Primer Parcial de HPC

Sebastian López Martínez

1088299807

Universidad Tecnológica

Pereira 30 de septiembre de 2015

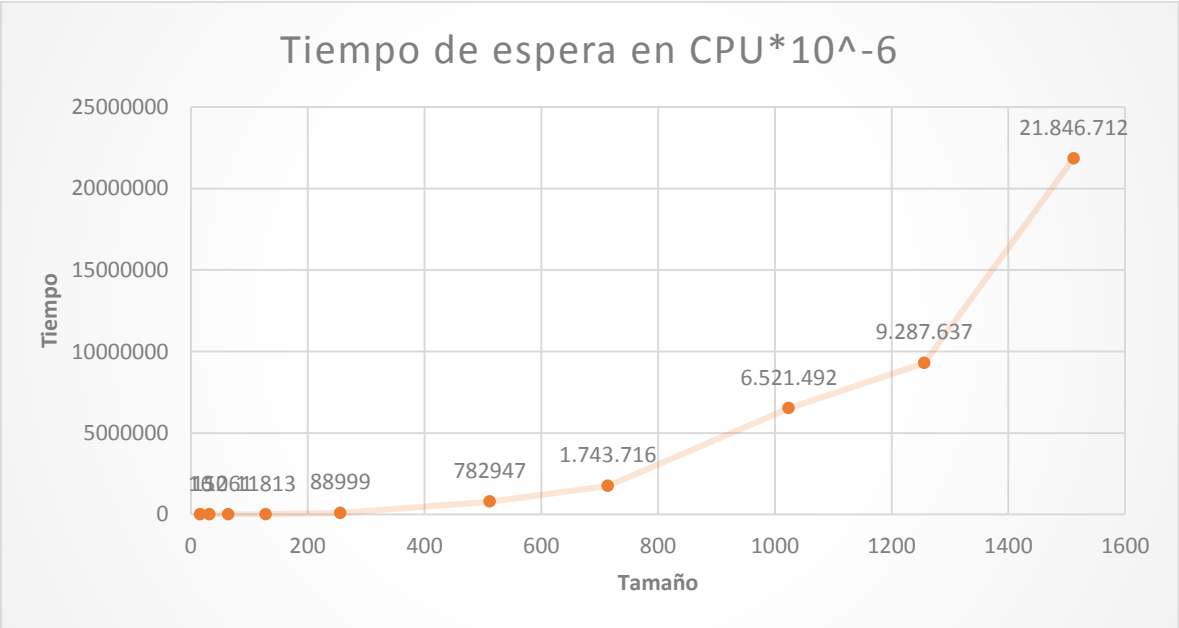
Introducción

El siguiente documento contiene una comparación estadística y grafica de los tiempos de ejecución de un mismo programa pero con distintas metodologías de uso de memoria en la ejecución. La funcionalidad de dicho programa es sobre la multiplicación de matrices cuadradas de diferentes tamaños.

El primer cuadro muestra los datos obtenidos de la ejecución secuencial del programa y los tiempos que le tomaron a la CPU resolver el problema, Los tiempos varían en función del tamaño del problema.

Tamaño	Tiempo de CPU *10^-6
16	16
32	152
64	1061
128	11813
256	88999
512	782947
714	1.743.716
1024	6.521.492
1256	9.287.637
1512	21.846.712

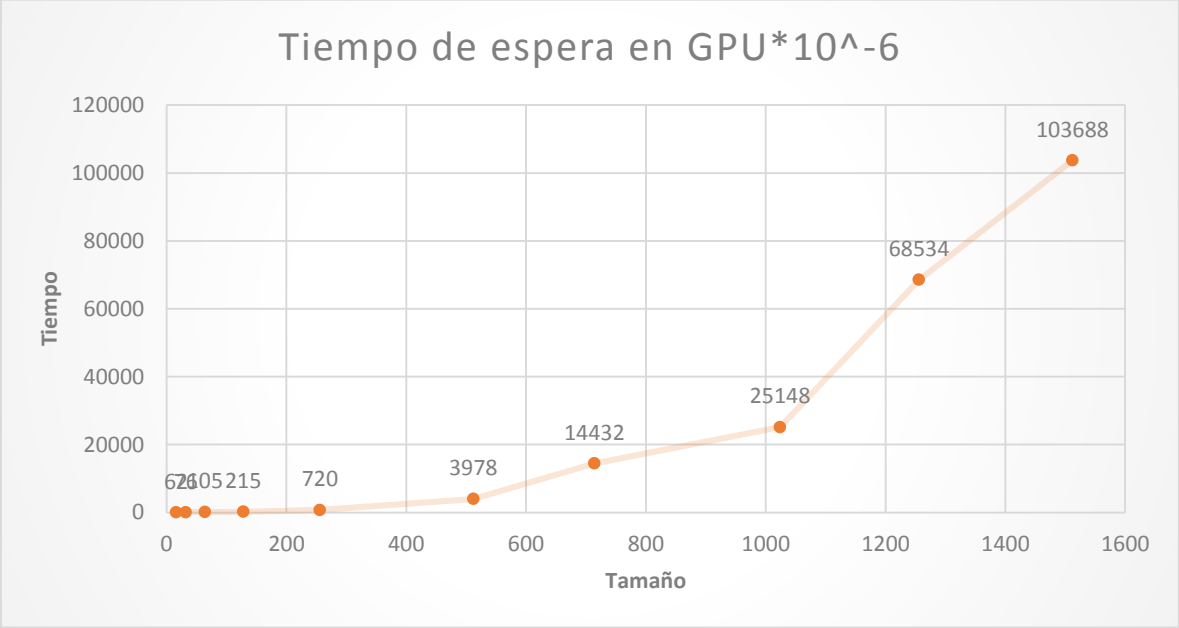
Estos datos revelan como los tiempos de ejecución van incrementando a medida de que el tamaño del problema aumenta alcanzando tiempos de varios segundos en el mayor tamaño en vez de micro segundos como en el primer caso. La siguiente grafica muestra el crecimiento de los tiempos de la ejecución:



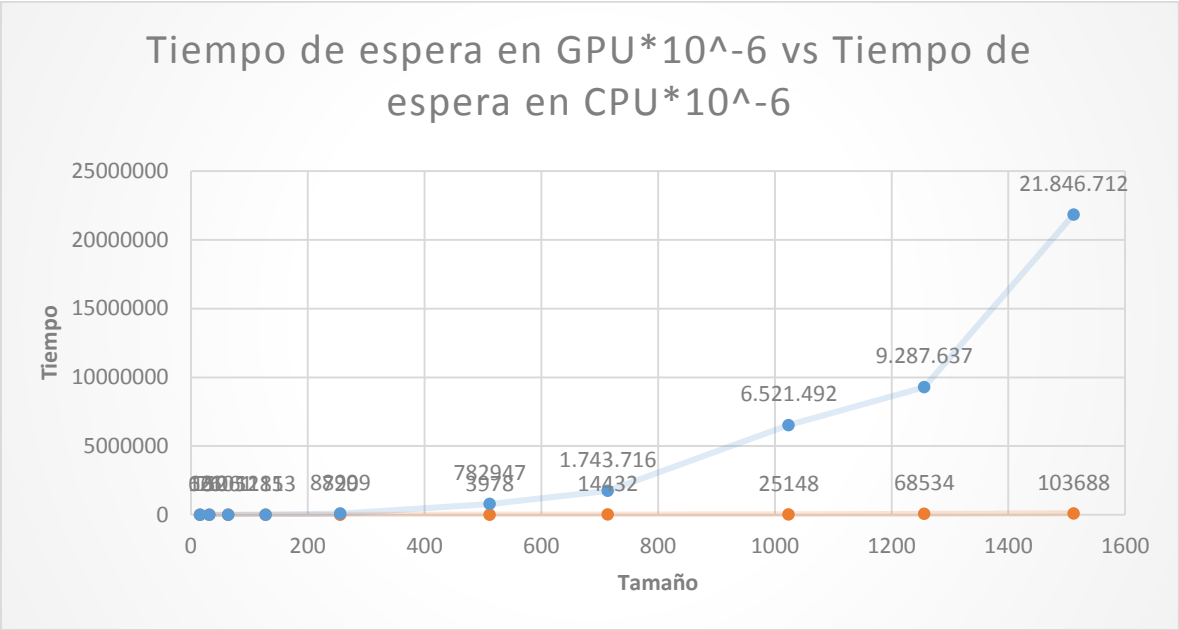
El siguiente cuadro de análisis representa los tiempos de ejecución del programa sobre la GPU incluyendo los tiempos de trasferencia de datos.

Tamaño	Tiempo de GPU *10^-6
16	62
32	76
64	105
128	215
256	720
512	3978
714	14432
1024	25148
1256	68534
1512	103688

A diferencia de los datos anteriores los tiempos de ejecución de la GPU en memoria global muestran una reducción en los tiempos de ejecución exceptuando en el menor tamaño del problema debido a que se pierde tiempo en la transferencia de los datos en ambas direcciones.



Sin embargo, a medida que el tamaño del problema aumenta los tiempos de ejecución se vuelven mucho más eficientes en comparación a los de la CPU demostrando una diferencia en tiempos significativa así como muestra la siguiente grafica



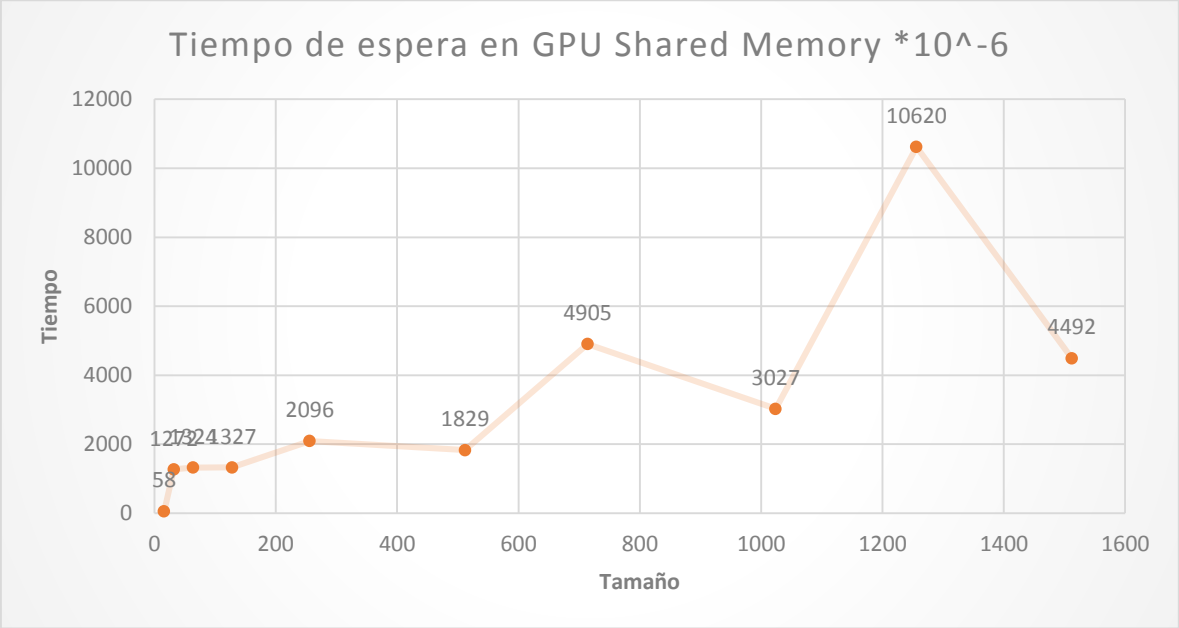
Los tiempos de ejecución de la GPU no crecen en igual medida que los de la CPU a pesar del tiempo de transferencia de datos del host al device, demostrando una aceleración considerable en la última medida del problema.

Por último los siguientes datos corresponden a la ejecución en la GPU pero en memoria compartida, donde se realiza una copia de los datos los cuales se deben utilizar varias veces durante la operación en la GPU. Al igual que en memoria global la transferencia de datos consume tiempo el cual es agregado a la medición de la ejecución.

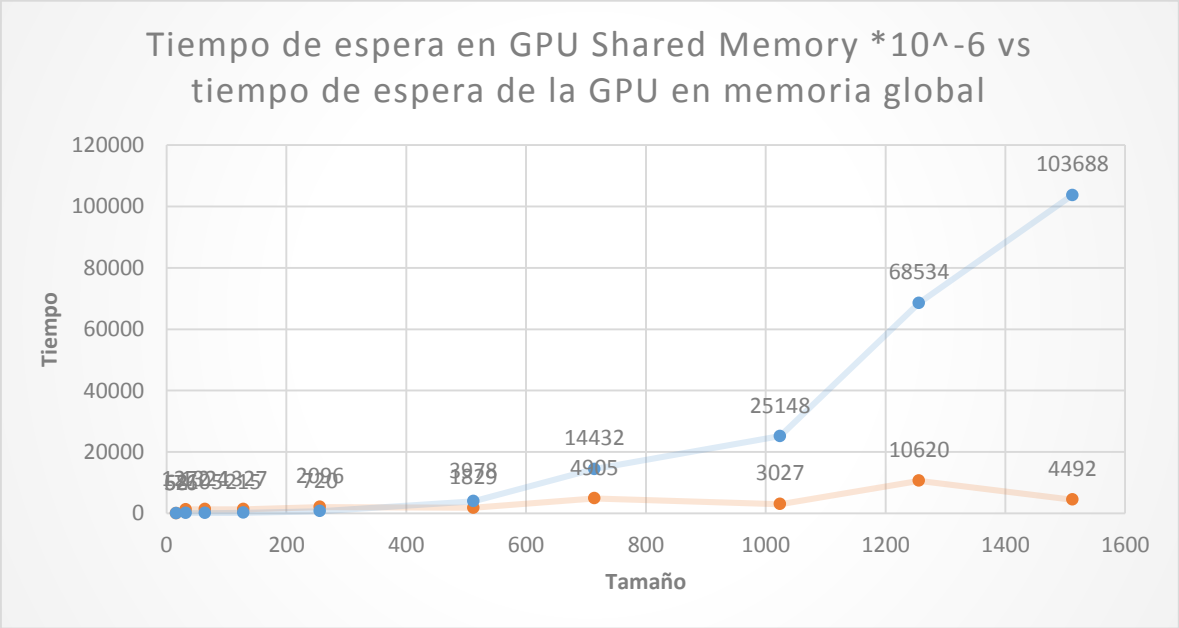
Tamaño	Tiempo de GPU en memoria compartida *10^-6
16	58
32	1272
64	1324
128	1327
256	2096
512	1829
714	4905
1024	3027
1256	10620
1512	4492

Los tiempos de ejecución de este método muestran un crecimiento en los tamaños pequeños debido al coste de transferir los datos de la memoria global a la

memoria compartida, sin embargo, a medida que los tamaños son más grandes se mantiene un crecimiento más reducido y los tiempos se van viendo afectado por el envío y recepción de los datos del host al device al igual que la memoria global.



Aunque a diferencia de la memoria global, en la memoria compartida se mantiene la persistencia de ciertos datos los cuales son usados varias veces dentro de la operación y también se hace una repartición del problema dentro de la memoria permitiendo que se vea la diferencia de tiempos con respecto a la global, tal y como lo demuestra la gráfica.



La diferencia se hace notable a pesar de que se ejecutó en la GPU pero el uso de diferente memoria del dispositivo en las operaciones remarca una reducción de tiempos en los dos métodos. Haciendo más grande la diferencia entre la GPU y la CPU.