Dustin (Ting-Hsuan) Ma
ME 2054
November 7th, 2018

Project 1: Finite Difference Solution of a Vibrating 2D Membrane

Introduction

Graphics processing unit (GPU) computing, also known as parallel computing is utilizing Nvidia's graphics computing codes to parallelize the serial codes that occurs on our central processing units (CPU). The benefits of parallelizing a serial portion of the code is to divide up the task for multiple different fast processing graphics cards to return a speed up in time needed for a large size calculation. Although CPU computing is sufficient for most of today's codes, scientists who are dealing with a large amount of data and needing to do numerous calculations can benefit from using GPU computing. By parallelizing a program, GPU computing carries out processes in a parallel manner (simultaneously), instead of the serial way of the CPU. Due to the ability to handle and compute large amount of data in a short amount of time, GPU computing has major contributions when it comes to performing large scale computations such as Fluid dynamics.

Usually these GPUs are kept inside a well ventilated cool room. Figure 1, shows an example of one of the supercomputing center in the world.

Figure 1. IBM's Gene/P massively parallel supercomputer.

Getting Started With CUDA

To parallelize one's code, the programmer must understand the basic functions that differentiates calling a function on the CPU and GPU. Figure 2, shows some basic functions and syntax programmers must use and learn to program using CUDA.

1) `__global__`
2) `cudaMallocManage();`
3) `cudaMemCpy();`
4) `Kernel<<<nBlocks,threadsPerBlock>>>();`

Figure 2. Examples of general codes and functions to know in CUDA.

While the new language may seem daunting, coders should know that CUDA works with C, therefore converting their daily C code into CUDA only requires an extra understanding of how the GPU computes, and process information.

Inside the GPU

Inside the GPU, there are two sub categories. The first one being blocks, and second ones being threads. Blocks and threads can be specified by the user. Nvidia gives their best practice value of 256 threats per block to have the most computational increase in speed up. While a desired number is given for a performance boost, users can set their own values to test the speed up in their code.

Dustin (Ting-Hsuan) Ma
ME 2054
November 7th, 2018

Procedures

In this project, we implemented the wave function of a 2d membranes whose boundaries are clamped. This 2d membrane vibrates up and down with the given formula below,

$$\phi_{i,j}{}^{t+1} = 2\phi_{i,j}{}^{t} - \phi_{i,j}{}^{t+1} + \frac{c^2\Delta t^2}{H^2}(\phi_{i+1,j}{}^{t} + \phi_{i-1,j}{}^{t} + \phi_{i,j+1}{}^{t} + \phi_{i,j-1}{}^{t} - 4\phi_{i,j}{}^{t})$$

We had to first initialized the wave function with the formula below,

$$\phi(x, y, t) = 0.1(4x - x^2)(2y - y^2)$$

After initialization, boundary conditions should be applied to all x and y that are at the boundary. In this problem the membrane size is Lx = 4 (units), and Ly = 2 (units). Therefor for all x = 0, x = Lx, y = 0, and y = Ly, the amplitude should be zero.

Methods

I first wrote my serial code and ran it on the CPU. For all Cases, I am running with an end time of 1 second. This gives me the first maximum amplitude. While coding, I used a mesh size of 41x21 to ensure that my exact solution matches with that of my numerical serial code. Initial condition visualization is shown in figure 2.

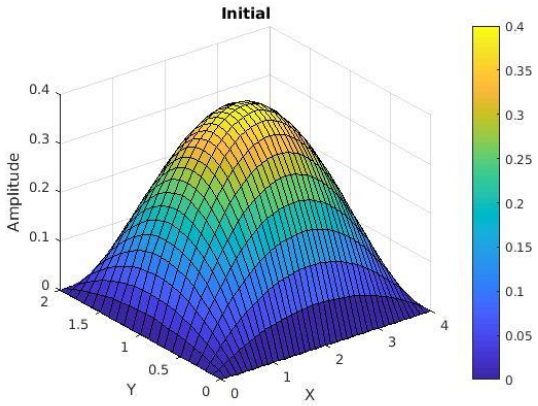

Figure 2. Initial condition of 2d vibrating membrane

After getting the CPU portion of the code working. I started implementing the GPU portion. In this portion I parallelized the code so that it will still only calculate within the boundaries.

Noticeable Speed Up

Speed up tables are shown below. With Table 1 comparing wall time between different sized meshes, and Table 2 changing in size block size. In Table2, code is ran using a mesh size of 1026 x 514. Comparing the different mesh sizes, speed up can be observed when larger block size is used to solve the mesh. 32x16 is more efficient than 32 x 32. This can be due to part of the problem parameter. Having a block size more closely to that of the problem statement may increase the efficiency of the program overall.

Table 1. Comparing different CPU Mesh size execution time, End time = 1 0[s].

Mesh Size	CPU Time [s]	GPU Time [s]
514 x 258	2.08393	0.01764
1026 x 514	16.66819	0.06144
2050 x 1026	133.3	0.30871

Table 2. Comparing Execution configuration of different Block Sizes, End time 10 seconds, and Mesh size 2050x1026

Block Size	GPU Time [s]
8 x 8	0.38446
16 x 16	0.29927
32 x 16	0.28260
32 x 32	0.28769

Figure 3 and figure 4 shows how close the match is between the final GPU output and the exact solution. During this end time, the amplitude is negative, but it is still bounded between -1 and 1.

Dustin (Ting-Hsuan) Ma
ME 2054
November 7th, 2018

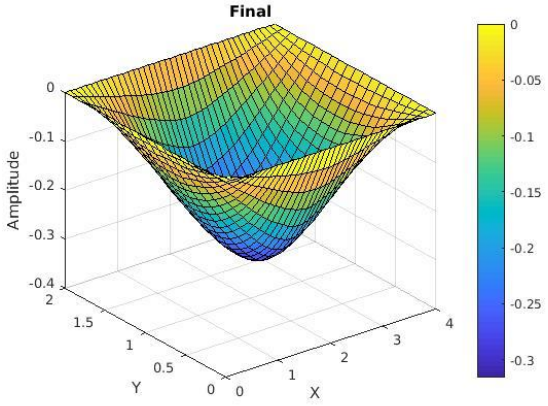

Figure 3. Final numerical output at ENDTIME = 1, and Mesh size = 41x21

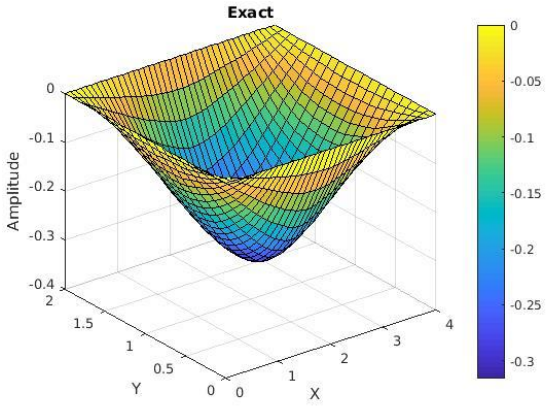

Figure 4. Exact solution to check final output

In Figure 5, figure 6, and figure 7, I took a slice in the z-direction to look at the amplitude accuracy of each computation method at different sections of the 2d membrane. In the graphs, I took it as 0.25Lx, 0.5Lx and 0.75Lx. You can see that the CPU and GPU lines almost fall exactly on top of each other, while they are a little off from the exact solution line.

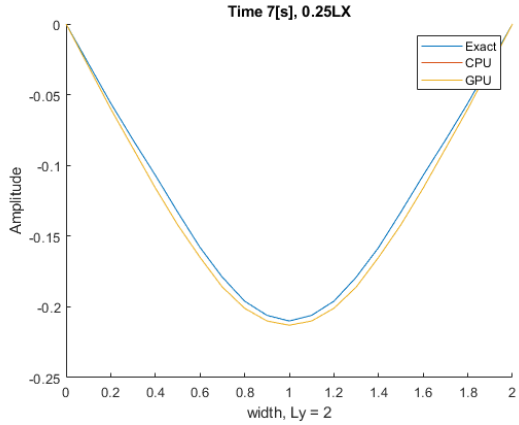

Figure 5. Time = 7[s], 0.25LX exact, CPU, and GPU comparison

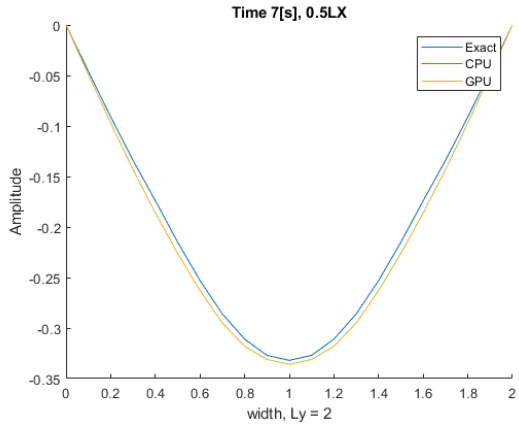

Figure 6. Time = 7[s], 0.5LX exact, CPU, and GPU comparison

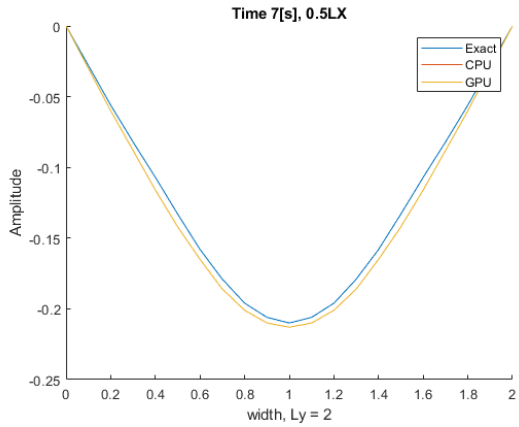

Figure 7. Time = 7[s], 0.75LX exact, CPU, and GPU comparison

Dustin (Ting-Hsuan) Ma
ME 2054
November 7th, 2018

Conclusion

In this lab, I implemented GPU parallelism into the oscillating wave formula for a 2d vibrating membrane. The results have shown that GPU is most beneficial when a large problem is at hand. I also noticed that when setting up the GPU environment, it is beneficial to design the environment to match some physical aspect of your problem to gain some speed up.

Some hiccups that I've encountered during this project was understand the usage of cudaMallocmanage() Function. I figured out that with such function, there is no need to copy memory back and from the GPU to read and receive data. Pointer address swapping also proved to be very useful. I learned how to linearly index a 2d array within the CUDA environment.

My GPU is also way faster than the CPU. This is seen to be way too fast, but if GPU outputs in [ms] and CPU outputs in [s]. I can compare the two by taking the multiple of a thousand. When I do that, my GPU will appear slower than the CPU, which is incorrect. To test whether this is true, I ran the CPU and GPU code separately, the CPU code takes some times, while the GPU code give me an output that matches the exact answer almost immediately.

References

Web.archive.org. (2018). HMPP Competence Center | Addressing the Many-Core Computing Challenge. [online] Available at: https://web.archive.org/web/20130616205308/http://openhmpp.org/ [Accessed 7 Nov. 2018].

W. H. (www.winkhosting.com), "OpenCL Compiler, Embedded Source, Encryption," ClusterChimps.org. [Online]. Available: https://web.archive.org/web/20111101184143/http://www.clusterchimps.org/ocltools.html. [Accessed: 07-Nov-2018].

NVIDIA Developer, 10-Oct-2018. [Online]. Available: https://developer.nvidia.com/nvidia-developer-zone. [Accessed: 07-Nov-2018]