Juan Siverio Rojas

20/01/2021

Proyecto gpu

Contenido

[Objetivo 2](#_Toc62032678)

[Consideraciones 2](#_Toc62032679)

[Previos 3](#_Toc62032680)

[Descripción 3](#_Toc62032681)

# Objetivo

Se realizará un histograma de un vector V de un número elevado N de elementos enteros aleatorios. El histograma consiste en un vector H que tiene M elementos que representan "cajas". En cada caja se cuenta el número de veces que ha aparecido un elemento del vector V con el valor adecuado para asignarlo a esa caja (normalmente cada caja representa un rango o intervalo de valores).

En nuestro caso, para simplificar la asignación del elemento de V a su caja correspondiente del histograma, vamos a realizar la operación ValorElementoV Modulo M, que directamente nos da el índice de la caja del histograma al que pertenecerá ese elemento y que deberemos incrementar. Se sugiere como N un valor del orden de millones de elementos y como M, 8 cajas.

Se utilizarán varios algoritmos basados en GPU Nvidia, con su programación basada en CUDA. También se ha programa la versión del algoritmo para CPU.

Se realizarán diferentes pruebas y se medirán los tiempos para hacer comparaciones.

# Consideraciones

## Previos

El programa se adecua a varios parámetros mediante unas definiciones previas. Se puede parametrizar lo siguiente:

**NUMELEMENTSV ->** número de elementos del vector de que se quiere calcular su histograma.

**NUMELEMENTSH ->** número de elementos del histograma.

**TREADSPERBLOCK ->** número de hilos por bloque. Siendo este adaptado al mayor posible que soporta la GPU en caso de indicar un valor superior a este.

**LOOPS ->** Indica el número de repeticiones que se ejecutará cada algoritmo, de esta forma obtendremos resultados más fiables y también el impacto de las memorias cache, en la ejecución de los diferentes algoritmos.

* **Los valores con los que se han realizado las pruebas de rendimiento son las siguientes.**

#define NUMELEMENTSV 33554432;

#define NUMELEMENTSH 8;

#define THREADSPERBLOCK 1024;

#define LOOPS 5;

**Nota:** Se debe considerar que los números de bloque sean Par, ya que con los algoritmos de reducción, aún no está completamente contemplada la posibilidad de los bloques impares. Para cumplir con esta restricción, solo hay que tener en cuenta que NUMELEMENSTV módulo por THREADSPERBLOCK, sea igual a cero.

## Descripción

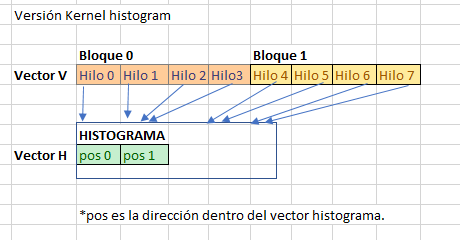
Se adjunta fichero **“histograma.cu”** que contiene toda la programación dividida en las siguientes funciones.

* **Kernels Cuda**

1. El siguiente Kernel realizará el cálculo del histograma pasado, y lo almacenará en el vector del histograma. Esto se realizará mediante operaciones atómicas por cada hilo de cada bloque, directamente a la memoria global de la GPU donde reside el vector del Histograma.

\_\_global\_\_ void histogram(int \*V, int \* H, int numElementsV, int numElementsH)

El siguiente esquema, se muestra como es el funcionamiento suponiendo 2 bloques de ejecución con 4 hilos cada uno, para un histograma de 2 elementos.

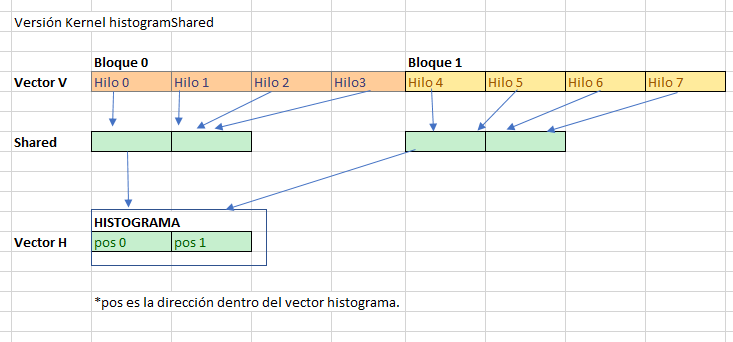


1. **Mejora Kernel 1:** El siguiente Kernel realizará lo mismo que el anterior (Kernel 1) pero con la diferencia de que las operaciones atómicas serán realizadas por los hilos de cada bloque a una memoria compartida **“\_\_shared\_\_”** inicializada a cero por el primer hilo de cada bloque.

Este mismo hilo, será el encargado de actualizar el vector histograma de la memoria global, también con operaciones atómicas, tras una sincronización de todos los hilos del bloque.

\_\_global\_\_ void histogramShared(int \*V, int \* H, int numElementsV, int numElementsH)

El siguiente esquema, se muestra como es el funcionamiento suponiendo 2 bloques de ejecución con 4 hilos cada uno, para un histograma de 2 elementos.



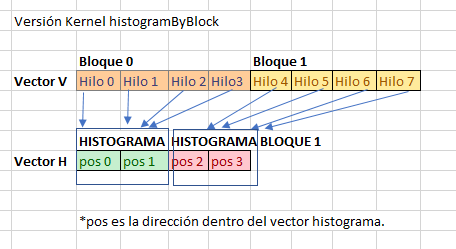
1. El siguiente Kernel, funciona mediante la creación y distribución de un histograma por cada bloque. Esto creará tantos vectores histogramas como bloques necesarios para cubrir todo el vector V, por lo que este Kernel proporciona solo una parte de la implementación, puesto que tras la ejecución de este, se debe de llamar a otro Kernel que se encargará de la suma de todos los Histogramas en solo uno.

Antes de la ejecución de este kernel, el vector Histograma será creado teniendo un tamaño en un único vector, equivalente al número de elementos del histograma multiplicado por el número de bloques que ocupa el vector V.

La idea es que cada hilo de un mismo bloque, pueda actualizar su propio histograma utilizando operaciones atómicas directamente a la memoria global de la GPU donde está almacenado el vector histograma por cada bloque. De esa forma se evita que millones de hilos intenten acceder a pocas porciones de memoria y quede distribuida las operaciones atómicas entre un vector mucho más grande.

\_\_global\_\_ void histogramByBlock(int \*V, int \* H, int numElementsV, int numElementsH)

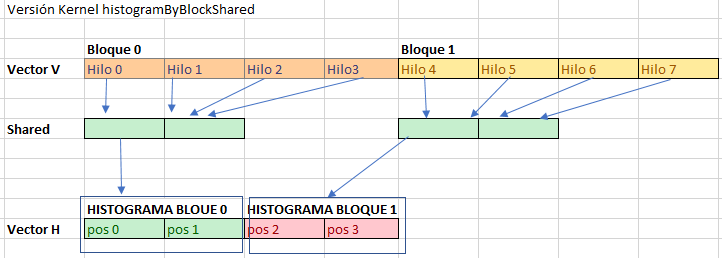
El siguiente esquema, se muestra como es el funcionamiento suponiendo 2 bloques de ejecución con 4 hilos cada uno, para un histograma de 2 elementos.



1. **Mejora Kernel 3:** Este Kernel realizará lo mismo que el anterior, pero utilizará memoria compartida **\_\_shared\_\_**  para cada bloque. Después de forma atómica el hilo 0 de cada bloque, se encargará de actualizar su propio histograma, distribuyendo las operaciones atómicas. Como con el Kernel anterior, se necesita de un segundo kernel encargado de realizar las sumas entre todos los histogramas generados.

\_\_global\_\_ void  histogramByBlockShared(int \*V, int \* H, int numElementsV, int numElementsH)

El siguiente esquema, se muestra como es el funcionamiento suponiendo 2 bloques de ejecución con 4 hilos cada uno, para un histograma de 2 elementos.



1. El siguiente Kernel se encarga de realizar una suma de un vector por el método por reducción. Es el Kernel utilizado tras la ejecución de los Kernels 3 ó 4. Devuelve la suma del vector en las primeras posiciones del mismo vector pasado.

Para el caso concreto de este programa, el vector pasado, será un vector en el que se encontrarán consecutivos y de tamaño **“número elementos histograma”** tantos histogramas como bloques hayan sido necesarios para la ejecución de los kernels 3 ó 4.

Este kernel se ejecuta con número de bloques igual al número de histogramas creados dividido entre 2, puesto que en el método de reducción, con la mitad de los hilos y sabiendo el tamaño del vector, es suficiente.

El número de hilos por bloque es igual al número de elementos de un histograma.

\_\_global\_\_ void sumHistogram(int \* h, int blocksPerGrid)

Por ejemplo, para el caso de un vector Histograma de 4 histogramas de dos elementos cada uno, la operativa sería como se muestra en el siguiente esquema.

