Politecnico di Milano

Dipartimento di Elettronica, Informazione e Bioingegneria

prof. Fabrizio Ferrandi

Parallel Computing– 2nd mid-term test– Thursday, December 22nd 2022

Polimi ID _____ _____

Surname _____ _____ **Name** _____

- This is a closed-book examination. You cannot use computers, phones, or laptops during the exam.
- Paper will be provided, but you should bring and use writing instruments that yield marks dark enough to be read easily. Erasable pens can be used.
- Total available time: 1h:30m.

Exercise 1 (4 points) _____

Exercise 2 (4 points) _____

Exercise 3 (4 points) _____

Exercise 3 (4 points) _____

Exercise n. 1

Answer the following questions about parallel patterns and briefly explain (without an explanation, the answer will be considered invalid)

A. Please define what a parallel pattern scan is: assumptions, complexity, and example. (1)

B. Please describe how a parallel pattern zip works. (1)

C. Please describe how priority scatter works. (1)

D. Can you remove an anti-dependency? How? (1)

Exercise n. 2

Answer the following questions about parallel programming with the shared memory model (without an explanation, the answer will be considered invalid).

A. What are the characteristics of a program that make it amenable to parallelization with multiple threads? (1)

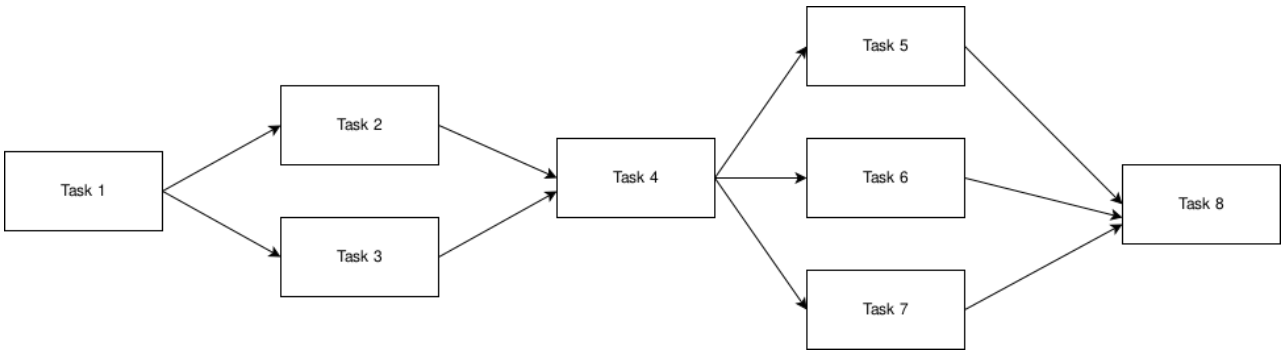
B. Why is synchronization necessary when programming with the shared memory model? (1)

C. What is the difference between joinable and detached threads in Pthreads? (1)

D. Which Pthreads API call is needed to ensure that spawned threads can finish their execution after main() returns? (1)

Exercise n. 3

Answer the following questions looking at the task graph (without an explanation, the answer will be considered invalid).



W(Task 1) = 50, W(Task 2) = 100, W(Task 3) = 100, W(Task 4) = 20, W(Task 5) = 10, W(Task 6) = 150, W(Task 7) = 175, W(Task 8) = 50

A. Calculate work and parallelism. (1)

B. Write an OpenMP implementation of the program. (2)

C. What is the improvement in performance with respect to a sequential implementation, assuming Task 4 and Task 5 represent the thread management overhead? (1)

Exercise n. 4

Answer the following questions about GPGPU computing with CUDA (without an explanation, the answer will be considered invalid).

A. Assume that a CUDA kernel is launched with 1000 thread blocks, each of which has 512 threads. If a variable is declared as a shared memory variable, how many versions of the variable will be created through the lifetime of the execution of the kernel? (1)

B. How does the impact of control divergence change for the following image processing kernel, executed with blocks of 16x16 threads and 32 threads per warp, when the input image size changes from 400x400 to 800x600? (3)

```
__global__ void PictureKernel(float* d_Pin, float* d_Pout, int n, int m) {
 // Calculate the row id of the d_Pin and d_Pout element to process
 int Row = blockIdx.y*blockDim.y + threadIdx.y;
 // Calculate the column id of the d_Pin and d_Pout element to process
 int Col = blockIdx.x*blockDim.x + threadIdx.x;
 // each thread computes one element of d_Pout if in range
 if ((Row < m) && (Col < n)) {
  d_Pout[Row*n+Col] = 2*d_Pin[Row*n+Col];
 }
}
```