5549: Assignment 1

1. Write cuda code for computing matrix matrix product of the form C=A*B where A is an m*k (double) matrix and B is k*n (double) matrix and C is an m*n (double) matrix. The code should be tiled and should shared memory. Make sure that the code works for matrices of all sizes not just square matrices. Follow the given template code.

- The name of the cuda kernel and the binary executable must be “matmul_double”.
- The executable must accept m,n and k (in that order) as command line arguments
- Report GFLOPS for the following sizes:
  - for i in 4, 8, 10, 14
    for j in 4, 8, 10, 14
    m,n=power(2,i)
    k = power(2,j)
  - m=4095,n=4097,k=125
- Report the number of DRAM transactions (read and write) and shared memory requests per transactions (read and write) for the following sizes:
  - m=4095,n=4097,k=125
  - m=4096,n=4096,k=128

2. Write cuda code for computing matrix matrix product of the form $C=A*B^T$ where A is an m*k (double) matrix and B is n*k (double) matrix and C is an m*n (double) matrix; without explicitly transposing B. The code should be tiled and should shared memory. Make sure that the code works for matrices of all sizes not just square matrices. Modify the given template code and use it.

- The name of the cuda kernel and the binary executable must be “matmul_double_t”
- The executable must accept m,n and k (in that order) as command line arguments
- Report GFLOPS for the following sizes:
  - for i in 4, 8, 10, 14
    for j in 4, 8, 10, 14
    m,n=power(2,i)
    k = power(2,j)
  - m=4095,n=4097,k=125
- Report the number of DRAM transactions (read and write) and shared memory requests per transactions (read and write) for the following sizes:
  - m=4095,n=4097,k=125
  - m=4096,n=4096,k=128