

Esercitazione 5 MNI

Studente: Giuseppe Cardaropoli

Matricola: 052251310

GPU: Tesla T4 (Colab)

| Compute Capability 7.5 |
|--|
| Maximum x-dim of grid of thread blocks = $2^{31}-1$ |
| Maximum number of resident threads per multiprocessor = 1024 |
| Maximum number of resident blocks per multiprocessor = 16 |
| Maximum number of threads per block = 1024 |
| Number of 32.bit registers per thread block = 64K |

Tempo di Esecuzione Sequenziale

| N | M | Tempo CPU(s) |
|-------|-------|--------------|
| 1024 | 1024 | 0,008102 |
| 2048 | 2048 | 0.030047 |
| 4096 | 4096 | 0.115502 |
| 8192 | 8192 | 0.459463 |
| 16384 | 16384 | 1.865172 |

Configurazione Ottimale Esercitazione n°5

GPU: NVIDIA Tesla T4

Compute Capability: 7.5

max numero di blocchi-x: $2^{31}-1$

max numero di blocchi-y: 65535

max num thread VSM: 1024

max num blocchi VSM: 16

max num thread Vblocco: 1024

num registri VSM: 64K

$$\left. \begin{aligned} \text{blockDim.x} &= \sqrt{1024/16} = \sqrt{64} = 8 \\ \text{blockDim.y} &= \sqrt{1024/16} = \sqrt{64} = 8 \end{aligned} \right\} 8 \times 8 = 64 < 1024 \Rightarrow \text{vincolo sulla dimensione del blocco rispettato}$$

$$\text{thread usati VSM} = (8 \times 8) \times 16 = 1024 \Rightarrow \text{OTTIMALE}$$

$$\text{num registri usati da ogni thread} = 10$$

$$\text{registri totali usati VSM} = 1024 \times 10 = 10240 < 64K \Rightarrow \text{vincolo sui registri rispettato}$$

Configurazione 4×4

Numero di thread per blocco: $4 \times 4 = 16$.

Numero di blocchi residenti per ogni SM: $\frac{1024}{16} = 64$, (possiamo avere al massimo **16** blocchi).

Numero di thread attivi per ogni SM: $16 \times 16 = 256$.

Con questa configurazione si attiveranno solamente **256** thread per ogni SM, su un totale di **1024**.

| N | M | Tempo GPU(s) | Speedup |
|-------|-------|--------------|---------|
| 1024 | 1024 | 0.000234 | 34,62 |
| 2048 | 2048 | 0.001005 | 29,89 |
| 4096 | 4096 | 0.004454 | 25,82 |
| 8192 | 8192 | 0.018672 | 24,60 |
| 16384 | 16384 | 0.121833 | 15,31 |

Configurazione 8×8 (ottimale)

Numero di thread per blocco: $8 \times 8 = 64$.

Numero di blocchi residenti per ogni SM: $\frac{1024}{64} = 16$.

Numero di thread attivi per ogni SM: $64 \times 16 = 1024$.

Con questa configurazione utilizziamo tutti i thread e blocchi a disposizione per ogni SM.

| N | M | Tempo GPU(s) | Speedup |
|-------|-------|--------------|---------|
| 1024 | 1024 | 0.000168 | 48,22 |
| 2048 | 2048 | 0.000720 | 41,73 |
| 4096 | 4096 | 0.003399 | 33,98 |
| 8192 | 8192 | 0.013229 | 34,73 |
| 16384 | 16384 | 0.060192 | 30,98 |

Configurazione 16×16

Numero di thread per blocco: $16 \times 16 = 256$.

Numero di blocchi residenti per ogni SM: $\frac{1024}{256} = 4$.

Numero di thread attivi per ogni SM: $256 \times 4 = 1024$.

Con questa configurazione utilizziamo tutti i thread ma non tutti i blocchi a disposizione per ogni SM. Con questa configurazione otteniamo un parallelismo ridotto.

| N | M | Tempo GPU(s) | Speedup |
|-------|-------|--------------|---------|
| 1024 | 1024 | 0.000302 | 26,82 |
| 2048 | 2048 | 0.001172 | 25,63 |
| 4096 | 4096 | 0.004511 | 25,60 |
| 8192 | 8192 | 0.018040 | 25,46 |
| 16384 | 16384 | 0.071918 | 25,93 |

Speed Configurazioni

