Shiny Patel CSE 5441 Section 12:45 – Lab 4 Report

Part 1

```
------------------------------------------------------------
Time taken for parallel version: 36.3 ms
Gigaflops/s: 59.23

Time taken for serial version: 3070.0 ms
Gigaflops/s: 0.70
------------------------------------------------------------
```

- **Report your results in estimated GFlops.**
  GFlops/s for serial version: 0.70
  for parallel version: 59.23

- **Measure both serial and parallel performance.**
  There was significant performance improvement since the serial version took 3070 ms while the parallel version only took 36.3 ms.

- **Report the CUDA compute structure (Grid, Block and Thread) you used and explain your results.**
  There are 1024 blocks in x dimension in grid and 1024 threads in x dimension (per block) which results in 1024 * 1024 (no. of elements in C) threads. Each thread runs the k-for loop (given in the serial version) to compute 1 element of C.

Part 2

- **Provide a timing and threshold convergence summary for all images.**
  Summary for coins.bmp

```
---------------------------------------------
Image Info ::
       Height=246 Width=300
Time taken for Cuda Sobel Operation: 3.7 ms
Threshold: 49

Time taken for Serial Sobel Operation: 100.0 ms
Threshold: 49
---------------------------------------------
```

- **Explain your cuda organization (grid, block, thread) distribution**.
  The program is designed to have (img_height - 2) blocks in x dimension in grid and (img_width - 2) threads in x dimension (per block). There are total (img_height - 2) * (img_width - 2) iterations in the serial version and each thread handles one such iteration. The x-index of block and thread is used to calculate the equivalent i and j values (given in the serial version).

- **Did you see any performance improvement in using GPU?**
  Yes, there was improvement in the time taken for Cuda execution (3.7 ms) vs serial execution (100 ms). Based on this, it can be concluded that for larger images there would be significant time difference.

  Note: Both lab parts 1 and 2 are designed to have the number of blocks in x dimension equivalent to the number of iterations of the outer-for loop *and* the number of threads in x dimension (per block) equivalent to the number of iterations of the 2nd for loop.