CSCI-GA.3033-004
Graphics Processing Units (GPUs): Architecture and Programming
Homework Assignment 1

(total: 25 points)

1. [11] To design the next generation GPU, a company has several choices to make:
    1. Increasing number of SM
    2. Increasing number of SPs (or cuda cores) per SM
    3. Increasing memory bandwidth
    4. Increasing shared memory per SM
    5. Increasing L2 cache size

Discuss the pros and cons for each one of the following scenarios (at least one positive and one negative issue for every scenario to get full credit):

a) Doing 1
b) Doing 2
c) Doing 3
d) Doing 4
e) Doing 5
f) Doing 1 and 2
g) Doing 2 and 3
h) Doing 1 and 4
i) Doing 2 and 4
j) Doing 4 and 5
k) Doing 3 and 5

Assume the GPU has L1 cache and shared memory in each SM and there is also L2 cache shared among all SMs.

2. [2 points] Let's assume an application has a lot of independent and similar operations to performed on data. Does **the amount of data** has anything to do with the expected performance of the GPU?

3. [8 points] For each of the following applications state whether it is beneficial to implement them on a GPU, and justify your answer.
    a) Finding whether a number exist in an array of 10M integers
    b) Calculating the first 1M Fibonacci numbers
    c) Multiplying two 100x100 matrices
    d) Multiplying 1Mx1M matrices

4. [4 points] We said in class that communication between system memory and CPU memory is very expensive in terms of latency, which negatively affects performance. Why not having one main memory for both the CPU and GPU?