Introduction to CUDA Parallel Programming

Homework Assignment 4

B07901069 電機四 劉奇聖

May 4, 2022

1 README

This file is `report.pdf`. `src/` folder contains two folders, `dot_product` and `heat_diffusion` , which are the source code for the corresponding problems. `result/` folder contains two folders, `dot_product` and `heat_diffusion`, which are the execution results for the corresponding problems.

In `src/dot_product` folder or `src/heat_diffusion` folder, executing `make` to compile the program.

To execute the program, run `./vecDot_ngpu` or `./heat_diffusion` and follow the instructions to enter the corresponding parameters.

2 Result

2.1 Dot-product

The file named `Input` contains the input parameters. The file named `Output` contains the program standard output.

The optimal block size is 128 and optimal grid size is 160000.

From the result we see that the gpu execution and the cpu execution result are very close. The error is due to different float point precisions between CPU and GPU.

2.2 Heat Diffusion

There are three running result: g1, g2a, g2b. g1 means only 1 GPU is used. g2a means 2 GPU are used and NGx = 1 and NGy = 2. g2b means 2 GPU are used and NGx = 2 and NGy = 1.

For these three results, `Input_[i]` are the standard input parameters, `Output[i]` are the standard output. `phi_initial.dat` is the initial temperature distribution. `phi_GPU.dat_[i]` are the final temperature distribution computed by GPU. Where i is g1, g2a, or g2b.

The optimal block size is (16, 16).

The computed phi result is the same for all running result.

3 Discussion

For dot product, the speed up of GPU is only 0.538518. This may because 40960000 is too small. We can see from result that the total GPU time is dominated by the I/O time.

For heat diffusion, the execution time of g2a and g2b are approximately the same. The execution time of 2 GPU is about 0.57 of the execution time of 1 GPU. The speed up is not 2 since some part of the computation cannot be parallelized.