ECEN 5593: ADVANCED COMPUTER ARCHITECTURE

FINAL PROJECT

Matrix Multiplication Research and Comparison

SANJANA KALYANAPPAGOL saka2821@colorado.edu

12th May, 2017

SPRING 2017

ABSTRACT:

Utilizing graphics hardware for general purpose numerical computations has become a topic of considerable interest. The implementation of streaming algorithms, typified by highly parallel computations with little reuse of input data, has been widely explored on GPUs. The matrix multiplication is the key operation for many computationally intensive algorithms. It is a kernel operation used in many transform, image and discrete signal processing application as well as robotic applications. Matrix mathematics applies to several branches of science, as well as different mathematical disciplines. It is also used in the field of Computer Graphics. Nowadays, we can observe the results of matrix mathematics in every computer-generated image that has a reflection, or distortion effects such as light passing through rippling water, etc. Before computer graphics, the science of optics used matrix mathematics to account for reflection and for refraction and It also helps calculate the electrical properties of a circuit, with voltage, amperage and resistance. Because of its trending applications, new algorithms and new techniques are being developed on configurable devices too. Its regular data access pattern and highly parallel computational requirements suggest matrix-matrix multiplication as an obvious candidate for efficient evaluation on GPUs but, surprisingly near optimal GPU implementations are pronouncedly less efficient than current cache-aware CPU approaches

The aim of the project is to develop Matrix Multiplication that is implemented using NVDIA's CUDA (Compute Unified Device Architecture). This project consists of the concepts of GPU, CUDA and C programming to implement Matrix Multiplication.

INTRODUCTION:

A matrix is represented as a rectangular array of numbers which has one of the most useful and fundamental implementation in the field of mathematics and scientific computation. For matrix multiplication, we have two matrices that have 'i' rows and 'j' columns. We use the SIMT code on a CUDA enabled device. SIMT refers to 'single instruction, multiple thread', a model where many threads running in parallel to execute same instruction at the same time. The CUDA kernels are used for parallel execution and scale from a manycore GPU to massive distributed system consisting of hundreds of processors.

The threads in CUDA consists of their own program counters and registers. They have a 'global memory' and a 'shared memory' that is more limited in size. Within the same block the thread shares the instructions and execute them in parallel. They diverge to execute serially and once that execution is successfully implemented they converge to start parallel execution. The threads are grouped into Blocks. Each block can hold up to 512-1024 threads. Hence multiple blocks are used to compute the result of the matrix multiplication. The blocks can be 1 - Dimensional, 2 - Dimensional and 3 - Dimensional. To make it all system compactible we use the size of 16*32.

The thread hierarchy is explained below in the figures. In the CUDA memory hierarchy each thread has an access to private local memory for storing large data. All threads have access to shared memory which can be accessed by the threads of that block, whereas the global memory is accessible to all the threads. A kernel can be launched by specifying the grid dimensions, block

dimensions and the kernel function to run on the device. These functions act as an entry for GPU computations. At the time of kernel launch on the GPU, a grid of thread blocks is created and the blocks are queued to run on the GPU. The CUDA programming model requires that these blocks can compute in any order, that is, the programmer may make no assumptions about the order in which the GPU schedules and runs the blocks of threads. The program must run correctly regardless of the order in which the blocks are scheduled.

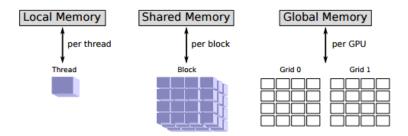


Fig: Memory hierarchy and its relation between thread hierarchy

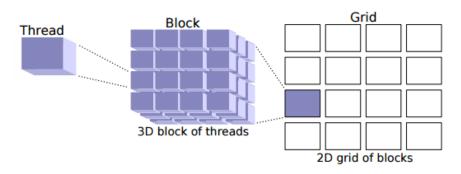


Fig: CUDA thread Hierarchy

METHODOLOGY:

The kernel function is declared as __global__ in the code and a global function on the device and passing parameters is used. The variables that are used can hold triple unsigned integers. An example syntax where the variables to which the threads have access is:

GlobalFunction<<<variables that hold 3 integers>>> (parameters)

Implementation of Matrix Multiplication:

The matrix multiplication will consist of the following:

- 1. __global__ keyword for declaring that it is an entry point function for running code
- 2. Declarations to reserve a register to hold the value where the product of the row and column entries.
- 3. Declaring the row and column of the matrix and its boundaries.

- 4. A condition check that terminated if the row and column is outside the boundary of the product matrix.
- 5. Loop to iterate over the values of the row and column of that matrices that must be multiplied to obtain the product matrix.

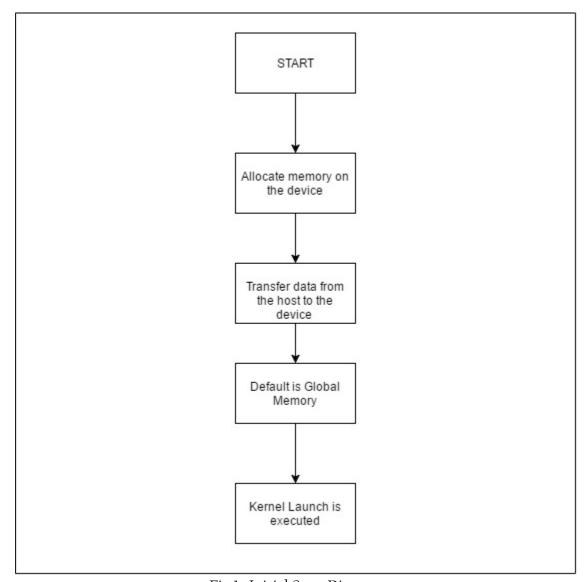


Fig 1: Initial Start Diagram

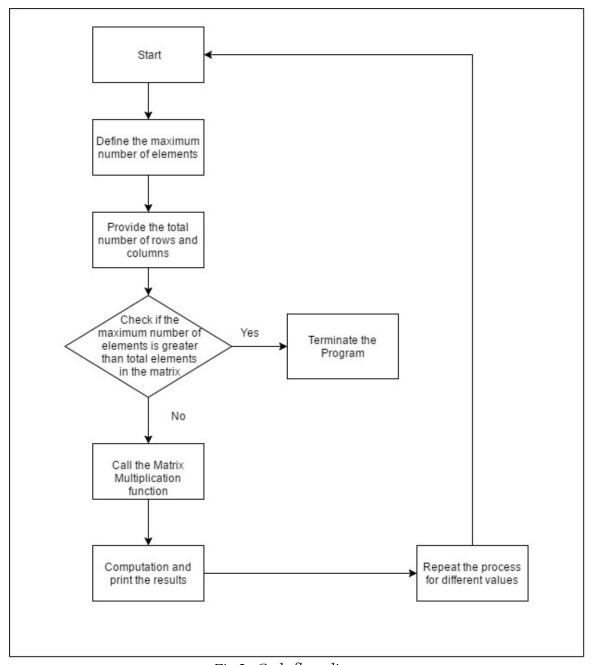


Fig 2: Code flow diagram

The two matrices say A and B are stored in the global memory as a 1 - D array with the first row followed by the second row, and so on. Thus to find the index in this linear array of the (i, j) - entry of matrix A, for example, we compute (i × width of A) to find the starting index of the ith row, and then add j to go to the jth entry in that row. Finally, the last line of the kernel copies this product into the appropriate element of the product matrix C, in the device's global memory.

In this project, I have computed the CPU and GPU timing for different sizes of the matrix and compared them with one another. The variations have also been indicated.

RESULTS AND ANALYSIS:

The computations for various matrix sizes 10*10 to 300*300 matrix were implemented and analyzed. The CPU and GPU times are presented in the table below supported by the screenshots.

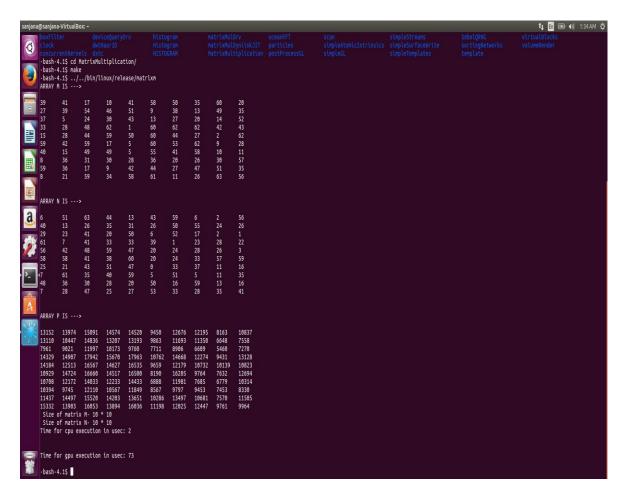


Fig 3:10*10 Matrix Multiplication

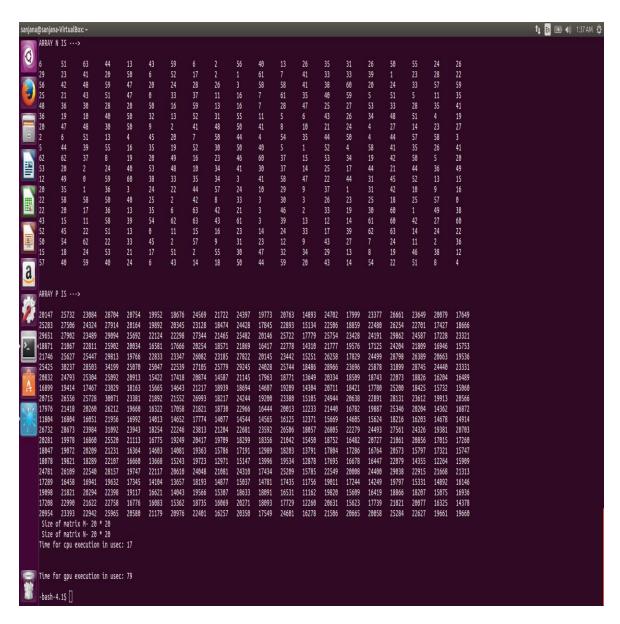


Fig 4: 20*20 Matrix Multiplication

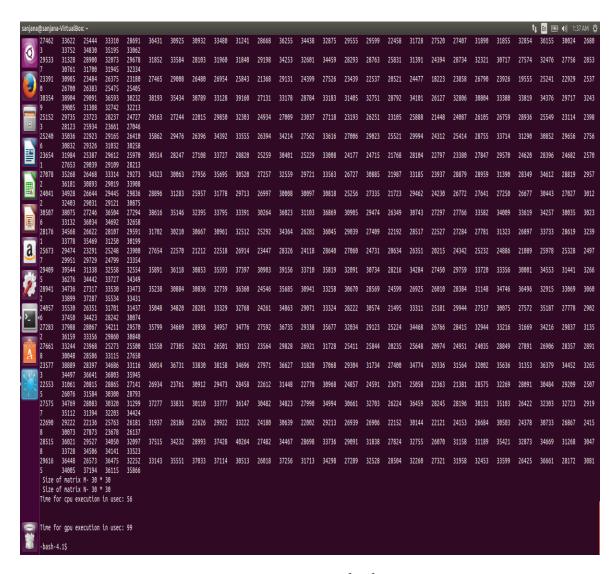


Fig 5: 30*30 Matrix Multiplication

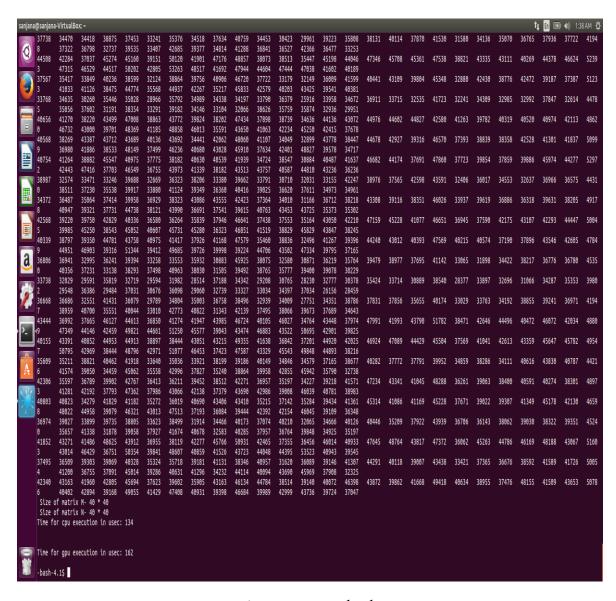


Fig 6: 40*40 Matrix Multiplication

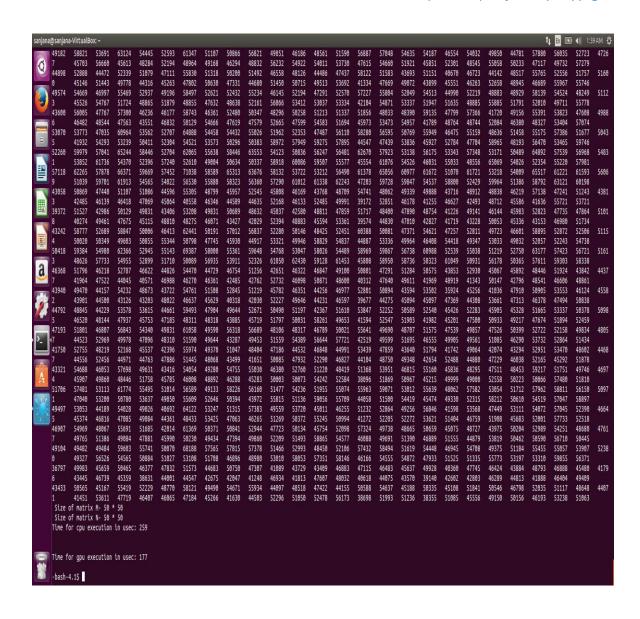


Fig 7: 50*50 Matrix Multiplication

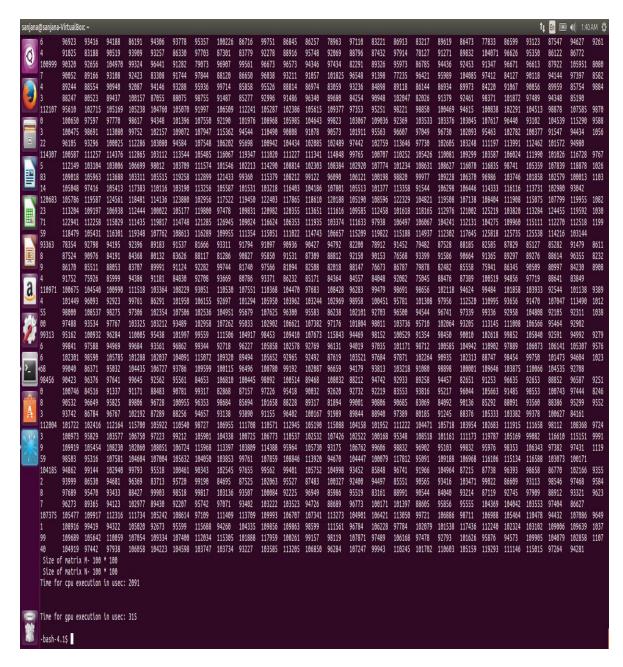


Fig 8: 100*100 Matrix Multiplication

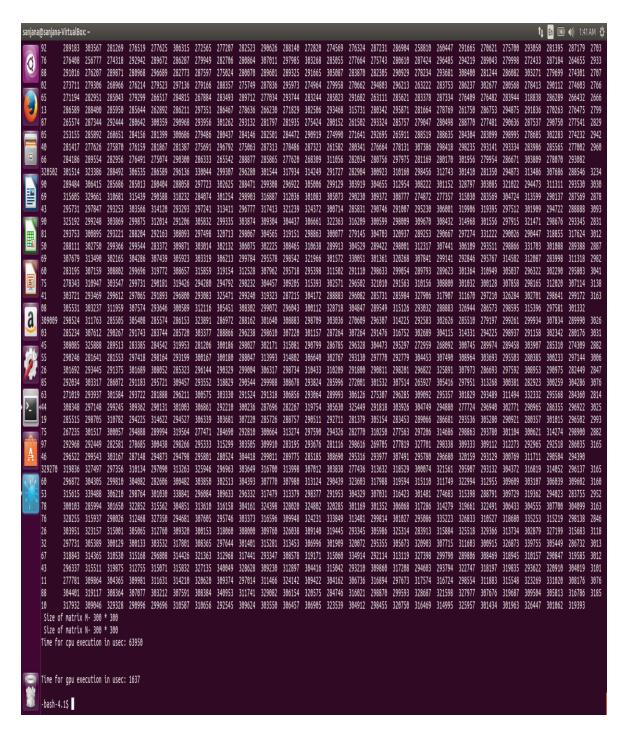


Fig 9: 300*300 Matrix Multiplication

Table 1: Execution times for different sizes of the matrix

SIZE OF MATRIX	CPU EXECUTION TIME IN MICROSECONDS	GPU EXECUTION TIME IN MICROSECONDS
10*10	2	73
20*20	17	79
30*30	56	99
40*40	134	162
50*50	259	177
100*100	2091	315
300*300	63950	1637

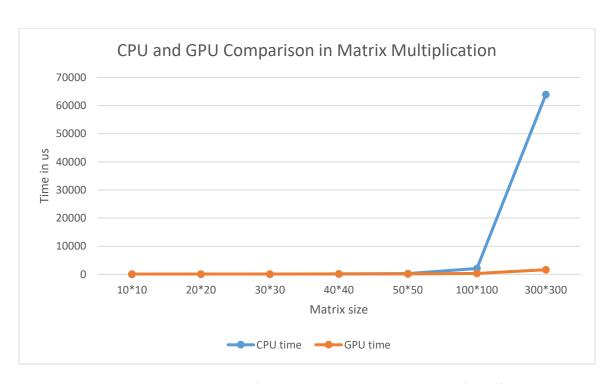


Fig 10: Graphical Representation of the CPU and GPU execution time for different sizes

APPLICATIONS:

There are various applications of Matrix Multiplication. Few of them are listed below:

- 1. Numerical Analysis.
- 2. Pattern Recognition.
- 3. Scientific Computing.
- 4. Used in the Computer Graphics industry involved in Video Gaming domain where construction and manipulation of realistic animation of a polygonal figure.

CONCLUSIONS:

- 1. CPU and GPU execution time varies with respect to the size of the matrix and it is observed that the CPU execution time changes widely as the size varies compared to the GPU execution time.
- 2. Using CUDA to implement scientific applications on the GPU we can fully exploit SIMD programming to populate with work the cores the GPU possess.
- 3. Floating point bandwidth from the closest cache GPU is several times lower than the current CPU.
- 4. There is a utilization of 17% of arithmetic resources on the NVIDIA hardware.

FUTURE SCOPE:

- 1. The code can be extended to bigger sizes and the execution times of CPU and GPU can be verified.
- 2. In terms of application wise it can be implemented for security which involve facial recognition or finger print recognition.
- 3. It can be extended to be implemented for performing mathematical computations for 3 Dimensional figures.

REFERENCES:

- https://graphics.stanford.edu/papers/gpumatrixmult/gpumatrixmult.pdf
- https://people.eecs.berkeley.edu/~sangjin/2013/02/12/CPU-GPU-comparison.html
- http://www.nvidia.com/object/cuda-home.html
- http://www.shodor.org/media/content/petascale/materials/UPModules/matrixMultiplicati on/moduleDocument.pdf
- The GPU on Matrix Matrix Multiply: Performance Study and Contributions by *Jose Maria Cecilia, Jose Manual Garcia*, *Manuel Ujaldon*.
- http://www.nvidia.com/content/nvision2008/tech_presentations/Game_Developer_Track/ NVISION08-Image_Processing_and_Video_with_CUDA.pdf
- http://supercomputingblog.com/cuda/advanced-image-processing-with-cuda
- http://homepages.math.uic.edu/~jan/mcs572/devmemdata.pdf
- http://cuda-programming.blogspot.in/2013/01/shared-memory-and-synchronization-in.html