Advanced Computer Architecture 3rd Term 2022-2023
GPU Project

All submissions via GitHub.
Topic: GPU-CPU memory transfer

a.) Video discussion on CUDA concept

GPU-CPU memory transfer.
CPU and GPU are two separate processors; each has its memory. There are several ways in which data are transferred between processors, as follows:

- Unified memory introduced in CUDA 6
- Prefetching of data with memory advice
- Data transfer or initialization as a CUDA kernel
- Old method of transferring data between CPU and memory (memCUDA copy)

Discuss:

- Explain the concepts of each method and how it is implemented
- Compare (a) the execution time of the kernel as well as (b) the memory transfer time (taking into consideration page faults, overhead time if it's a separate transfer time, or part of the kernel transfer time already).

CUDA programming project specifications:
Write a 1D convolution of vector *in* and place the result in vector *out*.
Implement using:

- Unified memory introduced in CUDA 6
- Prefetching of data with memory advice
- Data transfer or initialization as a CUDA kernel
- Old method of transferring data between CPU and memory (memCUDA copy)

1D convolution is defined as:

```
out[i] = (in[i] + in[i+1] + in[i+2])/3.0f;
```

Notes:
1.) Vectors are of **floating-point** type.
2.) Use Google Colab platform (submit link with appropriate access right)
3.) Obtain the execution time of a specified function/kernel. For CUDA, use nvprof apps.
4.) For each version, time the process/kernel for vector size {2^{20}, 2^{22}, 2^{24}} and threads per block {256,512,1024}
5.) You may initialize each vector with any values for the data.
6.) A routine to compare the correctness of your output.
7.) output:
a.) Google colab notebook
b.) Text/pdf file containing the comparative execution time.
c.) Analysis of the data (Compare (a) the execution time of the kernel as well as (b) the memory transfer time (taking into consideration page faults, overhead time if it's a separate transfer time, or part of the kernel transfer time already; effect of various block sizes; effect of varying elements, etc.)

Rubric:

Video presentation	40
Cuda program (Unified) with correct implementation and output	5
Cuda program (Prefetch+advice) with correct implementation and output	5
Cuda program (data initialization as kernel) with correct implementation and output	5
Cuda program (old style using memcuda copy) with correct implementation and output	10
Comparative result	10
Analysis	15
Follow instructions	5