EE/CSCI 451
Spring 2019
Programming Homework 5
Assigned: April 2, 2019
Due: April 20, 2019, before 11:59 pm
Total Points: 50

1 Login to HPC

- The host is: hpc-login**3**.usc.edu
- Username and password are the same as your email account
- **Do not** run your program in the login node.
- After login, use the 'srun' command to run your program on a remote node. For example:
  srun **-n**4 ./run

1.1 MPI Examples

The "vector_add.cu" is the source codes used in discussions. To run a CUDA program, for example, follow the steps:

1. Login to HPC
2. Setup CUDA toolchain: type 'source /usr/usc/cuda/default/setup.sh'
3. **nvcc** -o run -O3 vector_add.cu
4. srun -n1 –**gres=gpu:1** ./run
5. Note: The job might take a long time (minutes) to submit.

2 Matrix Multiplication [40 points]

In the previous assignment, we implemented two approaches for performing matrix multiplication of 1024×1024 matrices A and B: (a) unoptimized matrix multiplication and (b) blocked matrix multiplication using shared memory and a block size of $b = 32$.

In this assignment, we will first fill in the blanks in the example code and then analyze the effects of the grid/block configuration over the performance of both the approaches.

- Approach 1 (unoptimized implementation using global memory only) [15 points]:
  - Name this program as 'p1.cu'
  - The value of each element of A is 1
  - The value of each element of B is 2
  - Thread block configuration: $b \times b$. (b is a power of 2)
  - Grid configuration: $\frac{1024}{b} \times \frac{1024}{b}$
  - After computation, print the value of $C[451][451]$
- Approach 2 (block matrix multiplication using shared memory) [15 points]:
  - Name this program as 'p2.cu'
  - The value of each element of A is 1
  - The value of each element of B is 2
  - Thread block configuration: $b \times b$
  - Grid configuration: $\frac{1024}{b} \times \frac{1024}{b}$
  - More details of this algorithm can be found in the paper 'Matrix Multiplication with CUDA' available with this assignment.
  - After computation, print the value of $C[451][451]$
- Report [10 points]: For both the approaches discussed above, your report should contain the following:
  - The execution times for various values of b (8, 16 and 32) and a brief discussion on the observations.
  - The maximum value of b (power of 2) that can be successfully used for the execution. (Optional) discuss why a higher value of b cannot be used.

3 Submission

You may discuss the algorithms. However, the programs have to be written individually. Submit the code and the report via ee451spring2019@gmail.com. Please make sure to include your name, student ID and the homework number in the PDF, and name your PDF file lastname_firstname_pa#.