CODESIGN

LAB 1 : OPENCL Programming

Rq: For performances evaluation, processing Time (in seconds), or processing throughput (GFLOPS = Giga Floating Operations per Second) can be used.

Choose **COUNT=1**.

Report Deadline: April 9th 2025.

1- Give the characteristics of the opencl compatible devices installed on your PC: device reference, global, local, cache Memories sizes, Number of compute units (Streaming MultiProcessors), max number of work items, work groups, etc….

A) Matrix Multiplication Implementations: performances comparison *(8 pts)*

1- Compare the performances of the Matrix multiplication classic implementation on CPU (Sequential), and the openCL implementations .(only for CPU and for N =256; 512)

2- **For each** of the OpenCL compatible devices, give the performances (GFLOPS) of the 3 Implementations versions of matrix multiplication on GPUs (uncoalsced, coalsced and block_tiled), for following Matrix sizes (N) and Work-group sizes:

N = { 2048, 4096 , 8192}. For CPU only 2048

Work-group size = {2*2, 8*8, 16*16, 32*32}

- Compare the performances, ***interpret and explain*** the results (For ***each device separately***).
- For the values generating errors, explain the cause.

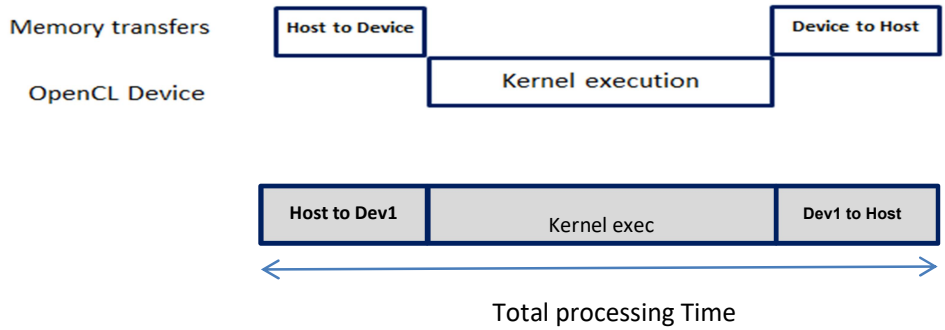
3- For the device with the best performance (NVIDIA Dedicated normally) and N=8192, compare the performance for the following implementations and Work-Group values:

- UNCOALSCED (1*32) vs COALSCED (32*32)
- COALSCED (1*32) vs UNCOALSCED (32*32)

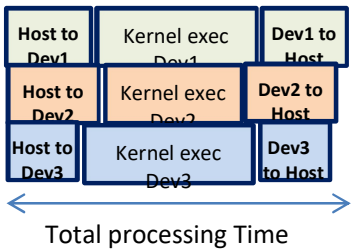
- **Interpret and explain** the results

B) Running the kernel on multiple OpenCL devices *(12 pts):*

When running a kernel on a OpenCL device, the total processing time is the durations sum of data transfers and kernel execution:



To improve the performance on a PC equipped with many **OpenCL** devices (3 generally: the CPU, the integrated GPU and the dedicated GPU), it is possible to split the processing (matrix multiplication in this case) on the 3 devices and execute the kernel instances in parallel (see fig below).



1- Our goal is to speed up the processing of the **UNCOALSCED** implementation on the dedicated GPU (NVIDIA) for **N=8192** and a **work-group size = 16*16**. The speedup will be evaluated:

Speedup = Processing Time (Nvidia) / Processing Time (3 openCL devices)

- Explain how you split the Matrix Multiplication on the 3 devices: justify the choice of the dimensions of the sub-matrices processed by each opencl device. (***5 pts)***

- Give the openCL Host code and the obtained results (speedup value). (*7* ***pts***)
  [Higher speed up → Higher Mark]
  Validation required only for the 10 best speedup values.