

Przetwarzanie równoległe PROJEKT 1 OMP studia dzienne kierunek Informatyka, rok ak. 2020/ 2021

Ogłoszony na 3 spotkaniu laboratorium Przetwarzania równoległego od 31.03.2021, w ramach nauczania zdalnego podczas pandemii COVID-19.

Przygotowanie projektu w grupach

Studenci wykonują projekt w grupach 2 osobowych (można również przygotować projekt indywidualnie). Grupy projektowe zastaną ustalone podczas zajęć. Warto aby powstające grupy miały dostęp do komputera wielordzeniowego z procesorem Intel (system Windows lub Linux) w trybie administratora.

Zakres projektu:

1. Zapoznanie się z technologią oceny efektywności kodu równoległego dostępną w ramach oprogramowania Intel® VTune™ Profiler – opis w plikach pomocniczych do projektu VTUNE_TEORIA.pdf VTUNE_PRAKTYCZNIE.pdf.
2. Analiza efektywności przetwarzania równoległego realizowanego w komputerze równoległym z procesorem wielordzeniowym z pamięcią współdzieloną. Przykładem takiego systemu jest komputer z procesorem 6 rdzeniowym umożliwiający jednocześnie uruchomienie 12 wątków dostępny w laboratorium 2.7.6 lub **inny komputer** pozwalający jednocześnie przetwarzanie co najmniej 2 wątków (ciekawsze efekty eksperymentów można zaobserwować przy użyciu procesorów z większą liczbą rdzeni). Ze względu na polecane do przygotowania projektu oprogramowanie Intel VTune preferowane są komputery z procesorami Intela.
3. Opracowanie kodów i przygotowanie opracowania projektu bazującego analizie problemu, przygotowanych kodach i na uzyskanych wynikach interpretowanych zgodnie i zasadami oceny jakości przetwarzania proponowanymi w Intel® VTune™ Profiler

Problem do analizy:

Problem znajdowania liczb pierwszych w podanym za pomocą parametrów problemu przedziału $<M..N>$. W odpowiedzi na parametry $<10,49>$ powinny zostać wykonane niezbędne obliczenia potrzebne do wyznaczenia liczb 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47.

Algorytmy do implementacji

Należy przygotować różne wersje algorytmów: sekwencyjne i równoległe.

Wersje podejść - koncepcyjne:

1. Testowanie każdej liczby z wskazanego do badań przedziału i sprawdzanie czy jest podzielna przez liczby mniejsze. Zakres liczb mniejszych używanych do sprawdzenia można ograniczyć do pierwszej w kolejności liczby, która jest dzielnikiem badanej liczby i liczb mniejszych od pierwiastka kwadratowego z badanej liczby np. aby sprawdzić liczbę 49 jest ona dzielona przez 2,3,5,7, a liczba 50 będzie dzielona tylko przez 2. Jako potencjalne dzielniki można wykorzystywać liczby pierwsze (jeśli są dostępne lub jeśli ich wyznaczenie i dostęp do nich nie spowodują wzrostu kosztu obliczeń - potrzebny test efektywności).
2. Usuwanie ze zbioru liczb do zbadania (przedział $<M..N>$ liczb będących liczbami złożonymi (metoda Sita, które służy do odsiania nieinteresujących nas liczb złożonych – pozostają liczby pierwsze). Aby wyznaczyć i usunąć ze zbioru liczby złożone wyznaczane są wielokrotności liczb. Analogicznie do podejścia 1 interesują nas wielokrotności liczb będących liczbami w zakresie od 2 do pierwiastka kwadratowego z wartości górnego zakresu badanych liczb. Aby zmniejszyć ilość pracy – liczbę wykreśleń, do usunięcia liczb złożonych wystarczy kolejno używać wielokrotności, które są niezbędne: wielokrotności liczby 2 (2,4,6,8...) , wielokrotności liczby 3 począwszy od $3*3, 3*5, 3*7, 3*9, 3*11...$ Wystarczy używać wielokrotności liczb pierwszych, jeśli są dostępne lub okaże się że ich wyznaczenie jest rozwiązaniem efektywnym. Może się zatem okazać, że aby wyznaczyć liczby pierwsze z przedziału $<M..N>$ potrzebne są liczby pierwsze z przedziału $<2,..., \text{pierwiastek}(N)>$ używane do usuwania ich wielokrotności.

Wersje podejść – warianty zrównoleglenia:

1. Uwzględnienie dwóch podejść do podziału pracy pomiędzy wątki: podejście podziału domenowe i funkcyjne. Różne podejścia podziału pracy (tworzenia zadań równoległych) dotyczą drugiego podejścia koncepcyjnego opisanego powyżej:
 - Zadaniem realizowanym równoległe przez jeden wątek w **pierwszym trybie podziału pracy** może być wykreślenie wszystkich wymaganych wielokrotności podzbioru liczb pierwszych.
 - Zadaniem realizowanym równoległe przez jeden wątek może być wykreślenie **wielokrotności mających ograniczony zakres wartości**, ale są to wielokrotności wszystkich wymaganych liczb obsługiwane przez jeden/każdy wątek.
2. Równoważenie pracy procesorów poprzez odpowiednie ustalenie liczby wątków OMP i **odpowiedni sposób przydziału pracy** dla wątków. Można zastosować wskazanie pracy dla dowolnego wątku (dyrektywy single, sections, section) i dyrektywę podziału pracy w pętli for w różnorodnych wariantach (statyczny blokowy, warianty statyczne cykliczne, statyczne dynamiczne) oraz inne efektywne rozwiązania. Równoważenie obciążenia powinno wziąć pod uwagę możliwość określenia i możliwy rozrzut kosztów obliczeń dla zadania wykonywanego równoległe. Pomocne mogą tutaj się okazać eksperymenty obliczeniowe określające czas upływający podczas przetwarzania oraz stopień zrównoważenia pracą poszczególnych procesorów (Vtune).

Przygotowanie systemu obliczeniowego

W celu wykonania projektu należy użyć komputera z procesorem wielordzeniowym z zainstalowanym oprogramowaniem:

- Visual Studio (testowano oprogramowanie Visual Studio 2019 dostępne dla studentów Informatyki Politechniki Poznańskiej przez serwer oprogramowania Politechniki Poznańskiej) oraz oprogramowaniem
- Intel® oneAPI DPC++/C++ Compiler i Intel® VTune™ Profiler udostępnianymi bezpłatnie przez Intel jako części pakietu Intel® oneAPI Base Toolkit przez stronę:

<https://software.intel.com/content/www/us/en/develop/articles/free-intel-software-developer-tools.html#student>

Wersje kodu

• Wymagane jest przygotowanie kilku wersji kodu zgodnie z wymaganiami zadania. Wersje algorytmów rozwiązujących program powinny powstawać częściowo jako wersje wstępne do wykonania wstępnych eksperymentów dotyczących;

- a) poprawności uzyskiwanych wyników
- b) pomiaru czasu upływającego podczas przetwarzania
- c) zrównoważenia pracy użytych procesorów.

Zbiór kodów należy uzupełnić kolejnymi wersjami procedury obliczającej, będącymi próbami ulepszenia wyników uzyskanych we wcześniejszych eksperymentach. Proszę realizując eksperyment kontrolować zależność pomiędzy numerem eksperymentu, trybem realizacji eksperymentu a testowanym kodem, aby właściwie móc opracować wyniki przetwarzania.

Wymagane jest aby wszystkie kody źródłowe używane w eksperymentach zostały umieszczone w sprawozdaniu w części wstępnej, gdzie zostaną scharakteryzowane pod względem użytego podejścia do organizacji przetwarzania – opis użytych struktur danych, dostęp do danych - współdzielenie, podział pracy i dominujące operacje arytmetyczne.

Uruchomienia przetwarzania

Uruchomienia przetwarzania proszę wykonywać w optymalnej dla efektywności konfiguracji pozwalającej wykonać eksperyment oceny efektywności (Release, x86)

Badane instancje powinny zapewnić (wyznaczenie parametry MAX):

- przetwarzanie trwające odpowiednio długo dla zapewnienia poprawności metody zbierania informacji o przebiegu przetwarzania przez program oceny jakości, lecz nie dłużej niż około 1 minuta (dla dowolnej liczby procesorów i wątków)
- możliwość zapisu wyników obliczeń w jednej strukturze: `byte matrix[MAX-MIN]`;
- warianty bazujące na różnych podejściach koncepcyjnych ze względu na różne wydajności tych podejść powinny być testowane dla takich samych i różnych instancji pozwalających porównać ich możliwości.
- Wymagane instancje testowe:
 - a) 2...MAX
 - b) 2...MAX/2
 - c) MAX/2...MAX
- Liczba użytych procesorów:
 - a) Przetwarzanie sekwencyjne – jeden procesor,
 - b) Przetwarzanie równoległe dla maksymalnej liczby dostępnych w systemie procesorów logicznych,
 - c) Przetwarzanie równoległe dla maksymalnej liczby procesorów fizycznych - jeśli jest inna niż liczba procesorów logicznych.

Uwagi do postaci kodu:

Ze względu na minimalizację kosztów związanych z synchronizacją w OpenMP:

- Wprowadzamy w kodzie jedną dyrektywę tworzącą wątki `#pragma omp parallel` dla minimalizacji narzutu związanego z ponownym uruchomieniem zbioru wątków i wstrzymaniem-synchronizacją pracy wątków.
- Należy unikać dynamicznych struktur danych (np. bibliotecznych), których użycie może charakteryzować się elastycznością i prostotą kodu, lecz których obsługa może powodować wzrost kosztów przetwarzania
- Wszystkie niepotrzebne synchronizacje wbudowane po analizie poprawności przetwarzania (obecności wyścigu) usuwamy za pomocą klauzuli `nowait`
- Minimalizacja unieważnień linii pamięci podręcznej w ramach współdzielonych danych przez różne procesory poprzez minimalizację zbędnych zapisów będących przyczyną unieważnienia kopii linii pp.
Na potrzeby testów poprawności opcja kodu pozwala na wyświetlenie liczb pierwszych na konsoli w wierszach po 10 liczb **oraz** wyświetla liczbę liczb pierwszych w badanym zakresie: np. w zakresie od 50 do 100 znaleziono 10 liczb pierwszych.

Problemy do podjęcia w trakcie: przygotowania kodu (ich konsekwencje dla tworzonego kodu oraz przebiegu eksperymentu) i analizy wyników jakości przetwarzania:

- prywatność /współdzielenie danych i lokalność dostępu do danych: wynikająca z przebiegu algorytmu, przetwarzania równoległego (unieważnianie kopii linii pamięci podręcznej procesora), podziału pracy na wątki,
- zrównoważenie obciążeń procesorów: uwzględnienie czasu wykonania i rozrzutu wartości czasu wykonania zadania równoległego oraz wynikających z tego faktu uzasadnień dla statycznego (cyklicznego, blokowego) i dynamicznego podziału pracy, ocena jakości zrównoważenia na podstawie wyników pracy oprogramowania testującego.
- uwzględnienie rodzaju wykonywanej operacji podstawowej podczas obliczeń: wariant 1 koncepcji - dzielenie, wariant 2 koncepcji – dodawanie.
- określenie parametrów uruchomień i mierzonych podczas eksperymentu zdarzeń procesora, które zostaną użyte do oceny przyczyn różnic prędkości i przyspieszenia oraz wartości pozostałych udostępnianych przez oprogramowanie miar efektywności.

Sprawozdanie z projektu i eksperymentu obliczeniowego

Wstęp:

- Nazwa zaliczanego przedmiotu: laboratorium z przetwarzania równoległego
- Imiona i nazwiska autorów sprawozdania, numery indeksów, numer grupy dziekańskiej i termin zajęć laboratoryjnych.
- Terminy: wymagany i termin rzeczywisty oddania sprawozdania.
- Informacja czy sprawozdanie jest wersją pierwszą czy też wersją poprawianą – w przypadku poprawiania sprawozdania wymagane jest zaznaczenie w sposób czytelny (kolorem tekstu lub wyodrębnieniem akapitu) modyfikowanych fragmentów sprawozdania.
- Krótki opis treści realizowanego zadania
- Adres email kontaktowy do autorów sprawozdania.

Punkt 1:

- Opis wykorzystanego systemu obliczeniowego:
 - oznaczenie używanego procesora i jego parametry: liczba procesorów fizycznych, liczba procesorów logicznych, oznaczenie typu procesora, wielkość i organizacja pamięci podręcznych procesora, wielkość i organizacja bufora translacji adresów procesora (jeśli te informacje są dostępne)
 - wersja systemu operacyjnego, wersja i nazwa oprogramowania użytego do przygotowania kodu wynikowego i przeprowadzenia testów.

Punkt 2:

- Prezentacja przygotowanych wariantów kodów z wyjaśnieniem przewidywanego przebiegu przetwarzania. W sprawozdaniu powinny się znaleźć, z wyjaśnieniem zawartości, kluczowe fragmenty kodów charakterystyczne dla poszczególnych wersji kodu. Kolejne wersje kodu powinny zostać oznaczone w sposób skrótowy jednoznaczny i nazwą nawiązującą do wariantu użytego algorytmu i sposobu podziału pracy wątków omówionego powyżej. Oznaczenia zostaną użyte w dalszej części sprawozdania: w prezentacji wyników eksperymentów i wniosków związanych z poszczególnymi wariantami kodu. W opisie przebiegu przetwarzania (dla poszczególnych wariantów kodu) należy uwzględnić:
 - zagadnienie **podziału** pracy – wielkość zbioru zadań, który powstaje w przetwarzaniu równoległym,
 - sposób **przydziału** zadań do procesorów – kiedy i w jakiej liczbie zadania są przydzielane do procesorów
 - uzasadnić wybrany sposób podziału przetwarzania na zadania i sposób przydziału zadań w powiązaniu ze spodziewanym zrównoważeniem/brakiem zrównoważenia procesorów przetwarzaniem.
 - omówić dyrektywy i klauzule Open MP użyte w kodzie i ich znaczenie dla poprawności przebiegu obliczeń.
 - omówić występujące w kodzie potencjalne problemy poprawnościowe: wyścig (co to jest wyścig, dlaczego w kodzie występuje/ nie występuje i jaki jest jego efekt na wyniki i prędkość przetwarzania).
 - omówić występujące w kodzie potencjalne problemy efektywnościowe:
 - false sharing (co to jest false sharing i jak często (w stosunku do podstawowej operacji dodawania lub dzielenia) będzie występował podczas obliczeń).
 - synchronizacja – czy i gdzie w kodzie występuje i ja wpływa na czas obliczeń,
 - brak zrównoważenia procesorów obliczeniami.

Punkt 3:

Prezentacja wyników i omówienie przebiegu eksperymentu obliczeniowo-pomiarowego:

- a) omówienie jakie kodu (wymienione w sprawozdaniu) były testowane, jakich instancji przetwarzanie dotyczyło, dla jakich instancji (reprezentatywnych do porównań charakterystycznych wariantów przetwarzania) uruchomiono ocenę jakości przetwarzania i prezentowane są wyniki szczegółowe przetwarzania na poziomie miar jakości obliczonych na podstawie zdarzeń procesora.

- b) opisanie własnymi słowami sposób zbierania przez oprogramowanie Intel Vtune informacji o efektywności przetwarzania, określenie znaczenia zebranych i prezentowanych w sprawozdaniu informacji; informacja o wykorzystanych trybach pracy narzędzi oprogramowania Intel VTune.
- c) tabela z wynikami z czytelnym określeniem jakich uruchomień dotyczy: testowany wariant kodu i parametrami instancji; (warto aby tabele były jedna lub dwie, zwarte, dobrze wykorzystujące przestrzeń strony/stron opracowania, tabela zbiorcza zorientowana portrait - tak samo jak tekst (łatwa doczytania bez konieczności obrotu) i w sposób czytelny prezentowała kluczowe parametry, które są z wynikami eksperymentu oraz miary wyliczone:
- czas przetwarzania – czas upływający podczas przetwarzania od momentu rozpoczęcia obliczeń do gotowości tablicy zawierającej jednoznaczne informacje o zbiorze liczb pierwszych Tobl lub Elapsed time (wyznaczony w ramach eksperymentu Microarchitecture Exploration dla Intel Vtune Profiler)
 - liczba instrukcji kodu asemblera (instructions retired)
 - liczba cykli procesora w czasie wykonywania badanego kodu – (clockticks)
 - udział procentowy wykorzystanych zasobów procesora do przetwarzania kodu (retiring)
 - „ograniczenie wejścia” - udział procentowy w ograniczeniu efektywności przetwarzania części wejściowej procesora (front-end bound)
 - „ograniczenie wyjścia” - udział procentowy w ograniczeniu efektywności przetwarzania części wyjściowej procesora (back-end bound)
 - „ograniczenie systemu pamięci” - udział procentowy w ograniczeniu efektywności przetwarzania systemu pamięci (memory bound)
 - „ograniczenie jednostek wykonawczych” - udział procentowy w ograniczeniu efektywności przetwarzania jednostek wykonawczych procesora (core bound)
 - efektywne wykorzystanie rdzeni fizycznych procesora – „effective physical core utilization” – dodatkowe informacje jeśli występuje niezrównoważenie – na podstawie widoku Effective CPU Utilization Histogram lub widoku Bottom-up w ramach eksperymentu Microarchitecture Exploration dla Intel Vtune Profiler.
 - przyspieszenie przetwarzania równoległego dla badanego wariantu kodu równoległego – parametr ten jest ilorazem czasu przetwarzania najlepszego dostępnego przetwarzania sekwencyjnego (dowolnym algorytmem w tym samym systemie) oraz czasu przetwarzania równoległego, dla którego przyspieszenie jest wyznaczane.
 - prędkość przetwarzania liczona jako liczba przetestowanych liczb w jednostce czasu,
 - efektywność przetwarzania równoległego jako iloraz przyspieszenia przetwarzania równoległego i liczby użytych w przetwarzaniu procesorów fizycznych.
- d) wartości parametrów powinny być prezentowane w formacie pozwalającym na ich czytelne porównanie - wartości wraz z jednostkami prezentowanych wartości: przykładowo: czas przetwarzania 0,0053 s, prędkość obliczeń 1,23E+05 1/s (liczba zbadanych liczb na sekundę obliczeń).
- e) oprócz powyżej wymienionych miar efektywności obliczeń można przedstawić dodatkowe miary szczegółowe dostępne w wynikach przetwarzania, które w przypadku gdy wąskim gardłem ograniczającym wzrost prędkości, będą określały jaki element systemu pamięci powoduje opóźnienie przetwarzania: dostęp do poziomów pamięci podręcznej L1, L2, L3, dostęp do danych w pamięci operacyjnej, czy ograniczenia wynikające z dostępu do adresu wirtualnego w buforze translacji (DTLB lub STLB). Taka analiza choć nieobowiązkowa (umożliwia uzyskanie wyższej oceny), przy dobrym zrozumieniu miar i ich prezentacji znaczenia pozwala na głębsze spojrzenie w przyczyny braku pełnej wydajności przetwarzania.

Punkt 4

Wnioski. W wnioskach należy:

- porównać jakość rozwiązań problemu przy użyciu różnych zaproponowanych w badanych kodach podejść – w oparciu o prędkość przetwarzania lub czas obliczeń przy jednakowych instancjach problemu.
- podsumować czy użyte podejścia efektywnie wykorzystują struktury wewnętrzne procesora, jakie wąskie gardła pojawiają się w poszczególnych wariantach kodu i czy jest możliwość poprawy w tym zakresie lub znaleziono rozwiązanie, które pozwala na lepsze wykorzystanie jednostek wewnętrznych procesora, czy są warianty kodu, które w lepszy sposób niż pozostałe wykorzystują zasoby procesora i dostarczają przetwarzania przewyższającego efektywnością inne podejścia.
- podsumować użyte podejścia pod kątem jakości zrównoleglenia przetwarzania. W tej analizie należy wykorzystać przede wszystkim przyspieszenie, efektywność oraz dodatkowo „efektywne wykorzystanie rdzeni fizycznych procesora”; ostatni parametr jest tutaj miarą pomocniczą gdyż efektywne - zrównoważone wykorzystanie rdzeni fizycznych nie zawsze doprowadza do obliczeń o najkrótszym czasie.
- określić, które podejście okazało się najlepsze/najgorsze pod względem prędkości przetwarzania, czy ta przewaga jest niezależna od liczby użytych procesorów i rozmiaru testowanej instancji. Jaką przyczyną tych wyników.

- Jakie podejście równoległe jest najbardziej efektywne, czy efektywność skaluje się ze wzrostem liczby procesorów (jest stała, w jaki sposób się zmienia), jak wygląda przegląd efektywności użytych podejść. Jakie są przyczyny tych wyników.
- Jakie ograniczenia efektywnościowe dominują w poszczególnych metodach rozwiązania problemu.
- proszę używać do porównania miar względnych – np. przetwarzane trwało 2 razy krócej, a nie o 2 sekundy krócej, prędkość wzrosła o 100%, a nie o 10000 zbadanych liczb na sekundę i w sposób jednoznaczny wskazywać porównywane warianty uruchomień

W próbie uzasadnienia przyczyn wysokiej lub niskiej efektywności poszczególnych podejść do realizacji kodu proszę odwoływać się w sposób jednoznaczny do **poszczególnych wartości** miar w tablicy wyników (należy określać położenie porównywanych wartości w tablicy wyników eksperymentu) i poszczególnych **linii kodu** omawianej wersji kodu (znaczenia występujących w tych liniach konstrukcji językowych), rysunków zawierających graficzne efekty pracy oprogramowania Intel Vtune (zrzuty ekranu z pełnym wyjaśnieniem zawartości). Proszę unikać sformułowań ogólnych typu szybki, wolny, lepszy, gorszy oraz uwzględnić znaczenie wartości parametru: jeżeli stosunek trafień do pamięci ma wartość 0,5 lub 0,00001 to jego dwukrotny wzrost ma diametralnie różne konsekwencje na czas przetwarzania, podobnie różne znaczenie dwukrotny wzrost wartości przyspieszenia przetwarzania równoległego wynoszącego 5 lub 0,3.

Pod koniec sprawozdania proszę zamieścić dodatkową tabelę podsumowującą (zgodną zazwyczaj z zawartością wcześniejszej tabeli pełnej), w której znajdują się najlepsze wyniki prędkości przetwarzania dla poszczególnych wariantów metod: dzielenie, usuwanie wielokrotności domenowe, usuwanie wielokrotności funkcyjne z określeniem wielkości instancji, użytych: wariantów o kodu, liczby procesorów, liczby wątków i wartości ewentualnych dodatkowych parametrów metody wpływających kluczowo na wynik..

Przygotowanie sprawozdania.

Forma: sprawozdanie w formie elektronicznej wymagane jest w terminie 2 tygodni od 4 spotkania laboratoryjnego (28.04. – 13.05.2021 - terminy dla każdej grupy będą podane indywidualnie w ramach systemu **ekursy**) W sprawozdaniu zawierającym powyżej wymienione elementy składowe:

- powinny być ponumerowane strony,
- każdy obiekt sprawozdania (poza tekstem ciągłym): listing kodu, rysunek, wzór, tabela powinien posiadać numer kolejny i podpis określający w sposób jednoznaczny i pełny przedstawioną w obiekcie zawartość.

Podsumowanie i ocena realizacji projektu

Po sprawdzeniu merytorycznej zawartości zostanie wystawiona ocena, możliwe jest skierowanie sprawozdania nie spełniającego powyżej opisanych wymagań do uzupełnienia i/lub rozmowa z autorami na temat przebiegu eksperymentu i zawartości sprawozdania. Ewentualne uzupełnienia sprawozdania powinny zawierać zaznaczone dodane elementy (sprawozdanie należy uzupełnić w terminie 7 dni od otrzymania informacji).

Nieuzasadnione opóźnienie w oddaniu sprawozdania obniża ocenę w stosunku do oceny przyznanej za jakość sprawozdania. Obniżają ocenę o ½ stopnia następujące opóźnienia - po terminie, po tygodniu, po 2 tygodniach, po 3 tygodniach). Przekroczenie terminu oddania o 4 tygodnie powoduje wystawienie oceny ndst z pierwszego terminu zaliczenia przedmiotu i konieczność indywidualnego wystąpienia o wydanie nowego zadania zaliczeniowego.

Konsultacje:

Trudności praktyczne i koncepcyjne związane z realizacją projektu mogą Państwo zgłaszać przez email. Spotkania konsultacyjne online odbywają się 2 x w tygodniu. Możliwe są dodatkowe spotkania w miarę potrzeb.

Literatura:

Wykłady i literatura przedmiotu

Opracowania: Vtune_koncepcja, praktyczne_wprowadzenie_VtuneProfiler

Dokumentacja Intel Vtune Profiler

Opis przygotowany: 29.03.2021.