej wykorzystania środowiska OpenCL z uwzględnieniem różnic pomiędzy wersjami oraz użycia narzędzi profilujących.

Rozdział 5 stanowi podstawowy, obszerny rozdział implementacyjny o dużej wartości poznawczej. Jest o podzielony na autonomiczne podrozdziały, z których każdy zawiera opis metodologii, modelu wykonania wraz z analizą wydajności i wnioskami, odnoszący się do badanych systemów obliczeniowych. Do badań zaadoptowano szeroki zestaw najbardziej nowoczesnych (w okresie tworzenia pracy) systemów masywnie wielordzeniowych przy obecności wielopoziomowej hierarchii pamięci. Zestaw używanych systemów został korzystnie dobrany, ze względu na wspólne ich cechy funkcjonalne, lecz też istotne różnice w realizacji szczegółów architektury. W niektórych przypadkach przedyskutowano różne wersje algorytmu (np. z dekompozycja domenową lub wektoryzacją). Pozwoliło to na wskazanie ważnych elementów, niezbędnych do uwzględnienia przy opracowaniu implementacji o możliwie największej efektywności. Część wyników jest interpretowana poprzez analizę kodu asemblera (np. w odniesieniu do PowerXCell8i) – tego typu podejście jest bardzo wartościowe dla niskopoziomowej analizy czynników wpływających na wydajność. W wielu wypadkach uzyskano bardzo korzystne wartości przyspieszenia.

W odniesieniu do architektury x86 badania zrealizowano dla dwóch wariantów architektury: procesora Sandy Bridge i procesora Haswell. Zostały one wykonane wyjątkowo starannie i obszernie, z wykorzystaniem benchmarków (z literatury), obliczeń parametrów teoretycznych procesorów, które wnikliwie przeanalizowano w odniesieniu do własnych algorytmów (Poissona oraz aproksymacji Galerkina), z wykorzystanie środowisk wcześniej omówionych w

Rozdziale 4 (wektoryzacja, Cilk, Intristic, Stride). Pokazano praktycznie różnice w wydajności, w zależności od przyjętych założeń eksperymentu, wskazano na czynniki ograniczające (dostęp do pamięci) oraz możliwość uzyskania wydajności zbliżonej do teoretycznej – choć warto podkreślić pracochłonność takiego podejścia od strony programistycznej.

W następnej kolejności badano realizację opracowanych algorytmów na procesorach graficznych Nvidia K20M oraz Radeon R9280X. Na podstawie wcześniej wprowadzonych charakterystyk intensywności arytmetycznej (Tabela 3.7) wykazano czynniki limitujące wydajność oraz przeprowadzono szereg eksperymentów wydajnościowych z różnymi kombinacjami realizacji algorytmu całkowania dla aproksymacji liniowej i opcji optymalizacyjnych, przy założeniu wykorzystania paradygmatu „jeden element-jeden wątek". Zastosowano metodę zapisu ciągłego do pamięci (tzw. coalesced), dobrze spisującą się w dawnych realizacjach obliczeń typu SIMD z wykorzystaniem procesorów macierzowych – jej wpływ na wydajność stanowi jeden z badanych elementów. Wyznaczono teoretyczne czasy wykonania algorytmów i oszacowano wydajność rzeczywistą, która w najlepszych przypadkach była rzędu 70% i więcej, choć zauważono także dość znaczne niekiedy rozbieżności. Ciekawy eksperyment dotyczył badania wydajności wersji OpenCL algorytmów realizowanych na architekturach x86 (uzyskano niewiele gorsze wyniki wydajnościowe) oraz porównań wydajności GPGPU z CPU.

Pozostałe podrozdziały dotyczą wyników uzyskanych dla koprocesora Knights Corner i układu AMD APU. W przypadku pierwszego z nich uzyskano wyniki potwierdzające obecne w literaturze opinie o małej efektywności praktycznej tego układu. Nawet wersja procesorowa Knights Landing o wydajności teoretycznej ponad 3TFlops nie znalazła uznania u użytkowników i program został zamknięty przez producenta. Zawarte w pracy wyniki szczegółowej analizy przyczyn potwierdzają tę decyzję. W przypadku drugiej z architektur wykazano eksperymentalnie znaczący udział komunikacji miedzy elementami CPU a GPU pomimo sprzętowego wsparcia HSA w układzie APU. Wykazano jednak, że zastosowanie HSA znacząco minimalizuje koszt komunikacji.

Rozdział 6 rozprawy stanowi podsumowanie osiągniętych wyników oraz wnioski dotyczące wykorzystania podobnych architektur w przyszłości. Podsumowanie zawiera także pewne informacje dotyczące dorobku naukowego Autora.

Podsumowując stwierdzam, że recenzowana rozprawa p. mgra inż. Filipa Krużela wnosi istotny wkład o charakterze zarówno teoretycznym, jak i praktycznym do zagadnień rozwoju algorytmów dla współczesnych i przyszłych systemów o masowej wielordzeniowości z ogromnym potencjałem utylitarnym. Rozprawa jest dobrze ulokowana w obszarze intensywnych badań prowadzonych na świecie nad wykorzystaniem współczesnych architektur komputerowych w badaniach naukowych, łącząc w sobie synergicznie zagadnienia algorytmiczne i implementacyjne.

III. Zasadnicze osiągnięcia Autora rozprawy

Rozprawa doktorska p. mgra inż. Filipa Krużela stanowi wartościową naukową pozycję, zawierając wiele interesujących elementów o charakterze teoretycznym i praktycznym. Jej niezaprzeczalną zaletą jest podejście synergiczne, gdyż tylko ono zapewnia możliwość uzyskania

wysokiej wydajności założonych aplikacji naukowych, zwłaszcza w odniesieniu do nowych technologii przetwarzania. Praca napisana jest na wysokim poziomie, tak merytorycznym jak i edycyjnym; na podkreślenie zasługuje bardzo dobry warsztat naukowy pozwalający na zaawansowany teoretyczny sposób przedstawienia problematyki i rozwiązań szczegółowych, świadczący o dobrym przygotowaniu Autora do prowadzenia badań naukowych.

Do najważniejszych osiągnięć Autora rozprawy zaliczyć należy następujące elementy:

- Wybór i dobre sfomułowanie problemu naukowego wraz z propozycją algorytmów oraz ich teoretycznym zbadaniem pod kątem złożoności obliczeniowej i pamięciowej.
- Wprowadzenie i wykorzystanie pojęcia „intensywność arytmetyczna" w celu dyskusji i wyboru własności proponowanych algorytmów.
- Wykorzystanie różnych środowisk i porównanie wyników, w tym dla OpenMP, Cilk, Intrinsincs, opracowanie metody profilowania aplikacji w OpenCL za pomocą Nvidia Visual Profiler.
- Zaplanowanie, przeprowadzenie i szeroka dyskusja wyników eksperymentów w odniesieniu do poszczególnych systemów wybranych do testowania.

Wymienione osiągnięcia są oryginalne i znaczące, wspierają zatem *ogólnie pozytywną ocenę pracy*. Wnioski z przeprowadzonych badań są aktualne i dobrze wpisują się w prace naukowe w zakresie metod, środków i narzędzi informatyki i nauk pokrewnych. Ponownie należy podkreślić wysoki poziom rozprawy, dobrze świadczący o dojrzałości Autora i jego przygotowaniu do pracy naukowej.

IV. Uwagi dyskusyjne i krytyczne

Oprócz niewątpliwych walorów, rozprawa zawiera pewną liczbę elementów, które mogą być przedmiotem dyskusji. Wydaje się celowe podanie następujących spostrzeżeń:

- Arbitralne przyjęcie środowiska OpenCL dla realizacji większości prac implementacyjnych z pominięciem innych środowisk (OpenACC, CUDA).
- Brak informacji na temat statystyki wyników (jaki jest średni błąd kwadratowy przy pomiarach?); uzasadnienie poprawności przedstawienia niektórych wartości w ns z dokładnoścla do 3 miejsc znaczących (np. Tabela 5.21).
- W dyskusji algorytmów całkowite pominięcie zagadnienia kosztu energetycznego poszczególnych rozwiązań sprzętowych oraz związanego z przesyłaniem danych.
- Czy istniałaby możliwość przeprowadzenia testów z wykorzystaniem modułów optymalizowanej biblioteki MKL? Jakich wartości wydajności można byłoby się spodziewać?

Należy wyraźnie zaznaczyć, że podane uwagi nie kwestionują słuszności przyjętych koncepcji ani też nie wpływają w sposób istotny na poznawcze i utylitarne wartości zrealizowanych badań. Ich uwzględnienie może okazać się korzystne w dalszej działalności naukowej dotyczącej zbliżonych zagadnień.

V. Wniosek końcowy

Wspomniane w recenzji uwagi dyskusyjne nie umniejszają zasług Autora ani nie kwestionują jego osiągnięć. W przedstawianej rozprawie postawiono ważny współcześnie problem tworzenia efektywnych algorytmów na współczesne masywnie wielordzeniowe procesory i zaproponowano metodykę jego rozwiązania w zależności od dostępnej architektury w sposób świadczący o biegłości teoretycznej i implementacyjnej Autora rozprawy i jego dojrzałości naukowej. Praca wnosi istotny wkład w rozwój obliczeń symulacyjnych dużej skali. Podstawowe cele i zadania pracy zostały zrealizowane. Tematyka dobrze wpisuje się we współczesny nurt badań w tym zakresie.

Stwierdzam zatem z przekonaniem, że opiniowana rozprawa Pana mgra inż. Filipa Krużela spełnia wszystkie wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej Ustawie o Tytule Naukowym i Stopniach Naukowych i stawiam wniosek o dopuszczenie jej Autora do dalszych etapów przewodu doktorskiego. Ponadto, biorąc pod uwagę bardzo wysoki poziom naukowy rozprawy, biegłość Autora w operowaniu zaawansowanym aparatem matematycznym, znaczącą wartość osiągniętych wyników oraz wszechstronność przeprowadzonych badań wnoszę o jej wyróżnienie.