Reporte Técnico

Arthur Alves Araujo Ferreira

Septiembre 2018

Resumen

Este reporte habla de la multiplicación matricial implementada en c++, c++ con hilos de openmp, y cuda. Se miden los tiempos de ejecución de cada instrucción y se hace un análisis de los resultados.

1. Introducción

El problema de multiplicación matricial es muy importante para la computación porque tiene una multiplicidad de usos muy grande desde ambientes tridimensionales, a cálculo de física de juego y manejo de imágenes, entre otras. Por esta razón, poder hacer cálculos matriciales de forma rápida es algo muy deseado. Los cálculos en la computadora típicamente se hacen con el procesador, pero la CPU solamente puede hacer una cierta cantidad de cálculos a la vez. Para este ejercicio, se busca hacer los cálculos tanto en CPU como con la tarjeta gráfica que puede hacer miles de cálculos a la vez, aunque no originalmente se haya diseñado con este propósito en mente.

2. Desarrollo

Se usarán las librerías y desarrollo de GPU para tarjetas gráficas de Nvidia con CUDA. Serán medidos y comparados los tiempos con diferentes tamaños de matrices en la CPU, CPU con threads y GPU en distintas configuraciones. Se corrió el programa en una CPU i7 de laptop y GPU GeForce 840M.

3. Resultados

En todos casos se prueba con 20 repeticiones y se reporta el promedio en milisegundos.

CPU

1000 * 1000	**2000 * 2000**	**4000 * 4000**
7487.82	83602.9	760759

CPU con openmp threads

1000 * 1000	**2000 * 2000**	**4000 * 4000**
4911.59	49558.7	434886

GPU con grid 1D

Dimensión de bloque	**1000 * 1000**	**2000 * 2000**	**4000 * 4000**
32	515.2222	2636.3089	10008.5068
64	515.0919	2610.4669	16161.6474
128	516.0321	2615.4845	16944.8085

GPU con grid 2D y bloque de 1D

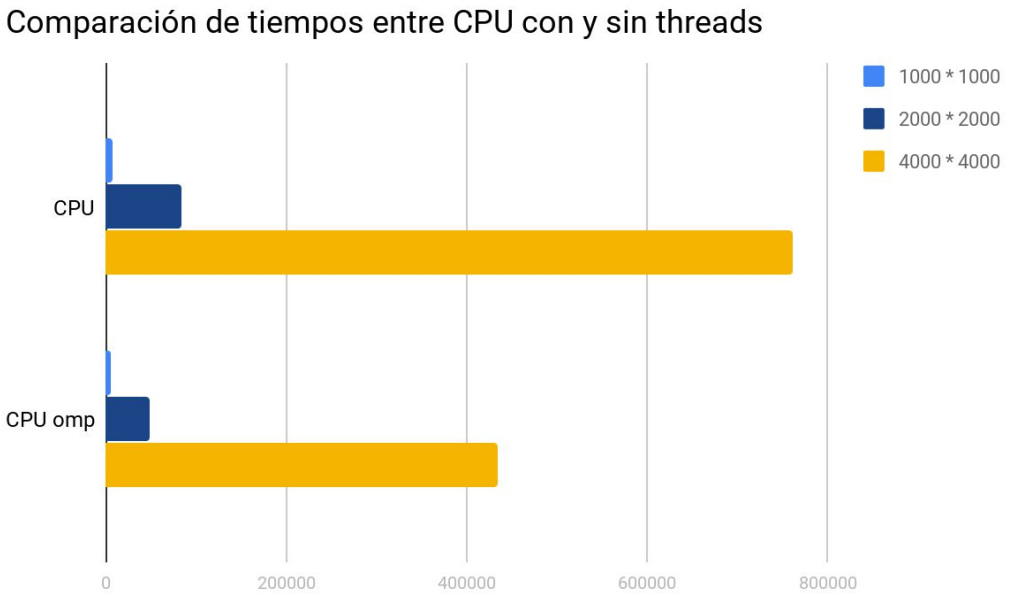
Dimensión de bloque	**1000 * 1000**	**2000 * 2000**	**4000 * 4000**
32	247.8148	2426.3592	19677.6931
64	202.6043	2308.9283	18974.5734
128	183.6006	2214.2006	18986.9453

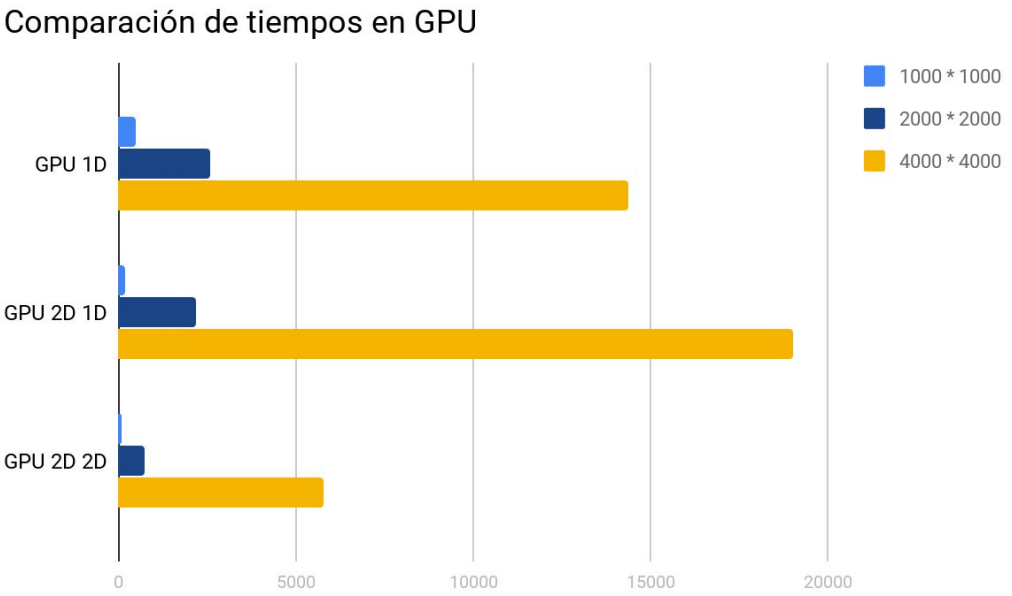
GPU con grid 2D y bloque de 2D

Dimensión de bloque	**1000 * 1000**	**2000 * 2000**	**4000 * 4000**
32 x 32	96.7550	738.3350	5793.7845

4. Análisis

Cómo se puede notar de las tablas de los tiempos y con las gráficas que se ven en la siguiente página, el tiempo crece exponencialmente con el tamaño de las matrices. Se puede notar que CPU con hilos logra bajar tiempos casi a la mitad. En GPU la reducción es mucho mayor, en el caso de la matriz de 4000 x 4000, con GPU en su peor caso logró una disminución de 97% del tiempo comparado con CPU y en el mejor caso logró una reducción del 99%.





Gráfica 1. Escala mayor 800000 ms (13 ⅓ min.); Gráfica 2. Escala mayor 20000 ms (20 seg.).

5. Conclusiones

Los resultados de tiempos en tarjeta gráfica son impresionantes porque pueden llegar a hacer el mismo cálculo que en el procesador pero en menos del 1% del tiempo. En el peor de los casos lo logró en 3% del tiempo que en cpu. La gpu tiene mucho mejor desempeño para cálculos repetitivos y pesados para cpu.

6. Referencias