Université de Pau et des Pays de l'Adour
Département de Mathématiques
2022-2023

M2-BigData : GPGPU
Chapter 4 – Exercice 1

Objectives

Implement a basic dense matrix multiplication kernel. This is the first version of several optimizations to follow (see next exercises).

The program computes :

$$C = AB$$

where A, B and C are general rectangular matrices.

Instructions

Edit the code in the given code to perform the following :
— allocate device memory
— copy host memory to device
— initialize thread block and kernel grid dimensions
— invoke CUDA kernel
— copy results from device to host
— deallocate device memory

The program needs only the dimensions of matrix A and the columns number of B. These parameters are read from program arguments.

Your program is supposed to work with rectangular matrices of any size.

Questions

1. **Before coding** : What are the relations between the three matrices dimensions to have a well defined multiplication ?
2. How many floating operations are being performed in your matrix multiply kernel ? explain.
3. How many global memory reads are being performed by your kernel ? explain.
4. How many global memory writes are being performed by your kernel ? explain.
5. Compute the arithmetic intensity of your kernel. The arithmetic intensity is a FLOP/Byte number standing for the number of floating point operations performed per byte of global memory accessed.