dr hab. inż. Krzysztof Banaś
Katedra Informatyki Stosowanej i Modelowania
Wydział Inżynierii Metali i Informatyki Przemysłowej
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Kraków, 31.01.2018r.

Recenzja rozprawy doktorskiej

mgra inż. Pawła Jarzębskiego

"Zastosowanie algorytmów wielowątkowych i rozproszonych do zwiększenia efektywności Metody Elementów Skończonych"

Promotor: prof. dr hab. inż. Krzysztof Wiśniewski

Przygotowana na zlecenie Rady Naukowej Instytutu Podstawowych Problemów Techniki PAN na podstawie decyzji z dnia 30 listopada 2017 r.

Kontekst i tematyka rozprawy

Rozprawa doktorska mgra inż. Pawła Jarzębskiego mieści się w dyscyplinie Informatyka, w szczególności w takich jej obszarach jak: obliczenia naukowo-techniczne, nauki obliczeniowe, obliczenia równoległe, obliczenia wysokiej wydajności i analiza numeryczna. Dotyczy metody elementów skończonych (MES), która w warstwie teoretycznej należy do analizy numerycznej. W warstwie praktycznej MES stosowana jest w wielu dziedzinach nauki i techniki. Jednak w ostatnich latach szczególnego znaczenia nabiera efektywne wykorzystanie możliwości oferowanych przez współczesny sprzęt komputerowy, co związane jest z optymalnym odwzorowaniem algorytmów na sprzęt. Analizy tego zagadnienia należą do obszarów obliczeń równoległych oraz obliczeń wysokiej wydajności i do niech w szczególności należy przypisać omawianą rozprawę.

Rozprawa dotyczy istotnego z praktycznego punktu widzenia oraz trudnego teoretycznie zagadnienia zwiększania wydajności obliczeń w metodzie elementów skończonych. Praktyczna istotność tematyki związana jest z pojawiającym się w wielu dziedzinach zastosowań wymaganiem przeprowadzania coraz bardziej złożonych symulacji coraz większych problemów, co w przypadku niezadowalającej wydajności prowadzi do zbyt długich czasów obliczeń. Jedną z najważniejszych możliwości zwiększania wydajności symulacji MES na nowoczesnym sprzęcie obliczeniowym jest stosowanie takich technik jak: przetwarzanie wektorowe (SIMD), wielowątkowość oraz klasyczna równoległość dla maszyn z pamięcią rozproszoną.

Trudność analizowanego zagadnienia wynika po części z jego istotności – zagadnienia związane z wydajnością obliczeń MES, w szczególności z rozwiązywaniem wielkich, rzadkich układów równań liniowych były i są nadal intensywnie badane przez wielu naukowców i praktyków na całym świecie. W takiej sytuacji osiągnięcie postępu jest trudne i wymaga przyswojenia rozległej wiedzy uzyskanej we wcześniejszych badaniach.

W przypadku optymalizacji wydajności symulacji MES zakres powiązanej problematyki obejmuje zagadnienia:

- sformułowania MES dla modelowanego problemu, w tym wyboru dyskretyzacji przestrzennej i metod aproksymacji, co decyduje o właściwościach rozwiązywanego układu równań liniowych
- tworzenia algorytmów rozwiązywania układów równań liniowych i ich implementacji w

specjalnych programach (solwerach)
- realizacji podstawowych, z punktu widzenia wydajności, fragmentów kodu, obejmujące szczegółową analizę interakcji sprzętu i oprogramowania

Autor w rozprawie skupia się na badaniach drugiego z wymienionych aspektów, wybierając z góry zadane rozwiązania dla obu pozostałych, co stanowi jeden z elementów dyskusyjnych rozprawy.

Zawartość rozprawy

Praca liczy 178 stron, poczynając od kilku informacji wstępnych, poprzez 7 rozdziałów zasadniczych (w tym "Wstęp" i "Podsumowanie"), spis wykorzystanej literatury oraz 9 dodatków zawierających informacje uzupełniające.

"Wstęp", będący rozdziałem pierwszym przedstawia motywację do podjęcia badań oraz cel i zawartość rozprawy. W rozdziale drugim Autor przedstawia szereg pojęć związanych z obliczeniami równoległymi, w tym ujednoliconą prezentację podstawowych miar wydajności równoległej (przyspieszenia obliczeń i wydajności zrównoleglenia) posługując się pojęciem procentowego udziału części sekwencyjnej (*serial fraction*) w całości czasu obliczeń, jako funkcji liczby procesorów (rdzeni) i rozmiaru zadania. Poza tym omawia model "*Roofline*" oceny wydajności obliczeń na pojedynczym węźle obliczeniowym oraz podstawowe elementy programowania w standardach OpenMP i MPI, wraz z wybranymi technikami optymalizacji wydajności obliczeń.

Rozdział 3 omawia jedno z osiągnięć pracy, efektywne zrównoleglenie pętli po elementach przy wielowątkowym tworzeniu globalnej macierzy sztywności w znanym programie obliczeń MES – FEAP. Autor przedstawia kilka możliwych rozwiązań problemu wyścigu (*data* race), który pojawia się przy agregacji elementowych macierzy sztywności do macierzy globalnej. Ostatecznie wybrane rozwiązanie (realizacja wzajemnego wykluczania wątków poprzez stosowanie dyrektywy OpenMP ATOMIC) wykazuje w przeprowadzonych testach mały narzut obliczeń równoległych i prawie idealną skalowalność dla rozważanych liczb wątków.

W rozdziale czwartym Autor omawia zrównoleglenie dla maszyn z pamięcia wspólną algorytmu bezpośredniego rozwiązywania układów równań liniowych z wykorzystaniem faktoryzacji LU i Choleskiego. W pierwszej części porównuje, teoretycznie i eksperymentalnie, szereg istniejących implementacji, komercyjnych i publicznie dostępnych, analizując takie aspekty jak metoda realizacji algorytmu (superwęzły, wielofrontalność), grafowy model zadań jako sposób organizacji obliczeń, formaty przechowywania macierzy rzadkich, wybór elementu głównego w trakcie faktoryzacji, itp. Szczególna uwaga zwrócona jest na algorytmy przenumerowania węzłów (stopni swobody), mające decydujący wpływ na liczbę wyrazów niezerowych w ostatecznie otrzymywanej w wyniku faktoryzacji macierzy, a w konsekwencji na wymagania pamięciowe implementacji, liczbę operacji przy realizacji obliczeń i ostateczną wydajność programu.

Druga część rozdziału czwartego omawia wielowątkową realizacje obliczeń w procedurze HSL MA86, wybranej z porównywanych wcześniej implementacji, wraz z zaproponowanymi przez Autora modyfikacjami algorytmu. Najistotniejszą ze zrealizowanych modyfikacji (pozostałe mają raczej charakter zmian konfiguracji obliczeń) jest przejście na obliczenia pojedynczej precyzji (uważane standardowo za zbyt mało dokładne dla obliczeń naukowo-technicznych) połączone z dodaniem fazy iteracyjnego poprawiania rozwiązania, realizowanej w podwójnej precyzji. Modyfikacja ta, motywowana powszechną w ostatnich latach wyższą wydajnością procesorów w miarę zmniejszania precyzji zmiennych, okazuje się być najskuteczniejsza dla rozważanych przez Autora przykładów.

Ostatnim zasadniczym rozdziałem pracy jest rozdział piąty przedstawiający wkład Autora w rozwiązywanie układów równań liniowych na maszynach z pamięcią rozproszoną. Autor koncentruje się na metodach bezpośrednich: uzupełnienia Schura i FETI-DP. Dla pierwszej z metod wyprowadza i implementuje wzory na hierarchiczną faktoryzację dla 8 podobszarów obliczeniowych, uzyskując w zadaniach przykładowych przyspieszenia ok. 3,5-4, przy

zastosowaniu opracowanego przez siebie solwera z mieszaną precyzją obliczeń. Jeszcze lepsze przyspieszenia udaje się uzyskać metodą FETI-DP, która jednak stosuje algorytm iteracyjny do rozwiązania problemu interfejsu.

Krótki rozdział 6 zawiera opis dwóch przykładów obliczeniowych dla zadań o znaczeniu praktycznym, dla których porównane są wydajności algorytmów opisanych w rozprawie. Zadania związane są z modelowaniem wieloskalowym złożonych materiałów, takich jak pianki i kompozyty ceramiczne.

Podsumowanie kończy właściwą część pracy. Następuje po nim bogaty spis wykorzystanej literatury (119 pozycji) oraz 7 dodatków zawierających materiał uzupełniający. Najciekawsze z dodatków, z informatycznego punktu widzenia, wydają się być Dodatek G, dotyczący porównania zrównoleglenia hybrydowego (MPI-OpenMP) ze zrównolegleniem w czystym modelu MPI oraz dodatek H, omawiający wykorzystanie liczników sprzętowych, dokładniej biblioteki PAPI, udostępniającej wartości liczników.

Najważniejsze osiągnięcia rozprawy

Autor w pracy wymienia szereg nowatorskich osiągnięć, z których szczególnie istotne są:

- opracowanie zmodyfikowanego wielowątkowego solwera z mieszaną precyzją obliczeń, na bazie procedury HSL MA86, oraz
- opracowanie solwera dla modelu z pamięcią rozproszoną i przesyłaniem komunikatów, bazującego na dekompozycji obszaru i obliczaniu uzupełnienia Schura za pomocą techniki częściowej faktoryzacji, także z możliwym użyciem obliczeń mieszanej precyzji.

Praktyczne znaczenie ma efektywne zrównoleglenie całości obliczeń w programie FEAP, obejmujące zastosowanie opracowanych solwerów oraz uzyskanie dzięki odpowiednim technikom programistycznym prawie idealnej skalowalności tworzenia globalnej macierzy układu równań liniowych.

Elementy dyskusyjne, uwagi krytyczne

1. Autor analizując wydajność programów równoległych opiera się na powszechnie przyjmowanym podstawowym założeniu, że wydajność programów zależy od wydajności w pojedynczym węźle oraz skalowalności. Analizując tę ostatnią wyprowadza szereg wzorów na przyspieszenie względne obliczeń oraz wydajność (efektywność) przyspieszenia.
    1. Na str. 17 Autor formułuje tezę, że przyspieszenie obliczeń nie może przekroczyć liczby procesorów, a tym samym wyklucza zjawisko przyspieszenia ponad-liniowego (*superlinear speed-up*). Dowód tej tezy opiera się na dyskusyjnym założeniu o możliwej równoważności wykonania równoległego programu oraz sekwencyjnego wykonania na jednym procesorze zadań poszczególnych procesorów (założenie takie nie uwzględnia np. komunikacji w wykonaniu równoległym). Założenie to wymaga komentarza, podobnie jak fakt obserwowanego w praktyce występowania zjawiska przyspieszenia ponadliniowego.
    2. Prawa Amdahla (str. 18) i Gustafsona-Barsisa (str. 19) wyprowadzane są jako konsekwencje tych samych wzorów ogólnych na przyspieszenie obliczeń, przy założeniu, że czas części sekwencyjnej jest zawsze stały. Prawo Amdahla powiązane jest z dodatkowym założeniem stałości rozmiaru zadania, natomiast przy wyprowadzaniu prawa Gustafsona-Barsisa brakuje jawnego wskazania dodatkowego wyróżniającego założenia (założeniem tym może być stałość współczynnika $\beta(n,p)$ przy stałym czasie $T(n,p)$, co byłoby zgodne z rys. 2.3 (str. 20), zakładając, że dla różnych krzywych na wykresie symbol β oznacza właśnie stałe $\beta(n,p)$. Jest to jednak sprzeczne z rys. 2.4 (str. 21), gdzie czas rozwiązania zadania nie jest stały – mimo podpisu stwierdzającego stałość czasu...). Wyjaśnienia wymaga zwłaszcza rys. 2.4 i jego powiązanie z założeniami wyprowadzenia prawa Gustafsona-Barsisa.

3. Podstawą wzoru 2.16 (str. 20), definiującego efektywność w sensie słabym (zwaną także efektywnością przeskalowaną, *scaled efficiency*) jest niezamieszczona jawnie w pracy relacja: $T(p*n,1)=p*T(n,1)$, gdzie n jest wcześniej określane jako wielkość (rozmiar) zadania. Wzór ten jest słuszny tylko dla specyficznych definicji rozmiaru zadania. W pracy brak jest precyzyjnej definicji rozmiaru zadania, brak też dyskusji wpływu definicji rozmiaru zadania na rozmaite miary wydajności i sposoby wyprowadzenia związanych z nimi wzorów (w tym wzorów związanych z prawami Amdahla i Gustafsona-Barsisa).
4. Brak precyzyjnej definicji rozmiaru zadania powoduje, że dla różnych algorytmów w pracy za rozmiar najczęściej przyjmowana jest domyślnie liczba stopni swobody (tak wynika np. z sekwencji wartości N dla badania słabej skalowalności w teście kostki (str. 53), gdzie rozmiar zgodnie z analizami w pracy powinien rosnąć proporcjonalnie do liczby procesorów). Przyjęcie takiej definicji rozmiaru jest uzasadnione tylko dla pewnych algorytmów (np. całkowania numerycznego i agregacji układu równań liniowych), natomiast nie spełnia założeń wymaganych dla analizy słabej skalowalności dla bezpośredniego rozwiązywania układów równań liniowych, które jest głównym tematem pracy. Dlatego analizy słabej efektywności w p. 6.1.3 (str. 134-135) oraz 6.2.2 (str. 138-139) wymagają uzupełnienia. Szczególnie ważna jest odpowiedź na pytanie jak zmienia się liczba operacji i czas obliczeń sekwencyjnych dla różnych algorytmów rozwiązywania układów równań liniowych w miarę zwiększania się liczby niewiadomych (stopni swobody).

2. Jako podstawowy test wydajności dla obliczeń na pojedynczym węźle Autor przyjmuje model *roofline*.
   1. Autor ogranicza się tylko do porównania wydajności analizowanych programów z testami wzorcowymi (benchmarkami) dla wydajności przetwarzania (LINPACK) i przepustowości pamięci (STREAM). Ciekawe byłoby także uwzględnienie teoretycznych maksymalnych możliwości wykorzystywanych węzłów obliczeniowych.
   2. Do ustalenia wydajności przetwarzania i przepustowości pamięci (zwanej w pracy transferem danych, choć ściśle oznacza szybkość transferu) Autor korzysta z liczników sprzętowych. W pracy brakuje szerszej dyskusji wykorzystania liczników sprzętowych, które oferują różne możliwości analizy wykonania programów przez procesory, a także znane są z rozmaitych niedokładności funkcjonowania i innych trudności związanych ze złożonym charakterem współczesnych procesorów.

3. W zadaniu zrównoleglenia pętli po elementach przy tworzeniu globalnej macierzy sztywności w programie FEAP kluczowe znaczenie ma efektywność różnych metod synchronizacji pracy wątków.
   1. W rozprawie Autor porównał kilka metod synchronizacji i wybrał jako optymalną technikę dającą najlepsze wyniki praktyczne. Brakuje w rozprawie dyskusji niskopoziomowych mechanizmów związanych z technikami synchronizacji, co uczyniłoby analizy i wnioski zawarte w rozprawie bardziej ogólnymi i głębiej uzasadnionymi.
   2. Autor nie przedstawia technik unikania wyścigu przy agregacji macierzy lokalnych bez wzajemnego wykluczania wątków, takich jak np. kolorowanie. Porównanie tych technik z zaprezentowanymi przez Autora wzbogaciłoby rozprawę i nadałoby rozważaniom większe znaczenie praktyczne.

4. W rozprawie Autor porównuje wydajność szeregu algorytmów równoległego rozwiązywania układów równań liniowych. Koncentruje się na technikach, w których główną role odgrywają różne typy faktoryzacji, a więc technikach bezpośrednich. Jednak w niektórych algorytmach pojawiają się elementy związane z technikami iteracyjnymi (poprawianie precyzji obliczeń, rozwiązanie problemu interfejsu w metodzie FETI-DP).

1. Skuteczność stosowania technik iteracyjnych zależy od uwarunkowania globalnej macierzy układu równań. Autor analizuje to zagadnienie wyłącznie jakościowo (np. na str. 93), brakuje analizy ilościowej.
2. Podobnie w przykładach obliczeniowych brakuje informacji o uwarunkowaniu występujących macierzy oraz o wpływie tego parametru na efektywność różnych solwerów. Zamieszczenie takiej dyskusji pozwoliłoby na bardziej ogólne spojrzenie na analizowane algorytmy oraz na próby ekstrapolacji zachowania solwerów w przypadku innych zagadnień obliczeniowych.

5. Rozprawa skupia się głównie na aspektach algorytmicznych oraz wybranych zagadnieniach odwzorowania na architekturę sprzętu.
   1. W tym ostatnim aspekcie brakuje analizy wpływu wektoryzacji na wydajność rozważanych programów (między innymi brak informacji o tym jaki był udział operacji wektorowych w całkowitej liczbie operacji zmiennoprzecinkowych w trakcie przeprowadzanych testów obliczeniowych). Postępowanie Autora polegające na korzystaniu ze zoptymalizowanych procedur BLAS jest słuszne i prowadzi do optymalnych rozwiązań. W rozprawie brakuje jednak dyskusji jakie cechy algorytmów umożliwiają takie działanie, co nastąpiłoby, gdyby algorytmy nie wykorzystywały wektoryzacji, jakie w związku z tym modyfikacje algorytmów są dopuszczalne, a jakie nie.
   2. Wykresy *roofline* dla paralelizacji pętli po elementach w programie FEAP, wskazują, że wydajność jest kilkukrotnie niższa od maksymalnej osiąganej w teście LINPACK. Brakuje w rozprawie analizy przyczyn takiego stanu.
6. Autor analizuje wydajność programów rozwiązywania układów równań linowych (solwerów) wykorzystujących faktoryzację, w szczególności LU i Choleskiego. W wielu miejscach rozprawy brakuje informacji, którą konkretnie faktoryzację badano i stosowano (np. rysunki 4.11–4.16 ilustrujące wpływ przenumerowania stopni swobody na wydajność faktoryzacji pokazują strukturę macierzy po dekompozycji LU, natomiast podstawowy stosowany w pracy algorytm MA86 w ogóle nie posiada opcji dekompozycji LU).

Podsumowanie - ocena rozprawy

Zaprezentowane powyżej uwagi krytyczne nie umniejszają pozytywnej oceny całościowej recenzowanej rozprawy. Składa się ona z szeregu analiz wybranych aspektów zrównoleglenia obliczeń metodą elementów skończonych, wraz z diagnozą występujących problemów oraz propozycjami ich rozwiązania, które następnie weryfikowane są w praktyce. Wartością rozprawy jest wysoki poziom merytoryczny rozważań, uwzględnienie złożonych aspektów problematyki oraz zaawansowanych osiągnięć innych badaczy. Biorąc to pod uwagę stwierdzam, że rozprawa mgra inż. Pawła Jarzębskiego "Zastosowanie algorytmów wielowątkowych i rozproszonych do zwiększenia efektywności metody elementów skończonych" przedstawia szereg samodzielnych i oryginalnych analiz oraz rozwiązań problemów związanych z optymalizacją wydajnościową obliczeń MES oraz świadczy o dogłębnym opanowaniu przez Autora szeregu tematów z dyscypliny Informatyka. Rozprawa spełnia w mojej ocenie wymagania stawiane przez „Ustawę o stopniach i tytule naukowym oraz o stopniach i tytule w zakresie sztuki" z dnia 14 marca 2003 r., wnioskuję o dopuszczenie Autora do dalszego toku przewodu doktorskiego.