Tematy, zadania i pytania zaliczeniowe "Wydajność oprogramowania"

1. Zakładając DP = 2/4/8 drożną pamięć podręczną o rozmiarze RP = 16384/32768/65536 B:
   1. ile jest zbiorów linii, jeśli pojedyncza linia ma długość `DL =` 64/128 B ?
   2. w pętli:
      ```
      for(blok=0; blok < LICZBA_ITERACJI*SKOK; blok+= SKOK){
          suma += a[ blok*(DL/sizeof(double))];
      }
      ```
      (dla małych wartości parametru SKOK, będących potęgami 2, mniejszymi od 16)
      - co ile bloków o rozmiarze pojedynczej linii pamięci podręcznej następuje kolejny dostęp do tablicy `a` zmiennych double w pamięci DRAM?
      - co ile zbiorów linii następuje kolejny dostęp do kopii tablicy `a` w pamięci podręcznej?
      - co ile iteracji następuje trafienie w ten sam zbiór linii (dla bardzo dużych wartości parametru `LICZBA_ITERACJI`)?
      - dla konkretnej (bardzo dużej) wartości parametru `LICZBA_ITERACJI` ile razy odwiedzany jest pojedynczy zbiór linii w pamięci podręcznej (oczywiście spośród w ogóle odwiedzanych)? [można pominąć zaokrąglenia i założyć podzielność odpowiednich wartości]
      - dla jakiej wartości parametru `LICZBA_ITERACJI` wystąpi nagły skok chybień w pamięci podręcznej?
        [obliczona wartość jest iloczynem drożności i liczby odwiedzanych zbiorów, oznacza liczbę wykorzystanych linii pamięci cache, jej stosunek do liczby wszystkich linii daje wartość SKOK - SKOK jest dla dużych wartości parametru `LICZBA_ITERACJI` odwrotnością procentu wykorzystanej pamięci podręcznej]
   1. **LZ = (RP/(DL*DP))**
   2. SKOK, SKOK, SI=max(1,LZ/SKOK), `LICZBA_ITERACJI/`SI, `LICZBA_ITERACJI`>DP*SI

---

1. W trakcie sekwencyjnego (jednowątkowego) wykonania fragmentu kodu, w którym o czasie wykonania decydował tylko czas realizacji ..... operacji zmiennoprzecinkowych, profiler podał czas wykonania jako czas ..... taktów rzeczywistych i ..... taktów referencyjnych. Jaką wydajność w [flop/takt] uzyskał rdzeń mikroprocesora? Zakładając nominalną (zawartą np. w pliku */proc/*cpuinfo) częstotliwość pracy równą ..... GHz, oblicz jaka była wydajność rdzenia w Gflop/s przy wykonaniu kodu.

2. Skonstruuj diagram *roofline* dla platformy sprzętowej charakteryzującej się maksymalną wydajnością potoków przetwarzania ..... Gflop/s i przepustowością pamięci ..... GB/s. Wskaż na diagramie punkt oznaczany jako *ridge point*, związany z parametrem określanym jako *machine balance*. Czym jest ten parametr? Jaka jest jego wartość dla przeciętnych współczesnych mikroprocesorów (a także ich pojedynczych rdzeni)?

3. Skonstruuj diagram *roofline* dla platformy sprzętowej charakteryzującej się maksymalną wydajnością potoków przetwarzania ..... Gflop/s i przepustowością pamięci ..... GB/s. Zaznacz na diagramie linie i ewentualne punkty związane z wydajnością dwóch programów: jednego, którego wydajność jest ograniczana przez maksymalną wydajność pobierania danych z pamięci i drugiego, którego wydajność jest ograniczana przez maksymalną wydajność potoków przetwarzania.

4. Skonstruuj diagram *roofline* dla platformy sprzętowej charakteryzującej się maksymalną wydajnością potoków przetwarzania ..... Gflop/s i przepustowością pamięci ..... GB/s. Zaznacz na diagramie linie i ewentualne punkty związane z wydajnością dwóch programów: jednego, którego wydajność jest ograniczana przez maksymalną wydajność pobierania danych z pamięci i drugiego, którego wydajność jest ograniczana przez maksymalną wydajność potoków przetwarzania.

5. Program podczas wykonania realizuje ..... razy pętle odpowiadającą kodowi asemblera:

```
..B2.2:
        vfmadd231pd %ymm10, %ymm6, %ymm9
        vmovupd   %ymm9, 320(%rsp,%rax,8)
        addq      $4, %rax
        cmpq      $16, %rax
        jb        ..B2.2
```

Podaj ile pojedynczych operacji zmiennoprzecinkowych (dodawań i mnożeń) wykonuje rdzeń procesora realizując pętlę? Ile danych jest przesyłanych do procesora z pamięci (dowolnego poziomu)?

6. Program podczas wykonania realizuje ..... razy pętle odpowiadającą kodowi asemblera:

```
.L5:
        movsd   (%rdi,%rax,8), %xmm0
        mulsd   (%rsi,%rax,8), %xmm0
        addq    $1, %rax
        cmpl    %eax, %ecx
        addsd   %xmm0, %xmm1
        jg      .L5
```

Podaj ile pojedynczych operacji zmiennoprzecinkowych (dodawań i mnożeń) wykonuje rdzeń procesora realizując pętlę? Ile danych jest przesyłanych do procesora z pamięci (dowolnego poziomu)?

7. Program podczas wykonania realizuje ..... razy pętle odpowiadającą kodowi asemblera:

```
.L23:
        vmovupd (%rdx,%rax), %ymm1
        vbroadcastsd    (%rsi), %ymm0
        vfmadd132pd     %ymm2, %ymm1, %ymm0
        addq    $32, %rax
        cmpq    $864, %rax
        jne     .L23
```

Podaj ile pojedynczych operacji zmiennoprzecinkowych (dodawań i mnożeń) wykonuje rdzeń procesora realizując pętlę? Ile danych jest przesyłanych do procesora z pamięci (dowolnego poziomu)?

8. Program podczas wykonania realizuje ..... razy pętle odpowiadającą kodowi asemblera:

```
..B1.2:
        lea       128(%rsp), %rax
        movsd     (%rax), %xmm1
        incl      %eax
        addsd     %xmm2, %xmm0
        cmpl      $1000000000, %eax
        jl        ..B1.2
```

Podaj ile pojedynczych operacji zmiennoprzecinkowych (dodawań i mnożeń) wykonuje rdzeń procesora realizując pętlę? Ile danych jest przesyłanych do procesora z pamięci (dowolnego poziomu)?

---

9. Analiza kodu źródłowego, asemblera, zliczania zdarzeń sprzętowych i symulacji za pomocą programów profilujących doprowadziła do wniosku, że w trakcie wykonania programu:
    1. procesor wykonuje ..2x10^9.... skalarnych rozkazów zmiennoprzecinkowych fadd, fmul, fma na zmiennych 64-bitowych oraz

2. procesor wykonuje ...6x10^9.. rozkazów skalarnych pobrania danych podwójnej precyzji do rejestrów 64-bitowych oraz
3. występuje ....0.4x10^9.... chybień w pamięci L2 (zliczanych np. przez zdarzenie L2_LINES_IN.ALL) powodujących pobranie danych z pamięci L3 - co oznacza transfer ..... GB danych z pamięci L3

Zakładając, że wszystkie inne operacje poza wymienionymi powyżej są wykonywane w tle i nie wpływają na czas wykonania programu, oblicz jaki jest minimalny czas wykonania na platformie charakteryzującej się maksymalną wydajnością potoków przetwarzania ...100... Gflop/s i przepustowością pamięci: ....10... GB/s dla pamięci DRAM, ...40.. GB/s dla pamięci L3, ...80.. GB/s dla pamięci L2 i ...120... GB/s dla pamięci L1.

10. Analiza kodu źródłowego, asemblera, zliczania zdarzeń sprzętowych i symulacji za pomocą programów profilujących doprowadziła do wniosku, że w trakcie wykonania programu:
    1. procesor wykonuje ..6x10^9.. wektorowych rozkazów zmiennoprzecinkowych fadd, fmul, fma na wektorach 256-bitowych oraz
    2. procesor wykonuje ..2x10^9.. rozkazów wektorowych pobrania danych do rejestrów 256-bitowych
    3. występuje ...10^9... chybień w pamięci L1 (zliczanych np. przez zdarzenie L1D.REPLACEMENT) powodujących pobranie danych z pamięci L2 (np. dla pamięci podręcznych typu *inclusive*) - co oznacza transfer ..... GB danych z pamięci L2

    Zakładając, że wszystkie inne operacje poza wymienionymi powyżej są wykonywane w tle i nie wpływają na czas wykonania programu, oblicz jaki jest minimalny czas wykonania na platformie charakteryzującej się maksymalną wydajnością potoków przetwarzania ...100... Gflop/s i przepustowością pamięci: ....10... GB/s dla pamięci DRAM, ...40.. GB/s dla pamięci L3, ...80.. GB/s dla pamięci L2 i ...120... GB/s dla pamięci L1.

11. Analiza kodu źródłowego, asemblera, zliczania zdarzeń sprzętowych i symulacji za pomocą programów profilujących doprowadziła do wniosku, że w trakcie wykonania programu:
    1. procesor wykonuje ..6x10^9... wektorowych rozkazów zmiennoprzecinkowych fadd, fmul, fma na wektorach 256-bitowych oraz
    2. procesor wykonuje ...4x10^9.... rozkazów wektorowych pobrania danych do rejestrów 256-bitowych
    3. występuje ...2x10^9... chybień w pamięci L1 (zliczanych np. przez zdarzenie L1D.REPLACEMENT) powodujących pobranie danych z pamięci L3 (np. dla pamięci podręcznych typu ex*clusive*) - co oznacza transfer ..... GB danych z pamięci L3

    Zakładając, że wszystkie inne operacje poza wymienionymi powyżej są wykonywane w tle i nie wpływają na czas wykonania programu, oblicz jaki jest minimalny czas wykonania na platformie charakteryzującej się maksymalną wydajnością potoków przetwarzania ...100... Gflop/s i przepustowością pamięci: ....10... GB/s dla pamięci DRAM, ...40.. GB/s dla pamięci L3, ...80.. GB/s dla pamięci L2 i ...120... GB/s dla pamięci L1.

12. Analiza kodu źródłowego, asemblera, zliczania zdarzeń sprzętowych i symulacji za pomocą programów profilujących doprowadziła do wniosku, że w trakcie wykonania programu:
    1. procesor wykonuje ...8x10^9... skalarnych rozkazów zmiennoprzecinkowych fadd, fmul, fma na zmiennych 64-bitowych oraz
    2. procesor wykonuje ...2x10^9... rozkazów skalarnych pobrania danych podwójnej precyzji do rejestrów 64-bitowych
    3. występuje ...0.2x10^9... chybień w pamięci L3 (zliczanych np. przez zdarzenie MEM_LOAD_UOPS_L3_MISS_RETIRED.LOCAL_DRAM) powodujących pobranie danych z pamięci DRAM - co oznacza transfer ..... GB danych z pamięci DRAM

    Zakładając, że wszystkie inne operacje poza wymienionymi powyżej są wykonywane w tle i nie wpływają na czas wykonania programu, oblicz jaki jest minimalny czas wykonania na platformie charakteryzującej się maksymalną wydajnością potoków przetwarzania ...100... Gflop/s i przepustowością pamięci: ....10... GB/s dla pamięci DRAM, ...40.. GB/s dla pamięci L3, ...80.. GB/s dla pamięci L2 i ...120... GB/s dla pamięci L1.

---

13. Program podczas wykonania realizuje ..... razy pętle odpowiadającą kodowi asemblera:

a)
```
..B2.2:
        vfmadd231pd %ymm10, %ymm6, %ymm7
        vfmadd231pd %ymm10, %ymm6, %ymm8
        vmovupd   %ymm9, 320(%rsp,%rax,8)
        vmovupd   %ymm10, 448(%rsp,%rax,8)
        addq      $4, %rax
        cmpq      $16, %rax
        jb        ..B2.2
```

b)
```
.L4:
        movl    -4(%rbp), %eax
        imull   -44(%rbp), %eax
        addl    %edx, %eax
        leaq    0(,%rax,8), %rdx
        movq    -24(%rbp), %rax
        addq    %rdx, %rax
        movsd   (%rax), %xmm1
        mulsd   %xmm1, %xmm0
        addl    $1, -8(%rbp)
.L3:
        movl    -8(%rbp), %eax
        cmpl    -44(%rbp), %eax
        jl      .L4
```

c)
```
.L5:
        movsd   (%rdi,%rax,8), %xmm0
        mulsd   (%rsi,%rax,8), %xmm0
        addq    $1, %rax
        cmpl    %eax, %ecx
        addsd   %xmm0, %xmm1
        jg      .L5
```

d)
```
.L7:
        vmovsd  (%rcx,%rax), %xmm0
        vfmadd213sd     (%rdx,%rax), %xmm1, %xmm0
        vmovsd  %xmm0, (%rdx,%rax)
        addq    $8, %rax
        cmpq    $288, %rax
        jne     .L7
```

e)
```
.L23:
        vmovupd (%rdx,%rax), %ymm1
        vmovupd (%rcx,%rax), %ymm2
        vbroadcastsd    (%rsi), %ymm0
        vfmadd132pd     %ymm2, %ymm1, %ymm0
        vmovupd %ymm0, (%rdx,%rax)
        addq    $32, %rax
        cmpq    $864, %rax
        jne     .L23
```

f)
```
.L6:
        vmovsd  (%rdx,%rax), %xmm3
        vmovsd  (%rsi), %xmm0
        vfmadd132sd     (%rcx,%rax), %xmm3, %xmm0
        vmovsd  %xmm0, (%rdx,%rax)
        addq    $8, %rax
        cmpq    $864, %rax
```

```
        jne     .L6
```

g)

```
..B1.2:
        lea       128(%rsp), %rax
        movsd     (%rax), %xmm1
        mulsd     %xmm1, %xmm0
        incl      %eax
        addsd     %xmm2, %xmm0
        cmpl      $1000000000, %eax
        jl        ..B1.2
```

h)

```
        lub dowolnemu innemu ...
```

Podaj ile pojedynczych operacji arytmetycznych (dodawań i mnożeń) wykonuje rdzeń procesora realizując pętlę? Ile danych jest przesyłanych do procesora z pamięci (dowolnego poziomu)?

---

14. Analiza kodu źródłowego, asemblera, zliczania zdarzeń sprzętowych i symulacji za pomocą programów profilujących doprowadziła do wniosku, że w trakcie wykonania programu:
    - procesor wykonuje ..... skalarnych rozkazów zmiennoprzecinkowych fadd, fmul, fma na zmiennych 64-bitowych oraz
    - procesor wykonuje ..... wektorowych rozkazów zmiennoprzecinkowych fadd, fmul, fma na wektorach 256-bitowych oraz
    - procesor wykonuje ..... rozkazów skalarnych pobrania danych podwójnej precyzji do rejestrów 64-bitowych
    - procesor wykonuje ..... rozkazów wektorowych pobrania danych do rejestrów 256-bitowych
    - występuje ..... chybień w pamięci L1 (zliczanych np. przez zdarzenie L1D.REPLACEMENT) powodujących pobranie danych z pamięci L2 (np. dla pamięci podręcznych typu *inclusive*) - co oznacza transfer ..... GB danych z pamięci L2
    - występuje ..... chybień w pamięci L1 (zliczanych np. przez zdarzenie L1D.REPLACEMENT) powodujących pobranie danych z pamięci L3 (np. dla pamięci podręcznych typu ex*clusive*) - co oznacza transfer ..... GB danych z pamięci L3
    - występuje ..... chybień w pamięci L2 (zliczanych np. przez zdarzenie L2_LINES_IN.ALL) powodujących pobranie danych z pamięci L3 - co oznacza transfer ..... GB danych z pamięci L3
    - występuje ..... chybień w pamięci L3 (zliczanych np. przez zdarzenie MEM_LOAD_UOPS_L3_MISS_RETIRED.LOCAL_DRAM) powodujących pobranie danych z pamięci DRAM - co oznacza transfer ..... GB danych z pamięci DRAM

    Zakładając, że wszystkie inne operacje poza wymienionymi powyżej są wykonywane w tle i nie wpływają na czas wykonania programu, oblicz jaki jest minimalny czas wykonania na platformie charakteryzującej się maksymalną wydajnością potoków przetwarzania ..... Gflop/s i przepustowością pamięci: ..... GB/s dla pamięci DRAM, ... GB/s dla pamięci L3, ... GB/s dla pamięci L2 i ..... GB/s dla pamięci L1.

15. Przeprowadź analizę skalowalności w sensie słabym dla implementacji z przesyłaniem komunikatów algorytmu:
    - obliczania iloczynu skalarnego dwóch wektorów
    - obliczania iloczynu macierz-wektor
        - dla macierzy gęstych
        - dla macierzy pasmowych

16. Wyprowadź wzory na czas realizacji operacji komunikacji grupowej rozgłaszania (*broadcast*) / redukcji (*reduction*) / rozpraszania (*scatter*) / zbierania (*gather*) dla topologii pierścienia / torusa 2D / hiperkostki