

CAV-EJERCICIOS DE CUDA

Francisco Jesús Díaz Pellejero



9 DE ENERO DE 2023

Contenido

Especificaciones del PC	2
· Ejercicio 2	
Ejercicio 3	
Ejercicio 4	
Ejercicio 6	4
Ejercicio 9	5
Ejercicio 10	9
· Fiercicio 11	

Especificaciones del PC

Esta práctica se ha ejecutado en un ordenador con una NVIDIA GEOFORCE GTX:

```
elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$ nvidia-smi
Mon Jan 9 19:18:08 2023
 NVIDIA-SMI 525.60.13
                     Driver Version: 525.60.13
                                             CUDA Version: 12.0
               Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC
 Fan Temp Perf Pwr:Usage/Cap|
                                Memory-Usage | GPU-Util Compute M.
                                                           MIG M.
0 NVIDIA GeForce ... On | 00000000:01:00.0 Off
                                                             N/A
                 4W / N/A |
           P3
                               10MiB / 4096MiB
                                                   0%
                                                          Default
                                                             N/A
 Processes:
                                                       GPU Memory
 GPU
      GI
           CI
                   PID Type Process name
       ID
                                                        Usage
                  1203
                          G
                              /usr/lib/xorg/Xorg
                                                             4MiB
      N/A N/A
   0
      N/A
                  1973
                           G
                              /usr/lib/xorg/Xorg
                                                             4MiB
          N/A
```

Ejercicio 2

Para este ejercicio únicamente tendremos que cambiar el parámetro N a 600.

a. Verificar si se produce error en cudaMalloc o cudaMemCopy

Dichas llamadas devuelven un código de error si se produce alguno y se pasa a imprimir el resultado correspondiente:

```
// reservamos espacio en la memoria global del device
error = cudaMalloc((void**)&DA, size);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));
error = cudaMalloc((void**)&DB, size);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));
error = cudaMalloc((void**)&DC, size);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));

// copiamos HA y HB del host a DA y DB en el device, respectivamente
error = cudaMemcpy(DA, HA, size, cudaMemcpyHostToDevice);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));
error = cudaMemcpy(DB, HB, size, cudaMemcpyHostToDevice);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));
```

Sin embargo, no se imprime ningún error por pantalla al no reservar en memoria ni copiar en esta una cantidad excesivamente alta de datos (cada vector ocupa solo 600 enteros * 4 bytes = 1800 bytes por vector).

b. Verificar si la llamada al kernel produce error:

El kernel es el siguiente

```
__global__ void VecAdd(int* DA, int* DB, int* DC)
{
    int i = threadIdx.x;
    DC[i] = DA[i] + DB[i];
}
```

No hay ningún tipo de error al llamar a dicha función

c. En caso de error, que se utilice cudagetErrorString():

Al no producirse ningún error no se llama a dicha función. Pero esta esta presente en todas las llamadas a cudaMalloc y cudaMemcpy:

```
// copiamos HA y HB del host a DA y DB en el device, respectivamente
error = cudaMemcpy(DA, HA, size, cudaMemcpyHostToDevice);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));
error = cudaMemcpy(DB, HB, size, cudaMemcpyHostToDevice);
if (error != cudaSuccess) printf("%s\n", cudaGetErrorString(error));
```

d. ¿Has detectado algún error durante la ejecución?

Para comprobar la correcta salida del programa se comprueba que la suma de los vectores se ha realizado de forma correcta:

Sin embargo, la ejecución del programa no da ningún resultado erróneo ya que no se imprime ningún mensaje de error

```
elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$ ./suma-vectores1b elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$
```

Ejercicio 3

Para utilizar un 1 bloque de 1 hebra, los parámetros que se pasan al kernel tendrá que ser 1 y 1 (Número de bloques, hilos por bloque).

Sin embargo, el kernel se tendrá que modificar para que ese solo hilo procese todos los elementos de los vectores:

```
__global__ void VecAdd(int* DA, int* DB, int* DC)
{
            for(int i = 0; i < N; i++)
            DC[i] = DA[i] + DB[i];
}
```

El resultado del programa sigue siendo correcto:

```
elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$ ./suma-vectores2
elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$
```

Ejercicio 4

Si queremos lanzar N bloques de una hebra, tendremos que llamar al kernel con los parámetros N y 1:

```
VecAdd <<<N, 1>>>(DA, DB, DC);
```

Sin embargo, el kernel del ejercicio anterior no funcionaría ya que se tendría que volver a procesar un elemento por hilo, al haber N hilos funcionando (N bloques de 1 hilo):

```
__global__ void VecAdd(int* DA, int* DB, int* DC)
{
    int i = blockIdx.x;
    DC[i] = DA[i] + DB[i];
}
```

Este es el único cambio por realizar con respecto al anterior código. El programa sigue funcionando correctamente:

```
elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$ ./suma-vectores3 elflautista@manuel-Nitro-AN515-55:~/Chechu/CAV-CUDA/exec$
```

Ejercicio 6

El código anterior no se podría usar ya que el parámetro N esta a 600 y no se pueden procesar suficientes elementos. Si ponemos N a 10000 el programa funcionaría pero solo tendríamos 100000 bloques de 1 elemento, resultando en un programa ineficiente.

Otra opción sería tener menos bloques de mayor tamaño, por ejemplo, un tamaño de bloque de 500 hilos, esto está representado por la constante tb:

```
#define N 100000
#define tb 500
```

Para calcular el número de bloques necesarios utilizamos la fórmula de la sesión 1, y realizamos la llamada al kernel:

```
int blocks = (N+tb-1)/tb;

VecAdd <<<blocks, tb>>>(DA, DB, DC);
```

Sin embargo, como los bloques tienen más de un hilo, la indexación de los elementos se tendrá que hacer también por bloque, por lo que habrá que modificar el kernel:

```
__global__ void VecAdd(int* DA, int* DB, int* DC)
{
    int i = threadIdx.x + blockIdx.x*blockDim.x;
    DC[i] = DA[i] + DB[i];
}
```

Ejercicio 9

a. ¿Hay coalescencia de lectura o escritura?

El código del kernel que se nos proporciona en la práctica es el siguiente:

```
__global__ void trspta1(int *dev_a, int *dev_b, int filas, int cols)

{ int ix = blockIdx.x * blockDim.x + threadIdx.x;
    int iy = blockIdx.y * blockDim.y + threadIdx.y;

    if ((ix<cols)&&(iy<filas))
    { dev_b[ix*filas+iy] = dev_a[iy*cols+ix];} // B[ix][iy] = A[iy][ix] (A matriz representada por dev_a, etc.)
}</pre>
```

En este caso se ha hecho un mapeo (x, y) en el que la 'x' hace referencia al eje horizontal y el 'y' al vertical. Por ejemplo, si tenemos parte de una matriz dev_a:

En el kernel se dipone de manera lineal de la siguiente forma:

```
0 1 2 3 4 5 6 7..... 8 9 10 11 12 13 14 15.....
```

Si estamos en un bloque de 8x8, el hilo con identificador ix = 0 e iy = 0 leerá el elemento 0*cols + 0 = dev[0] de la matriz. El hilo [0,1] leerá el elemento dev[1], el [0,2] el dev[2]..... el [1,0] el dev[8], el [1,1] el dev[9]. Por lo tanto, cada se leerán elementos contiguos de 8 en 8. Del hilo [0,0] al [0,7] se irá a memoria una vez y se aprovechará un cuarto del warp (32 hilos). Del [1,0] al [1,7], se realizará otro acceso y así sucesivamente.

Por lo tanto, tendremos coalescencia de hilos de 8 en 8 direcciones. Si bien no se aprovecha todo el warp, en el caso de querer usar bloques de 8x8 es la mejor opción que tenemos.

Sin embargo, no hay coalescencia en la escritura ya que, por ejemplo, el hilo [0,0] escribiría en la dirección dev[0], el hilo [0,1] escribiría en la dirección 1*filas + 0 = dev[filas], el hilo [0,2]

escribiría en dev[2*filas]. Dichas direcciones no son contiguas, ni lo serán para ninguno de los hilos del bloque.

b. Modificar código para que aproveche coalescencia

Se han propuesto 2 soluciones:

```
__global__ void trspta1(int *dev_a, int *dev_b, int filas, int cols)
{ int ix = blockIdx.x * blockDim.x + threadIdx.x;
    int iy = blockIdx.y * blockDim.y + threadIdx.y;
    __shared__ int s[8][8];

if ((ix<cols)&&(iy<filas))
{ s[threadIdx.y][threadIdx.x] = dev_a[iy*cols+ix];}

ix = blockIdx.y * blockDim.y + threadIdx.x;
    iy = blockIdx.x * blockDim.x + threadIdx.y;

if ((ix<filas)&&(iy<cols)){
        dev_b[iy*filas+ix] = s[threadIdx.x][threadIdx.y];
    }
}</pre>
```

Se puede utilizar memoria compartida del mismo tamaño que el número de hilos del bloque, de tal manera que cada hilo procese solo 1 elemento.

En el primer if se comprueba que el hilo tiene las dimensiones adecuadas y se guardan los elementos de la submatriz de dev_a en la memoria compartida. Por ejemplo, para el hilo [0,0] dentro del bloque 0 tendrá la siguiente memoria compartida con el resto de hilos de su bloque:

```
0 1 2 3 4 5 6 7

43 44 45 46 47 48 49 50

86 87 88 89 90 91 92 93

129 130 131 132 133 134 135 136

172 173 174 175 176 177 178 179

215 216 217 218 219 220 221 222

258 259 260 261 262 263 264 265

301 302 303 304 305 306 307 308
```

Después, se recalculará la posición del hilo para adecuarse a la posición a escribir en dev_b de acuerdo a la matriz traspuesta. Ahora ix indexara al bloque en la dirección 'y', y viceversa.

```
ix = blockIdx.y * blockDim.y + threadIdx.x;
iy = blockIdx.x * blockDim.x + threadIdx.y;
```

En el siguiente if las comprobaciones se harán a la inversa ya que ix e iy son inversos (la posición a escribir en dev_b) y se guarda el resultado. El hilo [0,0] del bloque (0,0) cogerá el número 0 de 's' y lo guardará en dev[0] el hilo [0,1] cogerá el elemento 43 de 's' y lo guardará en la posición 0*filas + 1 = dev[1], el hilo[0,1] guardará el resultado en dev[2], y así hasta el

hilo[0,7]. Por lo tanto, al igual que con las lecturas, hay coalescencia de 8 en 8 direcciones. No se aprovecha todo el warp, pero es la mejor opción si quieremos bloques de 8x8.

La otra solución aportada en el transp1b-coalesc.cu tiene el siguiente kernel:

```
_global__ void trspta1(int *dev_a, int *dev_b, int filas, int cols){

int ix = blockIdx.x * 8 + threadIdx.x;
int iy = blockIdx.y * 32 + threadIdx.y;

_shared__ int s[32][8];

for(int j = 0; j < 32; j+=8)
    if ((ix < cols) && ((iy+j) < filas))
        s[threadIdx.y + j][threadIdx.x] = dev_a[(iy+j)*cols+ix];

ix = blockIdx.y * 32 + threadIdx.x;
iy = blockIdx.x * 8 + threadIdx.y;

for(int j = 0; j < 32; j+=8){
    if (((ix+j) < filas) && (iy < cols)){
        dev_b[iy*filas+(ix+j)] = s[threadIdx.x + j][threadIdx.y];
    }
}</pre>
```

Aquí se utilizan bloques de 8x8 pero se utilizan 'tiles' de 32x8, por lo que cada hilo tendrá que procesar 4 elementos. De ahí el uso de 2 bucles. La memoria compartida pasará a tener 32 filas en lugar de 8, por ejemplo, para el hilo [0,0] del bloque (0,0):

```
0 1 2 3 4 5 6 7
43 44 45 46 47 48 49 50
86 87 88 89 90 91 92 93
129 130 131 132 133 134 135 136
172 173 174 175 176 177 178 179
215 216 217 218 219 220 221 222
258 259 260 261 262 263 264 265
301 302 303 304 305 306 307 308
344 345 346 347 348 349 350 351
387 388 389 390 391 392 393 394
430 431 432 433 434 435 436 437
473 474 475 476 477 478 479 480
516 517 518 519 520 521 522 523
559 560 561 562 563 564 565 566
602 603 604 605 606 607 608 609
645 646 647 648 649 650 651 652
688 689 690 691 692 693 694 695
731 732 733 734 735 736 737 738
774 775 776 777 778 779 780 781
817 818 819 820 821 822 823 824
860 861 862 863 864 865 866 867
903 904 905 906 907 908 909 910
946 947 948 949 950 951 952 953
989 990 991 992 993 994 995 996
1032 1033 1034 1035 1036 1037 1038 1039
```

En el primer bucle, el hilo [0,0] leerá los elementos [0,0] (iteración 1), [8,0] (iteración 2), [16,0] (iteración 3)... el hilo [0,1] leerá [0,1], [8,1], [16, 1] y así sucesivamente. Por lo tanto, en cada iteración seguirá habiendo coalescencia en 8 direcciones.

Se recalculan la posición de los índices para la traspuesta y en la escritura el hilo [0,0] escribirá los elementos [0,0] (iteración 1), [8,0] (iteración 2), [16,0] (iteración 3) de la memoria compartida en dev[0], dev[8*filas], dev[16*filas]...etc, y así sucesivamente. Seguirá habiendo coalescencia de escritura de 8 en 8 direcciones.

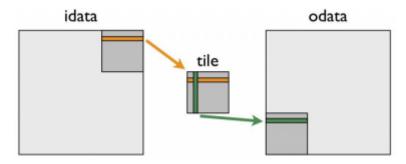
Para que la segunda solución funcione se tiene que calcular ix e iy en función del tamaño del 'tile' que estemos usando (32 x 8), por lo que:

```
int ix = blockIdx.x * 8 + threadIdx.x;
int iy = blockIdx.y * 32 + threadIdx.y;
__shared__ int s[32][8];
```

Además, el número de bloques cambia en función del tamaño del tile de la siguiente manera (se usa la fórmula de la sesión 1). El techo ya no está definido por las dimensiones del bloque (8x8) sino por la dimensión del 'tile':

```
int numBlf = (F+32-1)/32;
int numBlc = (C+8-1)/8;
```

Esta segunda solución sigue dicha abstracción:



Solo que el tile pasaría a ser recutangular, e idata y odata (dev_a y dev_b) no serían matrices cuadradas.

En la segunda solución habría únicamente un grid de 1x6 bloques debido a que se depende del tile.

Este ejercicio se ha realizado con la restricción de no cambiar el tamaño del bloque y dejarlo fijo a 8x8. Otra opción hubiese sido cambiar dicho bloque a uno más amplio a 32*8 y haber usado un 'tile' de 32*32. Sin embargo, para una matriz tan pequeña como la del enunciado, tendríamos únicamente un bloque de 32*8 hilos. Por lo que la decisión de utilizar determinados tamaños de bloques y de tiles dependen del tamaño de entrada y del criterio del programador.

Ejercicio 10

Es necesario la directiva de sincronización __syncthreads() para que el programa del ejercicio 9 funcione correctamente. El resultado de ejecución del ejercicio 9 es el siguiente:

Como se puede ver en la imagen, hay elementos que no se han colocado bien. Se debe colocar la directiva mencionada anteriormente tras la lectura de un elemento de la matriz dev_a:

```
__global__ void trspta1(int *dev_a, int *dev_b, int filas, int cols)
{    int ix = blockIdx.x * blockDim.x + threadIdx.x;
    int iy = blockIdx.y * blockDim.y + threadIdx.y;
    __shared__ int s[8][8];

if ((ix<cols)&&(iy<filas))
{    s[threadIdx.y][threadIdx.x] = dev_a[iy*cols+ix];}

__syncthreads();
    ix = blockIdx.y * blockDim.y + threadIdx.x;
    iy = blockIdx.x * blockDim.x + threadIdx.y;

if ((ix<filas)&&(iy<cols)){
        dev_b[iy*filas+ix] = s[threadIdx.x][threadIdx.y];
    }
}</pre>
```

Dicha llamada actuará como barrera, y hasta que todos los hilos no lleguen a dicha llamada ninguno podrá continuar. De esta manera, ningún hilo escribe en dev_b un elemento de la memoria compartida que todavía no ha sido completado por otro hilo que corresponda. De esta manera se consigue el resultado deseado:

Ya no tenemos ningún error.

Ejercicio 11

Si que habría conflicto. Para un tile de 8 x 8, todos los elementos de una columna se mapean al mismo banco de memoria. Tendríamos un peor caso en el que al leer una columna se produzcan 8 conflictos: 8 por columna que contiene cada 'tile'. La solución a esto es rellenar la memoria compartida:

```
__shared__ int s[8][8 + 1];
```

El resultado de ejecución sigue siendo el mismo que en el ejercicio anterior.