

Práctica 10. Suma de matrices N x N

$$\begin{bmatrix} -1 & 2 & 3 \\ 0 & 1 & 5 \\ 7 & 9 & 10 \end{bmatrix}_{3 \times 3} + \begin{bmatrix} 2 & 1 & -4 \\ 5 & 6 & 8 \\ 4 & -2 & 3 \end{bmatrix}_{3 \times 3} = \begin{bmatrix} 1 & 3 & -1 \\ 5 & 7 & 13 \\ 11 & 7 & 13 \end{bmatrix}_{3 \times 3}$$

U.A.Q. Fac. de Informática

Dra. Sandra Luz Canchola Magdaleno

Correo: sandra.canchola@uaq.mx

Dra. Reyna Moreno Beltrán

Correo: reyna.moreno@uaq.mx



Sean A y B matrices de dimensiones $n \times n$, la suma está definida como:

$$(A + B)_{i,j} = a_{i,j} + b_{i,j}$$

$$A_{n \times n} + B_{n \times n} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \dots & a_{nn} + b_{nn} \end{bmatrix}$$

Cálculo de suma de matrices $n \times n$

A

	0	1	2	...	n-1
0					
1					
2					
...					
n-1					

B

	0	1	2	...	n-1
0					
1					
2					
...					
n-1					

A+B

	0	1	2	...	n-1
0					
1					
2					
...					
n-1					

Elementos de matrices

Matriz 5 x 2

i	j	
	0	1
0		
1		
2		
3		✓
4		

Índice de elementos

`matriz[i][j]`

Ejemplo:

`mat1[3][1]`

Matriz 5 x 2 como apuntador a una memoria consecutiva de 10 elementos

0	1	2	3	4	5	6	7	8	9
							✓		

Índice de elementos

$\text{indice} = (i * \text{numCol}) + j$

Ejemplo:

`mat1[3][1]`

$\text{indice} = (3 * 2) + 1 = 7$

Elementos de matrices

Matriz 5 x 2 x 4

		j							
		0				1			
i	k=0	1	2	3	0	1	2	3	
	0								
	0	1	2	3	0	1	2	3	
	1						✓		
	0	1	2	3	0	1	2	3	
	2								
	0	1	2	3	0	1	2	3	
3									
0	1	2	3	0	1	2	3		
4									

Índice de elementos

`matriz[i][j][k]`

Ejemplo:

`mat2[1][1][2]`

Elementos de matrices

Matriz 5 x 2 x 4 como apuntador a una memoria consecutiva de 40 elementos

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...	36	37	38	39
														✓		...				

Índice de elementos

$$\text{indice} = (i * \text{numCol} * \text{numProf}) + (j * \text{numProf}) + k$$

Ejemplo:

`mat2[1][1][2]`

$$\text{indice} = (1 * 2 * 4) + (1 * 4) + 2 = 14$$

Proyecto CUDA



Create a new project

Choose a project template with code scaffolding to get started



CUDA 11.7 Runtime

A project that uses the CUDA 11.7 runtime

C++

CUDA

Windows

Linux

Cloud

Console

DataScience

Desktop

Machine Learning



CUDA 12.1 Runtime

A project that uses the CUDA 12.1 runtime

C++

CUDA

Windows

Linux

Cloud

Console

DataScience

Desktop

Machine Learning

Proyecto CUDA

Configure your new project

CUDA 12.1 Runtime

C++

CUDA

Windows

Linux

Cloud

Console

DataScience

Desktop

Machine Learning

Project name

Prog10_SumaMatNxN

Location

C:\TrabajoLaboratorio\CUDATopico2\Projects\Verano2024

Solution name ⓘ

Prog10_SumaMatNxN

☐

Place solution and project in the same directory

Project will be created in "C:\TrabajoLaboratorio\CUDATopico2\Projects\Verano2024
\Prog10_SumaMatNxN\Prog10_SumaMatNxN\"

Back

Create

Operaciones de memoria (CPU)

- **malloc.**- Reserva un bloque de memoria de un tamaño definido de bytes, retornando un apuntador al inicio de dicho bloque. El contenido de dicho bloque no se inicializa por lo que es indeterminado. Ejemplo:

```
void* malloc (size_t size);  
buffer = (char*) malloc (sizeof(char)*100);
```

- **memset.**- asigna valores en secciones de memoria. Ejemplo:
Memset(variable, valor_a_asignar, tamaño_de_memoria)
Donde: tamaño_de_memoria se define como n * sizeof(tipo)

- **memcpy.**- Copia el contenido de un bloque de memoria referenciado por un apuntador a otro apuntador. Ejemplo:

```
void* memcpy( void* dest, const void* src, std::size_t count );  
memcpy (ptrDest, ptrOrigen, sizeof(int)*100);
```

- **free.**- Liberar la memoria reservada con el comando malloc. Ejemplo:
free(pointerName);
free(array2);

Operaciones de memoria (GPU)

- **cudaMalloc**.- asigna una sección de memoria en GPU de acuerdo con el espacio solicitado.

Ejemplo:

```
cudaMalloc((void**) &apuntador, tamaño_de_memoria)
```

Donde: tamaño_de_memoria se define como $n * \text{sizeof}(\text{tipo})$

- **cudaMemset**.- asigna valores en secciones de memoria.

Ejemplo:

```
Memset(apuntador, valor_a_asignar, tamaño_de_memoria)
```

Donde: tamaño_de_memoria se define como $n * \text{sizeof}(\text{tipo})$

- **cudaMemcpy**.- copia memoria hacia y desde el device.

Ejemplo:

```
cudaMemcpy(destino, origen, tamaño_de_memoria, indicador_flujo_de_inf)
```

Donde Indicador= cudaMemcpyHostToDevice, cudaMemcpyDeviceToHost, cudaMemcpyDeviceToDevice

- **cudaFree**.- libera la memoria reservada por un apuntador.

Ejemplo:

```
cudaFree (apuntador)
```

Memoria

CPU (Host)

A01	width	50									
A05	epsilon	0.000001									
A07	maxN	16									
A10	maxM	20									
A15	A	0	1	2	3	4	5	6	7	...	(nxn)-1
		α	γ	ϕ	φ	η	χ	λ	ε	...	τ
B02	B	0	1	2	3	4	5	6	7	...	(nxn)-1
		τ	β	κ	θ	π	δ	ε	υ	...	η
B45	C	0	1	2	3	4	5	6	7	...	(nxn)-1
		α + τ	γ + β	ϕ + κ	φ + θ	η + π	χ + δ	λ + ε	ε + υ	...	τ + η
C30	C_host	0	1	2	3	4	5	6	7	...	(nxn)-1
		α + τ	γ + β	ϕ + κ	φ + θ	η + π	χ + δ	λ + ε	ε + υ	...	τ + η
E10	dev_A	J10									
F20	dev_B	J45									
G05	dev_C	J90									
H16											
H20											

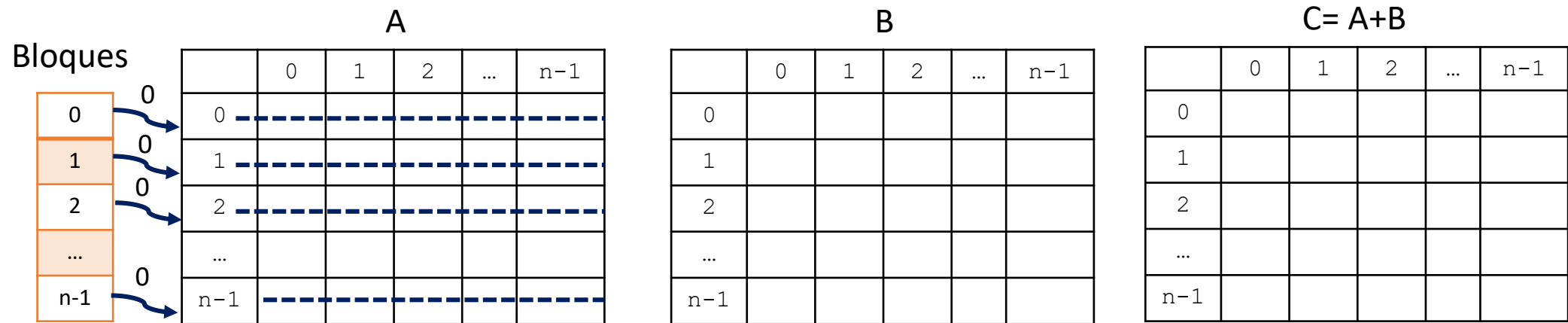
GPU (Device)

J01											
J05											
J10		0	1	2	3	4	5	6	7	...	(nxn)-1
		α	γ	ϕ	φ	η	χ	λ	ε	...	τ
J45		0	1	2	3	4	5	6	7	...	(nxn)-1
		τ	β	κ	θ	π	δ	ε	υ	...	η
J90		0	1	2	3	4	5	6	7	...	(nxn)-1
		α + τ	γ + β	ϕ + κ	φ + θ	η + π	χ + δ	λ + ε	ε + υ	...	τ + η
K01											
K05											
K10											
K15											
K20											
K30											
L07											
L10											

Memoria reservada

Apuntador

Caso 1. N bloques con hilo único.
Cada hilo calcula el resultado de una fila completa.



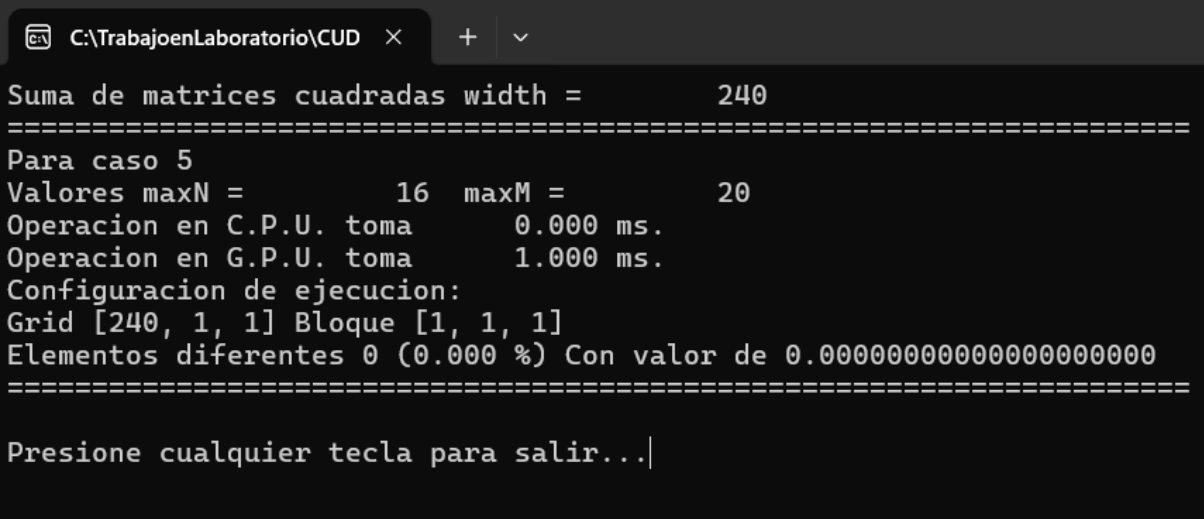
$tid = blockIdx.x$
 $primerElem = (blockIdx.x * n)$

blockIdx.x	threadIdx.x	tid	Elementos atendidos	
0	0	0	$C_{0,0}, C_{0,1}, C_{0,2} \dots C_{0,n-1}$	$0, 1, 2, \dots (n-1)$
1	0	1	$C_{1,0}, C_{1,1}, C_{1,2} \dots C_{1,n-1}$	$n, n+1, n+2, \dots, 2n-1$
2	0	2	$C_{2,0}, C_{2,1}, C_{2,2} \dots C_{2,n-1}$	$2n, 2n+1, 2n+2, \dots, 3n-1$
...
n-1	0	n-1	$C_{n-1,0}, C_{n-1,1}, C_{n-1,2} \dots C_{n-1,n-1}$	$n^2 - n, n^2 - n + 1, n^2 - n + 2, \dots, n^2 - 1$

Caso 1. N bloques con hilo único.

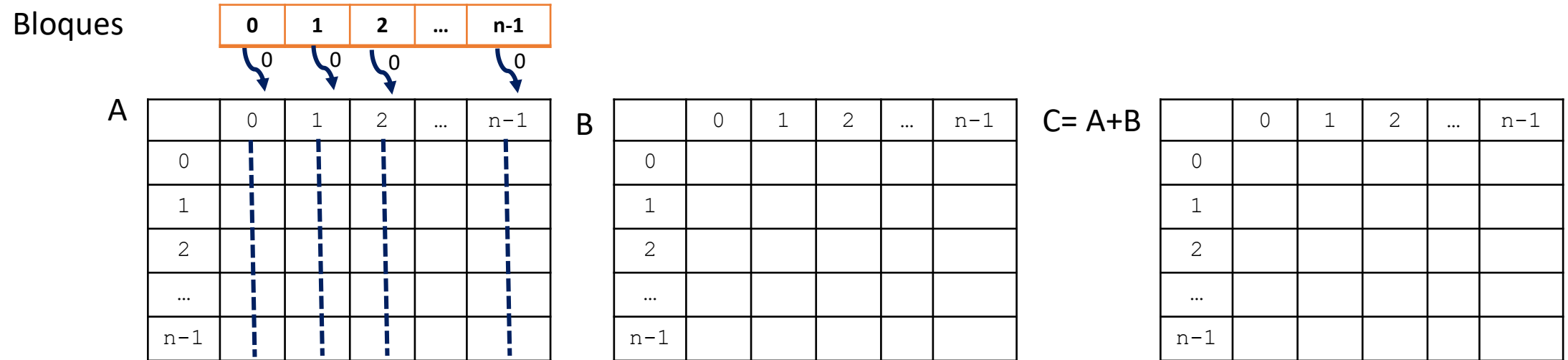
Cada hilo calcula el resultado de una fila completa.

```
#define width 240
#define epsilon float(0.00000001)
...
dim3 dimGrid(width);
dim3 dimBlock(1);
...
int fila = blockIdx.x;
int tid = fila * width; //primer elemento de la fila
for (int i = 0; i < width; i++){
    c[tid + i] = a[tid + i] + b[tid + i];
}
```



```
C:\TrabajoLaboratorio\CUD x + v
Suma de matrices cuadradas width = 240
=====
Para caso 5
Valores maxN = 16 maxM = 20
Operacion en C.P.U. toma 0.000 ms.
Operacion en G.P.U. toma 1.000 ms.
Configuracion de ejecucion:
Grid [240, 1, 1] Bloque [1, 1, 1]
Elementos diferentes 0 (0.000 %) Con valor de 0.00000000000000000000
=====
Presione cualquier tecla para salir...|
```

Caso 2. N bloques con hilo único.
Cada hilo calcula el resultado de una columna completa.

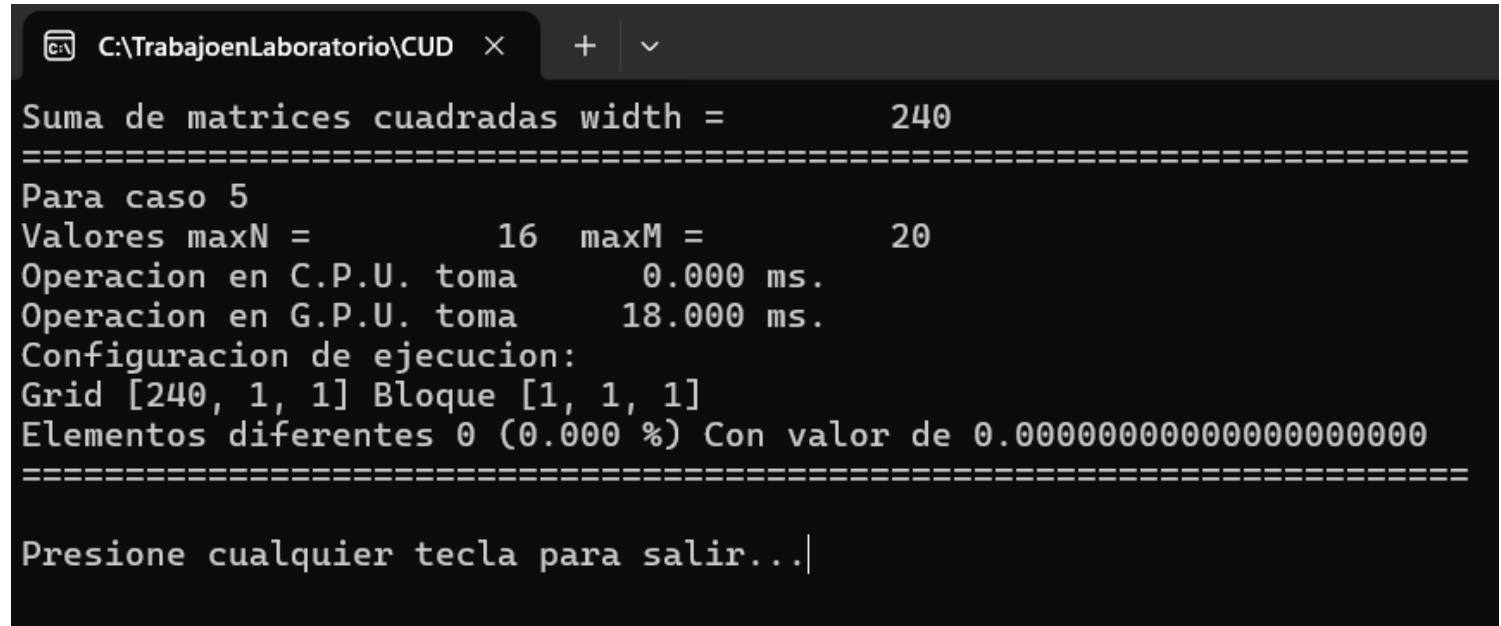


tid=blockIdx.x
primerElem=blockIdx.x

blockIdx.x	threadIdx.x	tid	Elementos atendidos	
0	0	0	$C_{0,0}, C_{1,0}, C_{2,0} \dots C_{n-1,0}$	$0, n, 2n, \dots, (n-1)n$
1	0	1	$C_{0,1}, C_{1,1}, C_{2,1} \dots C_{n-1,1}$	$1, n+1, 2n+1, \dots, (n-1)n+1$
2	0	2	$C_{0,2}, C_{1,2}, C_{2,2} \dots C_{n-1,2}$	$2, n+2, 2n+2, \dots, (n-1)n+2$
...
n-1	0	n-1	$C_{0,n-1}, C_{1,n-1}, C_{2,n-1} \dots C_{n-1,n-1}$	$n-1, 2n-1, \dots, (n \times n) - 1$

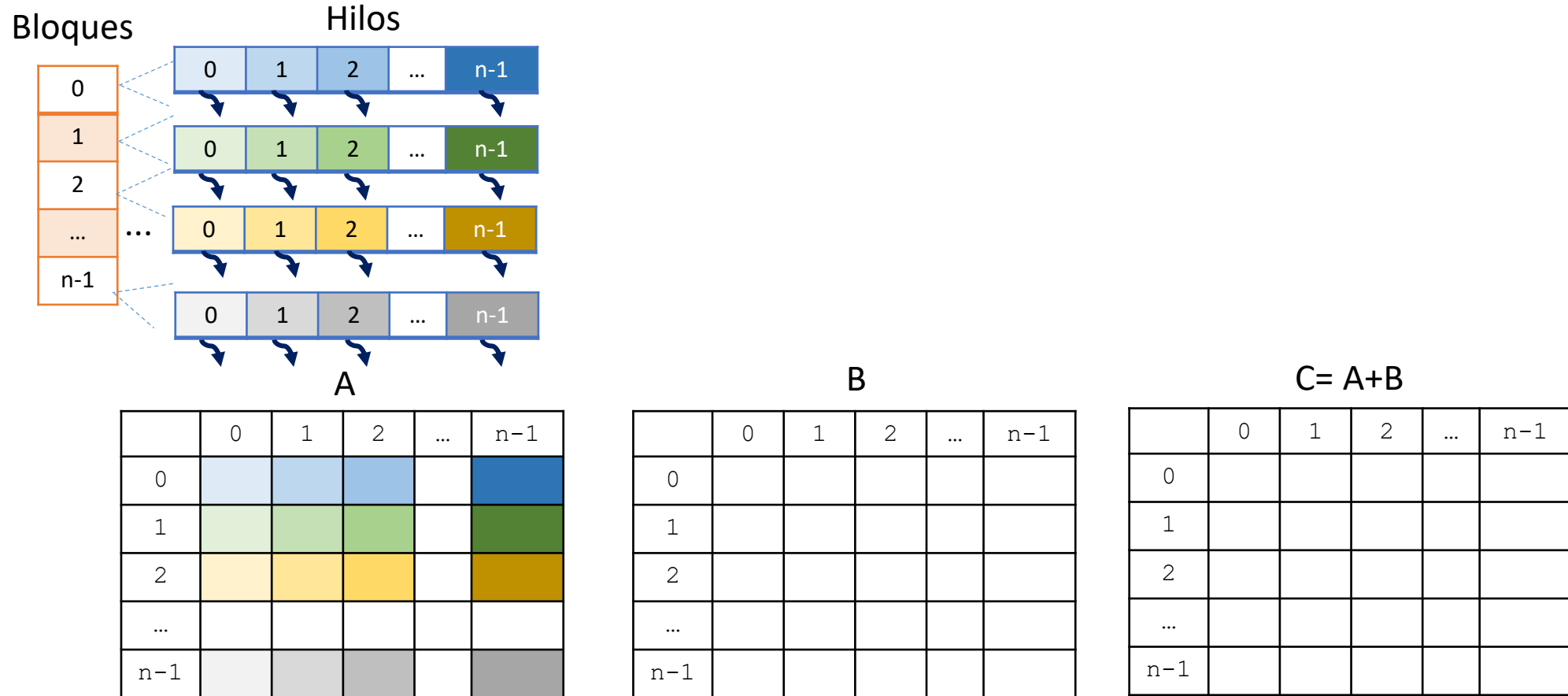
Caso 2. N bloques con hilo único.
Cada hilo calcula el resultado de una columna completa.

```
#define width 240
#define epsilon float(0.0000001)
...
dim3 dimGrid(width);
dim3 dimBlock(1);
...
int tid = blockIdx.x;
for (int i = 0; i < width; i++){
    c[tid+ (i * width)] = a[tid + (i * width)] + b[tid + (i * width)];
}
```

A screenshot of a terminal window titled 'C:\TrabajoLaboratorio\CUD'. The terminal displays the output of a CUDA program. The output includes a title 'Suma de matrices cuadradas width = 240', a separator line of equals signs, and then 'Para caso 5'. It shows 'Valores maxN = 16' and 'maxM = 20'. It reports 'Operacion en C.P.U. toma 0.000 ms.' and 'Operacion en G.P.U. toma 18.000 ms.'. It shows the 'Configuracion de ejecucion:' as 'Grid [240, 1, 1] Bloque [1, 1, 1]' and 'Elementos diferentes 0 (0.000 %) Con valor de 0.0000000000000000000000'. Another separator line of equals signs follows. The prompt 'Presione cualquier tecla para salir...|' is at the bottom.

```
C:\TrabajoLaboratorio\CUD x + v
Suma de matrices cuadradas width = 240
=====
Para caso 5
Valores maxN = 16 maxM = 20
Operacion en C.P.U. toma 0.000 ms.
Operacion en G.P.U. toma 18.000 ms.
Configuracion de ejecucion:
Grid [240, 1, 1] Bloque [1, 1, 1]
Elementos diferentes 0 (0.000 %) Con valor de 0.0000000000000000000000
=====
Presione cualquier tecla para salir...|
```

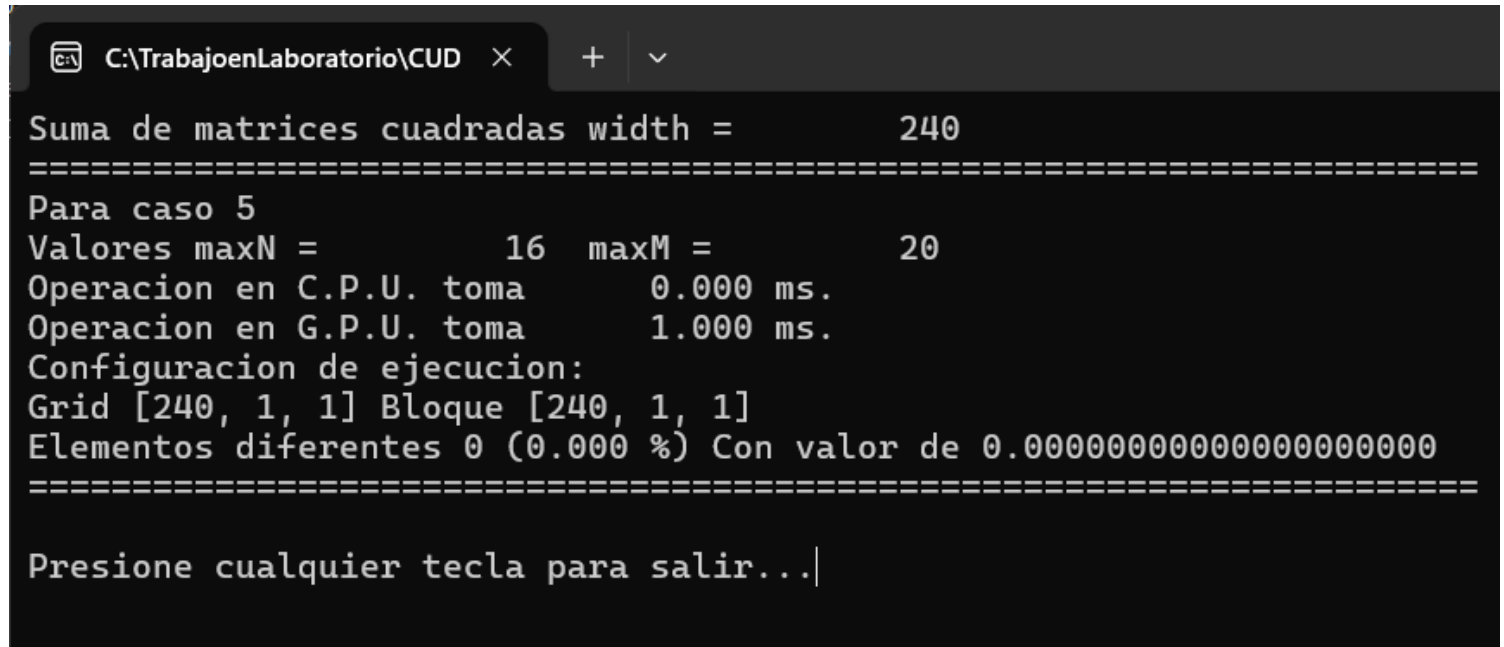
Caso 3. N bloques con N hilos cada uno.



$$tid = (blockIdx.x * blockDim.x) + threadIdx.x$$

Caso 3. N bloques con N hilos cada uno.

```
#define width 240
#define epsilon float(0.0000001)
...
dim3 dimGrid(width); // por cada renglon
dim3 dimBlock(width); // por cada columna
...
int tid = (blockIdx.x*blockDim.x)+threadIdx.x;
c[tid] = a[tid] + b[tid];
```

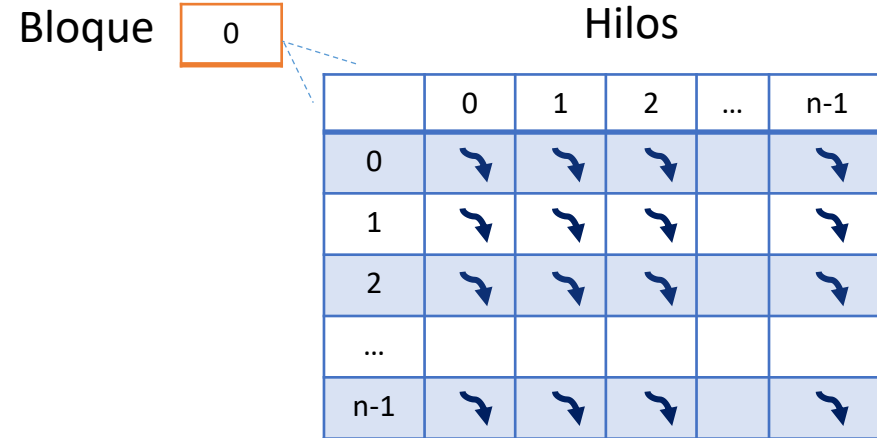


```
C:\TrabajoLaboratorio\CUD x + v

Suma de matrices cuadradas width = 240
=====
Para caso 5
Valores maxN = 16 maxM = 20
Operacion en C.P.U. toma 0.000 ms.
Operacion en G.P.U. toma 1.000 ms.
Configuracion de ejecucion:
Grid [240, 1, 1] Bloque [240, 1, 1]
Elementos diferentes 0 (0.000 %) Con valor de 0.000000000000000000000000
=====

Presione cualquier tecla para salir...|
```

Caso 4. Un bloque con NxN hilos.



A

	0	1	2	...	n-1
0					
1					
2					
...					
n-1					

B

	0	1	2	...	n-1
0					
1					
2					
...					
n-1					

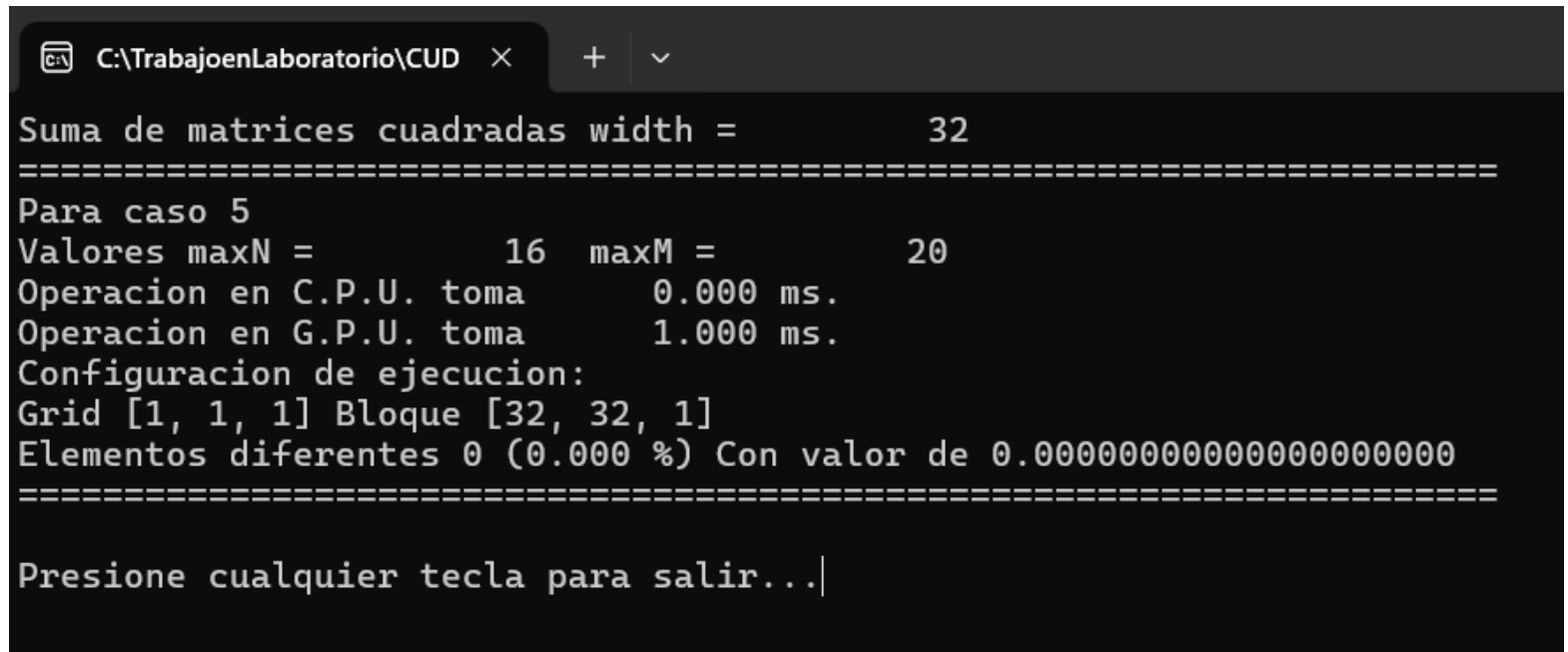
C= A+B

	0	1	2	...	n-1
0					
1					
2					
...					
n-1					

$tid = (threadIdx.x * blockDim.y) + threadIdx.y$

Caso 4. Un bloque con NxN hilos.













```
#define width 32
#define epsilon float(0.0000001)
...
dim3 dimGrid(1);
dim3 dimBlock(width, width); // por cada elemento
...
int tid = (threadIdx.x*blockDim.y) + threadIdx.y;
c[tid] = a[tid] + b[tid];
```



```
C:\TrabajoLaboratorio\CUD x + v
Suma de matrices cuadradas width = 32
=====
Para caso 5
Valores maxN = 16 maxM = 20
Operacion en C.P.U. toma 0.000 ms.
Operacion en G.P.U. toma 1.000 ms.
Configuracion de ejecucion:
Grid [1, 1, 1] Bloque [32, 32, 1]
Elementos diferentes 0 (0.000 %) Con valor de 0.00000000000000000000
=====
Presione cualquier tecla para salir...|
```

Caso 5. Bloque 2D con hilos 2D

Hilos

	0	1	2
0			
1			
2			
3			

Tamaño de la matriz
10 x 10

maxN=4, maxM=3

Cada bloque tiene $\text{maxN} \times \text{maxM}$ hilos.

Dimensión del grid

Primera dimensión

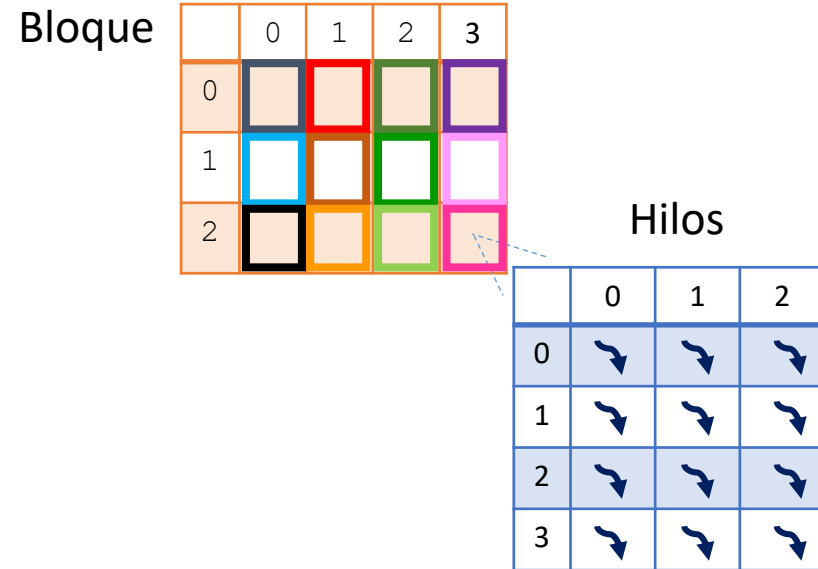
$$N / \max N = 10 / 4 = 2.5 \approx 3$$

Segunda dimensión

$$N / \max M = 10 / 3 = 3.3333 \approx 4$$

[illegible]

Caso 5. Bloque 2D con hilos 2D



maxN=4, maxM=3

Cada bloque tiene maxN x maxM hilos.

```
filaInicialBloque=(blockIdx.x*blockDim.x)
fila=filaInicialBloque+threadIdx.x
columnaIniciaBloque=(blockIdx.y*blockDim.y)
columna=columnaIniciaBloque+threadIdx.y
```

A

	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	10	11	12	13	14	15	16	17	18	19
2	20	21	22	23	24	25	26	27	28	29
3	30	31	32	33	34	35	36	37	38	39
4	40	41	42	43	44	45	46	47	48	49
5	50	51	52	53	54	55	56	57	58	59
6	60	61	62	63	64	65	66	67	68	69
7	70	71	72	73	74	75	76	77	78	79
8	80	81	82	83	84	85	86	87	88	89
9	90	91	92	93	94	95	96	97	98	99

blockI dx		thread Idx		Elemento		
x	y	x	y	fila	col	#
0	0	0	0	0	0	0
0	0	0	1	0	1	1
0	0	0	2	0	2	2
0	0	1	0	1	0	10
0	0	1	1	1	1	11
0	0	1	2	1	2	12
0	0	2	0	2	0	20
0	0	2	1	2	1	21
0	0	2	2	2	2	22
0	0	3	0	3	0	30
0	0	3	1	3	1	31
0	0	3	2	3	2	32
0	1	0	0	0	3	3
0	1	0	1	0	4	4
0	1	0	2	0	5	5
0	1	1	0	1	3	13
0	1	1	1	1	4	14
0	1	1	2	1	5	15
0	1	2	0	2	3	23
0	1	2	1	2	4	24
0	1	2	2	2	5	25
0	1	3	0	3	3	33
0	1	3	1	3	4	34
0	1	3	2	3	5	35

blockI dx		thread Idx		Elemento		
x	y	x	y	fila	col	#
0	2	0	0	0	6	6
0	2	0	1	0	7	7
0	2	0	2	0	8	8
0	2	1	0	1	6	16
0	2	1	1	1	7	17
0	2	1	2	1	8	18
0	2	2	0	2	6	26
0	2	2	1	2	7	27
0	2	2	2	2	8	28
0	2	3	0	3	6	36
0	2	3	1	3	7	37
0	2	3	2	3	8	38
0	3	0	0	0	9	9
0	3	0	1	0	10	10
0	3	0	2	0	11	11
0	3	1	0	1	9	19
0	3	1	1	1	10	20
0	3	1	2	1	11	21
0	3	2	0	2	9	29
0	3	2	1	2	10	30
0	3	2	2	2	11	31
0	3	3	0	3	9	39
0	3	3	1	3	10	40
0	3	3	2	3	11	41

blockI dx		thread Idx		Elemento		
x	y	x	y	fila	col	#
1	0	0	0	4	0	40
1	0	0	1	4	1	41
1	0	0	2	4	2	42
1	0	1	0	5	0	50
1	0	1	1	5	1	51
1	0	1	2	5	2	52
1	0	2	0	6	0	60
1	0	2	1	6	1	61
1	0	2	2	6	2	62
1	0	3	0	7	0	70
1	0	3	1	7	1	71
1	0	3	2	7	2	72
1	1	0	0	4	3	43
1	1	0	1	4	4	44
1	1	0	2	4	5	45
1	1	1	0	5	3	53
1	1	1	1	5	4	54
1	1	1	2	5	5	55
1	1	2	0	6	3	63
1	1	2	1	6	4	64
1	1	2	2	6	5	65
1	1	3	0	7	3	73
1	1	3	1	7	4	74
1	1	3	2	7	5	75

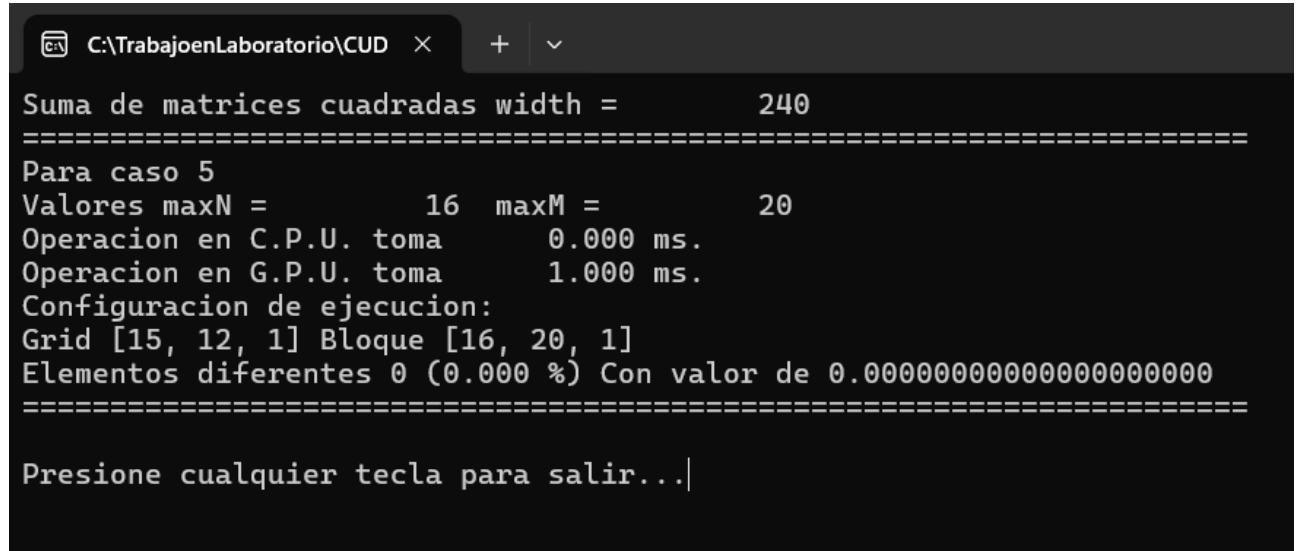
blockI dx		thread Idx		Elemento		
x	y	x	y	fila	col	#
1	2	0	0	4	6	46
1	2	0	1	4	7	47
1	2	0	2	4	8	48
1	2	1	0	5	6	56
1	2	1	1	5	7	57
1	2	1	2	5	8	58
1	2	2	0	6	6	66
1	2	2	1	6	7	67
1	2	2	2	6	8	68
1	2	3	0	7	6	76
1	2	3	1	7	7	77
1	2	3	2	7	8	78
1	3	0	0	4	9	49
1	3	0	1	4	10	50
1	3	0	2	4	11	51
1	3	1	0	5	9	59
1	3	1	1	5	10	60
1	3	1	2	5	11	61
1	3	2	0	6	9	69
1	3	2	1	6	10	70
1	3	2	2	6	11	71
1	3	3	0	7	9	79
1	3	3	1	7	10	80
1	3	3	2	7	11	81

blockIdx		threadIdx		Elemento		
x	y	x	y	fila	col	#
2	0	0	0	8	0	80
2	0	0	1	8	1	81
2	0	0	2	8	2	82
2	0	1	0	9	0	90
2	0	1	1	9	1	91
2	0	1	2	9	2	92
2	0	2	0	10	0	100
2	0	2	1	10	1	101
2	0	2	2	10	2	102
2	0	3	0	11	0	110
2	0	3	1	11	1	111
2	0	3	2	11	2	112
2	1	0	0	8	3	83
2	1	0	1	8	4	84
2	1	0	2	8	5	85
2	1	1	0	9	3	93
2	1	1	1	9	4	94
2	1	1	2	9	5	95
2	1	2	0	10	3	103
2	1	2	1	10	4	104
2	1	2	2	10	5	105
2	1	3	0	11	3	113
2	1	3	1	11	4	114
2	1	3	2	11	5	115

blockIdx		threadIdx		Elemento		
x	y	x	y	fila	col	#
2	2	0	0	8	6	86
2	2	0	1	8	7	87
2	2	0	2	8	8	88
2	2	1	0	9	6	96
2	2	1	1	9	7	97
2	2	1	2	9	8	98
2	2	2	0	10	6	106
2	2	2	1	10	7	107
2	2	2	2	10	8	108
2	2	3	0	11	6	116
2	2	3	1	11	7	117
2	2	3	2	11	8	118
2	3	0	0	8	9	89
2	3	0	1	8	10	90
2	3	0	2	8	11	91
2	3	1	0	9	9	99
2	3	1	1	9	10	100
2	3	1	2	9	11	101
2	3	2	0	10	9	109
2	3	2	1	10	10	110
2	3	2	2	10	11	111
2	3	3	0	11	9	119
2	3	3	1	11	10	120
2	3	3	2	11	11	121

Caso 5. Bloque 2D con hilos 2D

```
#define width 32
#define epsilon float(0.0000001)
#define maxN 16
#define maxM 20
...
int numBloquesN = divEntera(width , maxN);
int numBloquesM = divEntera(width , maxM);
dim3 dimGrid(numBloquesN, numBloquesM);
dim3 dimBlock(maxN, maxM);
...
int fila = (blockIdx.x * blockDim.x) + threadIdx.x;
int columna = (blockIdx.y * blockDim.y) + threadIdx.y;
if ((fila < width) && (columna < width)) {
    int tid = (fila * width) + columna;
    c[tid] = a[tid] + b[tid];
}
```



```
C:\TrabajoenLaboratorio\CUD x + v
Suma de matrices cuadradas width =      240
=====
Para caso 5
Valores maxN =      16  maxM =      20
Operacion en C.P.U. toma      0.000 ms.
Operacion en G.P.U. toma      1.000 ms.
Configuracion de ejecucion:
Grid [15, 12, 1] Bloque [16, 20, 1]
Elementos diferentes 0 (0.000 %) Con valor de 0.000000000000000000000000
=====
Presione cualquier tecla para salir...|
```

Bibliografía

- Documentación **CUDA C++ Programming Guide** NVIDIA. 2024
<https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html>
- Sitio **CUDA Toolkit Documentation** NVIDIA, 2024.
<https://docs.nvidia.com/cuda/index.html>
- Storti, Duane; Yurtoglu, Mete. **CUDA for Engineers: An Introduction to High-Performance Parallel Computing**. Addison Wesley. 2015.
- Cheng, John; Grossman, Max; McKercher. **Professional CUDA C Programming**. Edit. Wrox. 2014.
- Sanders, Jason; Kandrot, Edward. **CUDA by Example: An Introduction to General-Purpose GPU Programming**. Addison Wesley. 2011.
- Kirk, David; Hwu, Wen-mei. **Programming Massively Parallel Processors: A Hands-on Approach**. Elsevier. 2010.

Gracias por su atención



**U.A.Q. Fac. de Informática
Campus Juriquilla**

**Dra. Sandra Luz Canchola Magdaleno
sandra.canchola@uaq.mx
Cel. 442-1369270**

**Dra. Reyna Moreno Beltrán
reyna.moreno@uaq.mx**