# Assignment 5

TDT4200


## Stian Jensen


## Problem 1, CUDA:

*a) Briefly explain the difference in architecture between GPUs and CPUs.*

A GPU is more optimized for SIMD operations, ie. executing the same instruction on multiple data.

It is highly parallel and can execute many things at once.

*b) What is occupancy in CUDA?*

Occupancy is the number of currently active warps divided by the maximum number of warps

*c) Explain the difference between a block and a warp.*

A block can contain a number of threads which are able to communicate with each other.

A warp consists of 32 threads, which are all executing the same instruction.

*d) Explain the difference between local and shared memory.*

Local memory is local to the thread, while shared memory is shared between all threads in a block.

Local memory is considered slow, compared to shared memory.

## Problem 2:

CPU: $10 * n^2$
GPU: $n^2 + 2*n/r$

$10n^2 > n^2 + 2*n/r$
$10n^2 > n(n+2/r)$
$10n > n+2/r$
$9n > 2/r$
**$n > 2/9r$**

## Problem 3:

a)

The code might deadlock if a thread index's .x property is above 95. Then it won't call the __syncthreads() function. When not all threads call this function, threads that do call it will be waiting indefinitely.

b)

II is faster because all threads in the block are now executing the same code path.

## Problem 4:

II will be faster, because each thread works on a contiguous chunk of memory.