Homework 4 (never due)

Doing this homework will give you some experience with GPU programming through the CUDA framework. You will continue onwards from Homework 3's assignment on matrix-vector multiplication (MVM) and implement MVM on the GPU. After exploring the performance impact of various configurations, you will implement MVM using the GPU's shared memory and repeat your experiments.

You are provided a template code for matrix-vector multiplication **matrix_vector_mult.cu** that is ready to compile and run. You need to modify this template by implementing the following functions and kernels in CUDA and evaluate their performance.

1) The kernel for multiplying an nxn matrix, A, with a vector, x, to compute the vector y using n threads. You should assign one thread to the computation of each element of y.

2) Modify the matrix-vector multiplication kernel such that each thread block loads the vector x to its shared memory, thus avoiding repeated access to x in the global memory by each thread in the block.

Measure the execution time for different matrix sizes (1000x1000 and 5000x5000) and different values of BLOCK_SIZE (16, 64 and 256) and explain your observations from these results. You should compare the execution times in the two cases: when time for copying data from/to the GPU is included, and when this time is not included. You should also be aware that the time overhead for launching a kernel may dominate the time to execute the kernel.

Quick Start

- Log onto any of the following machines via SSH (the machine are in 6110 Sennott Square). These machines normally run linux on weekends and every night from 9:00PM to 8:00AM, but will run Linux continuously for the next 2 weeks to be used for this homework
  - pc6110-e1.cs.pitt.edu
  - pc6110-e2.cs.pitt.edu
  - pc6110-e3.cs.pitt.edu
  - pc6110-e4.cs.pitt.edu
  - pc6110-e5.cs.pitt.edu
- compile your cuda program using

```
/opt/cuda-7.0/bin/nvcc matrix_vector_mult.cu
```

- Run the a.out binary produced
- To know the number of SMs and cores in the GPU, you may use the Linux command "lspci" which will list all the devices attached to the PCI bus, including the GPU. Once you know the type/model of the GPU you can find its specs from the web. The GPUs connected to the above machines are GF119 (NVS 310).
- There are many cuda/gpu manuals. See, for example https://docs.nvidia.com/cuda/index.html