Lab Session 7

Submission deadline: April 10, 11:59pm

Please submit your lab results/code to CatCourses, including Makefile, source code and a short report (up to one page). You may find NVIDIA CUDA programming guide useful (https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html).

1. A simple test of your GPU hardware and CUDA programming environment

deviceQuery is a tool included in the CUDA toolkit. Using deviceQuery, you can check the basic hardware information of your GPU. Simply run "deviceQuery" and try to answer the following questions based on the output.

- How many GPU devices are there in your machine?
- What is the maximum amount of shared memory per thread block?
- What is the maximum dimension size of a thread block?
- What is the maximum number of registers available per thread block?
- What is the global memory size?

2. Find the maximum

In this task, you will write a CUDA program to find the largest element from an N-element input vector. Your program should take as input N and generate a randomized vector V of length N. Then it should compute the maximum value in V on CPU and GPU. The program should output the two computed maximum values and measure performance on CPU and GPU.

Remember that threads in different thread blocks cannot directly communicate. If possible, using shared memory will be helpful for performance optimization. Remember to synchronize all the threads within a thread block using the __syncthreads() function where necessary.