EECE5640
High Performance Computing
Homework 5
*Submit your work on Canvas in a single zip file

1. (40) Let us revisit the histogramming problem assigned in Homework 4. Your input to this program will be integers in the range 1-1,000,000 this time (use a random number generator that generates the numbers on the host). Your host-based data set should contain N integers, where N can be varied.
   a. This time you will implement a histogramming kernel in CUDA and run on a GPU. You can choose how to compute each class of the data set on the GPU. Attempt to adjust the grid size to get the best performance as you vary N. Experiment with N = 2^{10}, 2^{15}, 2^{20} and 2^{25}. When the GPU finishes, print each of the classes in ascending order on the host (do not include the printing time in the timing measurements, though do include the device-to-host communication in the timing measurement). Plot your results in a graph.
   b. Compare your GPU performance with running this same code on a CPU using OpenMP.

2. (40) The code below carries out a "nearest neighbor" or "stencil" computation. This class of algorithm appears frequently in image processing applications. The memory reference pattern for matrix **b** exhibits reuse in 3 dimensions. Your task is to develop a C/CUDA version of this code that initializes **b** on the host and then uses tiling on the GPU to exploit locality in shared memory across the 3 dimensions of **b**:

```
#define n 64
float a[n][n][n], b[n][n][n];
for (i=1; i<n-1; i++)
   for (j=1; j<n-1; j++)
      for (k=1; k<n-1; k++) {
      a[i][j][k]=0.8*(b[i-1][j][k]+b[i+1][j][k]+b[i][j-1][k]
      + b[i][j+1][k]+b[i][j][k-1]+b[i][j][k+1];
     }
```

   a.) Evaluate the performance of computing a tiled versus non-tiled implementation in your GPU application.
   b.) What other optimizations could accelerate your code on the GPU?

3. (20) Read the Pascal whitepaper provided, and then identify the key features that were introduced in the Pascal P100 architecture, comparing those features against the Volta V100 architecture (make sure to identify the source for the information you obtained

on the V100). Please do not just repeat what you read in the Pascal whitepaper, go into more detail on each of the features you identify.

4. (50 points extra credit for both UG and Grad) Revisit the computation of Pi with darts in Homework 4. This time compute Pi using CUDA on a GPU. Report on the accuracy based on:
    a. The number of darts that are thrown.
    b. Pi computed with first single-precision variables and compare against using double-precision variables.

    * Written answers to the questions should be included in your homework 5 write-up in pdf format. You should include your C/C++ programs and the README file in the zip file submitted.