Esercitazione 5: Somma Matrici Quadrate

GPU: Tesla T4
Compute capability: 7.5
Massimo numero di thread per blocco per SM: 1024
Numero massimo di blocchi residenti per SM: 16
Massimo numero di registri a 32 bit per multiprocessor/thread: 64k

Configurazione 1 : 8 x 8

N	Tempo CPU	Tempo GPU	Sp
1024	4,9456	0,0639	77,4080
2048	20,6178	0,2164	95,2763
4096	82,2401	0,8478	97,0041
8192	338,3068	3,2567	103,8802
15000	1.128,0279	10,8201	104,2530

8x8 = 64 thread: 1024/64 = 16 blocchi residenti.
Con un massimo di 16 blocchi per SM : 64x16 = 1024 thread per SM su un totale di 1024 disponibili. Piena occupazione dello SM!

Uso dei registri

Il numero di registri utilizzato da ogni thread è 8.
Dunque, moltiplicando il numero di registri, per il numero di thread e per il numero di blocchi ottengo: 8*64*16= 8192 < 64K

Configurazione 2 : 16 x 16

N	Tempo CPU	Tempo GPU	Sp
1024	4,9316	0,0828	59,5748
2048	20,3216	0,2150	94,5191
4096	81,2010	0,8314	97,6678
8192	324,0897	3,2562	99,5300
15000	1.087,2656	10,9832	98,9935

16x16 = 256 thread per blocco : 1024/256 = 4 blocchi residenti.
Con 4 blocchi: 256x4 = 1024 thread per SM. Occupazione parziale dello SM!

Uso dei registri

Il numero di registri utilizzato da ogni thread è 8.
Dunque, moltiplicando il numero di registri, per il numero di thread e per il numero di blocchi ottengo: 8*256*4= 8192 < 64K

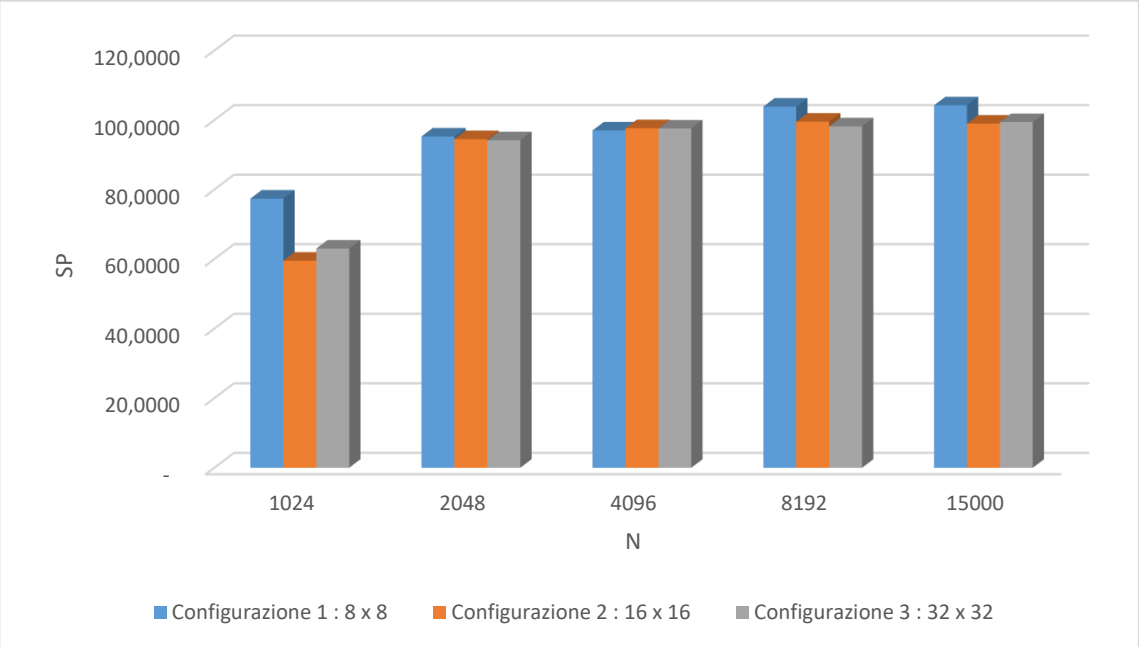
Configurazione 3 : 32 x 32

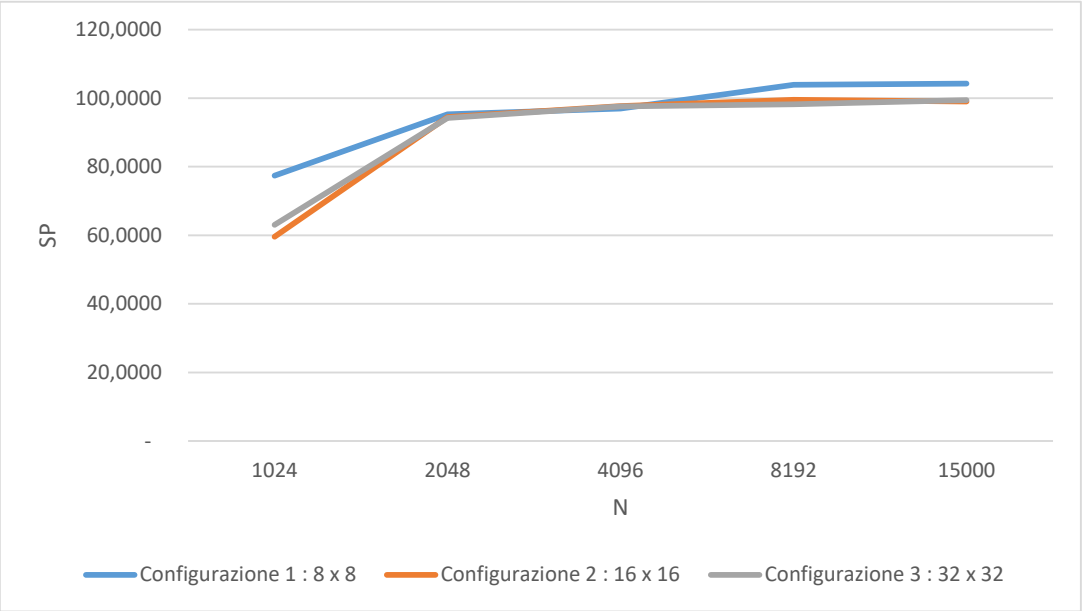
N	Tempo CPU	Tempo GPU	Sp
1024	5,8068	0,0921	63,0486
2048	20,2423	0,2149	94,2028
4096	81,1341	0,8314	97,5873
8192	317,4366	3,2324	98,2046
15000	1.117,8095	11,2425	99,4271

32x32= 1024 thread : 1024/1024 = 1 blocco residente.
Con 1 blocco: 1024 x 1 = 1024 thread per SM. Piena occupazione dello SM ma minore parallelismo.

Uso dei registri

Il numero di registri utilizzato da ogni thread è 8.
Dunque, moltiplicando il numero di registri, per il numero di thread e per il numero di blocchi ottengo: 8*1024*1= 8192 < 64K




120,0000
100,0000
80,0000
60,0000
40,0000
20,0000
-
SP
1024
2048
4096
8192
15000
N
Configurazione 1 : 8 x 8
Configurazione 2 : 16 x 16
Configurazione 3 : 32 x 32