# Assignment 2

## Cyclic Cross Correlation

CATARINA BARROQUEIRO - 103895
GONÇALO SILVA - 103244
GRUPO 1 (TP2) - AAD

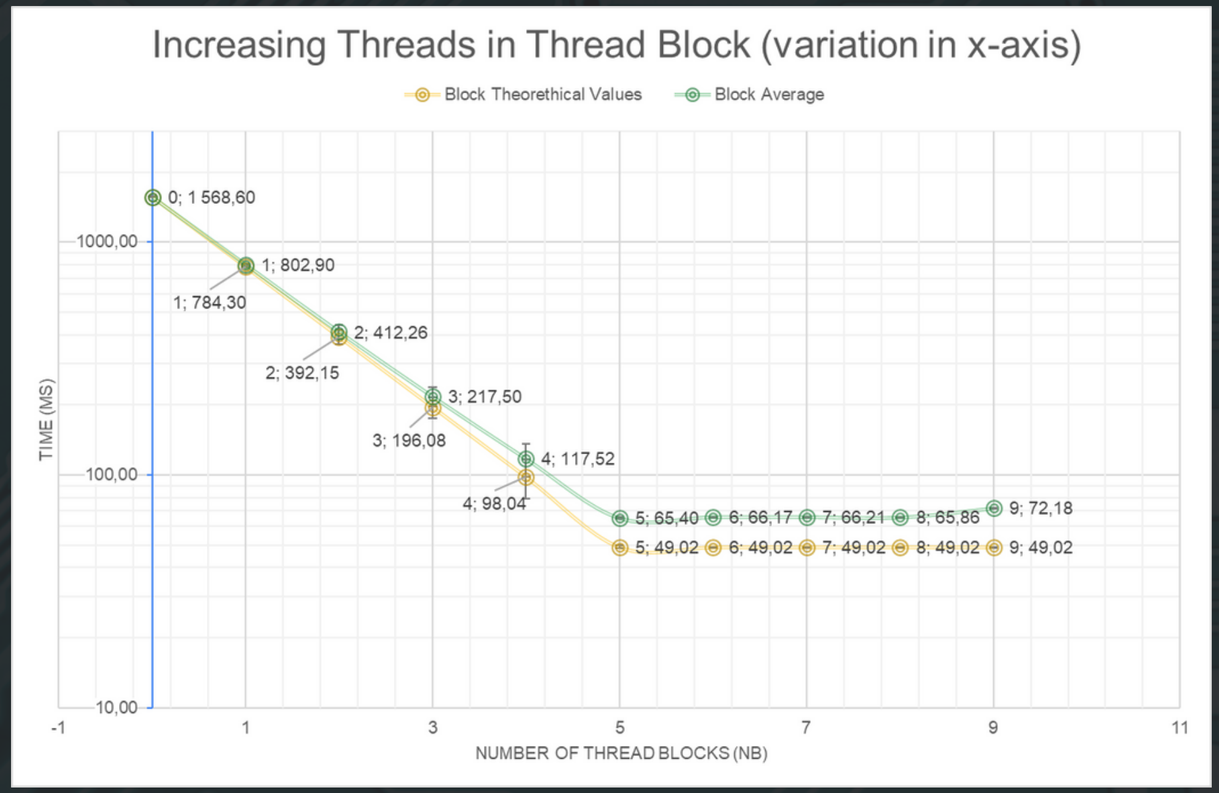
# GOAL & OPTIMAL NUMBER OF THREADS PER BLOCK

## GOAL

- Optimize circular cross-correlation thread launch grid, investigate GPU performance factors, and draw conclusions on computation efficiency when offloading to a GPU.

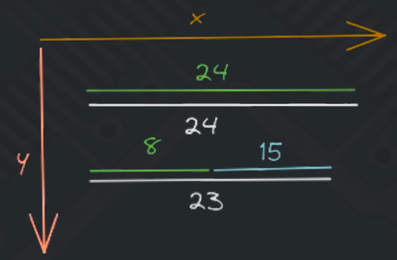
## OPTIMAL NUMBER OF THREADS PER BLOCK

- Obtained the optimal configuration with grid dimension 2048, block dimension 32, yielding an average time of 65.45ms with a standard deviation of 0.663ms.



# PROCESS

- Modifying Block dimensions to an **x-y** field generated huge differences in the average time of executions
- A cause can be the execution of each thread in a thread block, being processed linearly, even though appearing in a matrix format, it will always be executed by the SM in a 1D order
- Variance was also heavily impacted




- Another cause can be the execution order of each thread being on a first come first serve, which may be leading threads to always be in load while starving others
- Finally, with threads divided in the **x-y** axis, will lead the data of the program to be more disperse, causing more cache misses, thus slowing the program

| Messing with X-Y Threads per Block | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GridDimX | GridDimY | GridDimZ | BlockDimX | BlockDimY | BlockDimZ | Run 1 (ms) | Run 2 (ms) | Run 3 (ms) | Run 4 (ms) | Run 5 (ms) | Average | Variance |
| 2048 | 1 | 1 | 32 | 1 | 1 | 65,12 | 66,06 | 64,54 | 66,12 | 65,43 | 65,45 | 0,663 |
| 2048 | 1 | 1 | 16 | 2 | 1 | 60,86 | 68,79 | 68,80 | 68,80 | 68,81 | 67,21 | 3,5509 |
| 2048 | 1 | 1 | 8 | 4 | 1 | 51,91 | 69,09 | 69,12 | 69,10 | 69,11 | 65,67 | 7,6898 |
| 2048 | 1 | 1 | 4 | 8 | 1 | 83,49 | 83,41 | 63,18 | 62,88 | 83,34 | 75,26 | 11,165 |
| 2048 | 1 | 1 | 2 | 16 | 1 | 144,10 | 108,80 | 143,20 | 142,70 | 144,20 | 136,60 | 15,5533 |
| 2048 | 1 | 1 | 1 | 32 | 1 | 143,60 | 143,90 | 144,00 | 144,50 | 142,80 | 143,76 | 0,6269 |
| **Launching Number of Thread Blocks (nB)** | | | **Launching Thread per Block (nT)** | | | | | | | | | |

# PROCESS

- Modifying Grid dimensions to an **x-y** field generated small differences. This may be due to the execution being handled by the scheduller, which is more consistent
- Another factor is the grid only determining the number of thread blocks to be scheduled, not having impact in the handling of the threads and execution inside the block

| Messing with X-Y Block Grid | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **GridDimX** | **GridDimY** | **GridDimZ** | **BlockDimX** | **BlockDimY** | **BlockDimZ** | **Run 1 (ms)** | **Run 2 (ms)** | **Run 3 (ms)** | **Run 4 (ms)** | **Run 5 (ms)** | **Average** | **Variance** |
| 2048 | 1 | 1 | 32 | 1 | 1 | 65,12 | 66,06 | 64,54 | 66,12 | 65,43 | 65,45 | 0,663 |
| 1024 | 2 | 1 | 32 | 1 | 1 | 68,78 | 68,80 | 68,82 | 68,83 | 68,80 | 68,81 | 0,0195 |
| 512 | 4 | 1 | 32 | 1 | 1 | 68,81 | 68,81 | 68,80 | 68,83 | 68,80 | 68,81 | 0,0122 |
| 256 | 8 | 1 | 32 | 1 | 1 | 68,78 | 68,79 | 68,84 | 68,81 | 66,83 | 68,41 | 0,8835 |
| 128 | 16 | 1 | 32 | 1 | 1 | 68,82 | 68,80 | 68,78 | 68,79 | 68,81 | 68,80 | 0,0158 |
| 64 | 32 | 1 | 32 | 1 | 1 | 68,80 | 68,81 | 68,83 | 68,79 | 68,82 | 68,81 | 0,0158 |
| 32 | 64 | 1 | 32 | 1 | 1 | 68,78 | 68,80 | 68,79 | 68,82 | 68,82 | 68,80 | 0,0179 |
| 16 | 128 | 1 | 32 | 1 | 1 | 68,82 | 68,82 | 68,83 | 68,81 | 68,79 | 68,81 | 0,0152 |
| 8 | 256 | 1 | 32 | 1 | 1 | 68,79 | 68,78 | 68,80 | 68,81 | 68,79 | 68,79 | 0,0114 |
| 4 | 512 | 1 | 32 | 1 | 1 | 68,78 | 68,76 | 68,80 | 68,76 | 68,84 | 68,79 | 0,0335 |
| 2 | 1024 | 1 | 32 | 1 | 1 | 68,81 | 68,78 | 68,83 | 68,83 | 68,81 | 68,81 | 0,0205 |
| 1 | 2048 | 1 | 32 | 1 | 1 | 68,79 | 68,80 | 68,79 | 68,82 | 68,80 | 68,80 | 0,0122 |
| **Launching Number of Thread Blocks (nB)** | | | **Launching Thread per Block (nT)** | | | | | | | | | |

# GPU vs CPU

- **CPU EXECUTION:**
  - **Average Time :** 3103 ± 1.8257 ms
- **GPU:**
  - **Average Time :** 200.08 ± 3.754 ms
    - **Memory Host -> Device**: 134.50 ± 3.6968 ms
    - **Memory Device -> Host** : 0.130 ± 0.0034 ms
    - **Processing Time** : 65.45 ± 0.663 ms
- **SPEED-UP:**
  - **Upper bound:** 15.566 times
  - **Lower bound:** 15.533 times

**Executing the program in the GPU is around 16 times faster than on the CPU**

| CPU Execution | | | | | | |
|---|---|---|---|---|---|---|
| **Run 1 (ms)** | **Run 2 (ms)** | **Run 3 (ms)** | **Run 4 (ms)** | **Run 5 (ms)** | **Average** | **Variance** |
| 3120 | 3101,00 | 3102,00 | 3104,00 | 3105,00 | 3103,00 | 1,826 |

| Memory Transfer | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Direction** | **Run 1 (ms)** | **Run 2 (ms)** | **Run 3 (ms)** | **Run 4 (ms)** | **Run 5 (ms)** | **Average** | **Variance** |
| **Host -> Device** | 136 | 138,00 | 133,00 | 130,00 | 137,00 | 134,50 | 3,6968 |
| **Device -> Host** | 0.127 | 0,129 | 0,135 | 0,128 | 0,128 | 0,130 | 0,0034 |