Shiraz University	**Shiraz University** **Department of Computer Science & Engineering**
	Professor: *Dr. Farshad Khunjush*
	TAs: *Alireza Rostami*
	Course: GPU Programming **Semester:** Fall, 2024
	Due Date: December 28th
	Homework #3

1. (100 points) Given a matrix **M** of arbitrary parameters, write two implementations of matrix sort, namely one sequential and one parallel in CUDA C, profile the performances of two implementations, and compare them.

   Formally, we define matrix sort as follows:

   $$M[i][j] \leq M[i][j+1]$$
   $$M[i][j] \leq M[i+1][j]$$

   The specifications are as follows:

   (a) You are hereby explicitly prohibited from flattening the matrix and sorting it out as an array. You must find other ways to achieve the desired operation.

   (b) Explain why a multithreaded design would theoretically improve performance. If the performance of your CUDA implementation is not better than your sequential one, you should be able to explain why.

   (c) Implement multiple sorting algorithms using CUDA (e.g. merge sort, quick sort, etc.) and analyze their performance in a highly parallel processor like a GPU. Implement at least two sorting algorithms and explain why you have chosen them.

   (d) Analyze the scalability of the CUDA implementation by sorting datasets of multiple sizes. The following sizes are desired: *i*) 1024, *ii*) 2048, *iii*) 4096, *iv*) 8192, *v*) 16384, *vi*) 32768, *vii*) 65536. (Side note: your system may not allow you to run large matrices. Therefore, run instances up to the largest size value your system allows you. If your system is too slow, you may use smaller powers of two for your matrix size.)

   (e) Implement multiple data streams and compare BFS and DFS schemes.

   (f) Compare different memory management schemes provided by CUDA (`cudaMallocManaged`, `cudaMalloc`, etc.)

   (g) Profile your code with different thread sizes. Discuss the effect of thread size on performance.

   (h) Use different data types and profile their effect on performance. The following data types are expected: *i*) `int`, *ii*) `float`, *iii*) `double`.

   (i) Report every command you use. If you wish to write a bash script, then include it in the files you send.

   (j) Report the profiling results and try to justify and make sense of them. Raw results are of no value. Agglomerate your benchmarks' results. Use plots and other tools to visualize data.