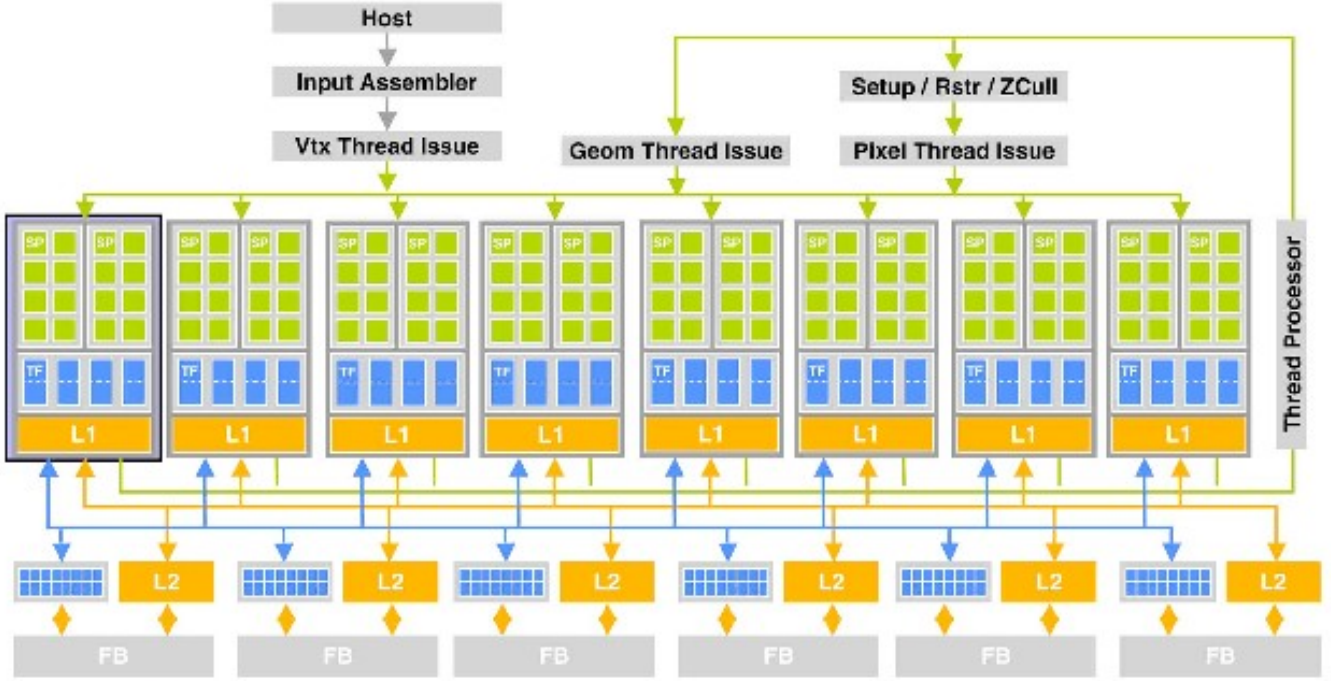
Multiplicação de Vetor por Matriz em CUDA

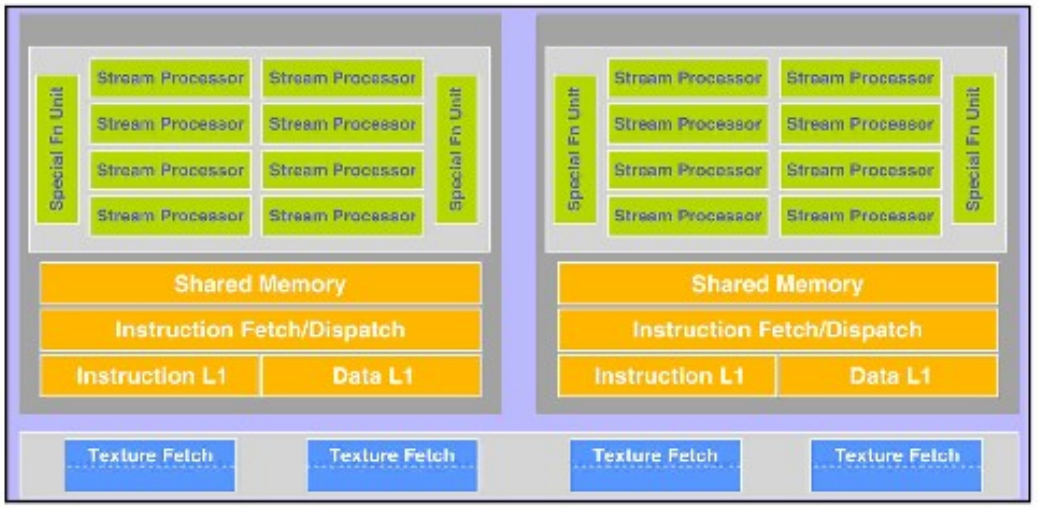
Disciplina de Programação Paralela
Universidade Federal de São Carlos - UFSCar

Erik Aceiro Antonio

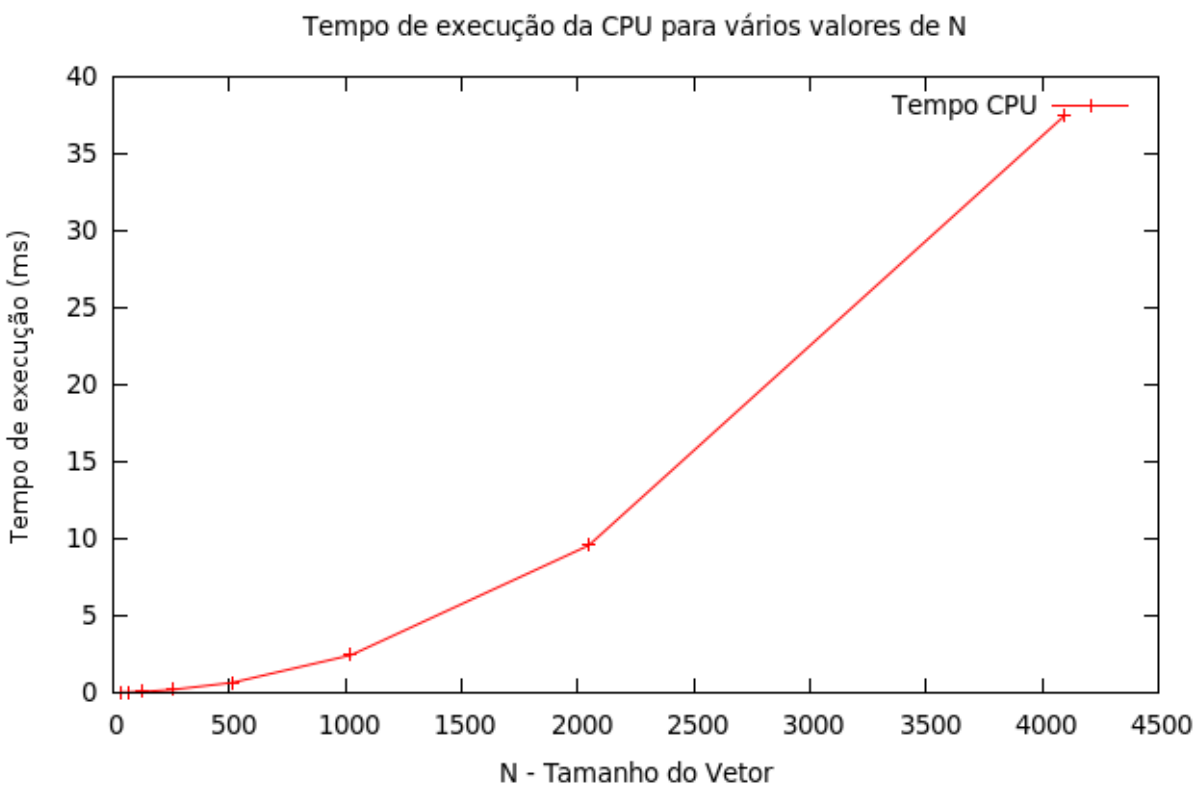
Prof. Dr. Hermes Senger

São Carlos
2010

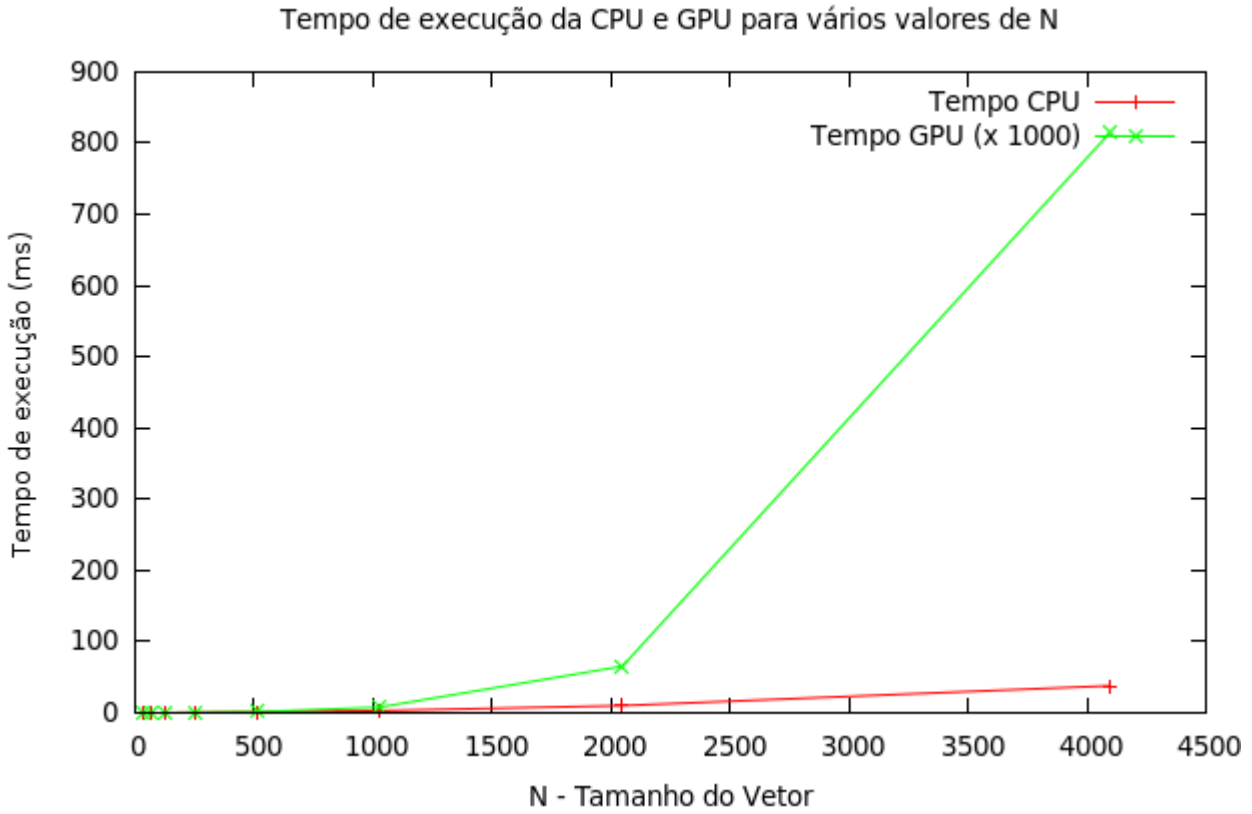




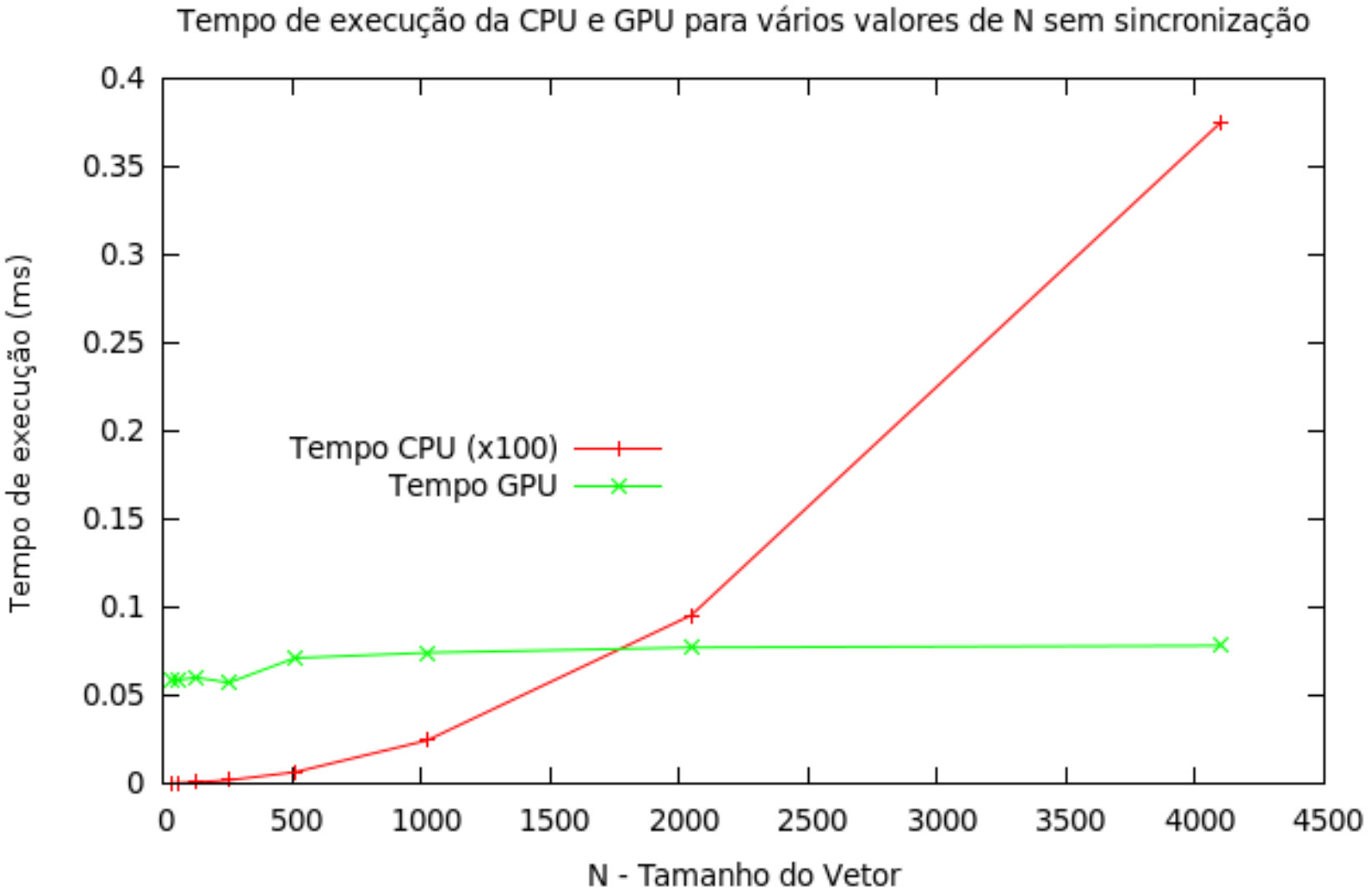
Atualmente AMD e NVIDIA possuem massiva, unidades programáveis em seus núcleos – A NVIDIA GeForce 8800 GTX com 16 streaming multiprocessors de 8 thread (stream) processadores cada.
Um par de stream de multiprocessadores é mostrado abaixo. Cada um contém instruções compartilhadas, cache de dados, ULA, 16 kB de memória compartilhada, 8 processadores e duas unidades especiais.

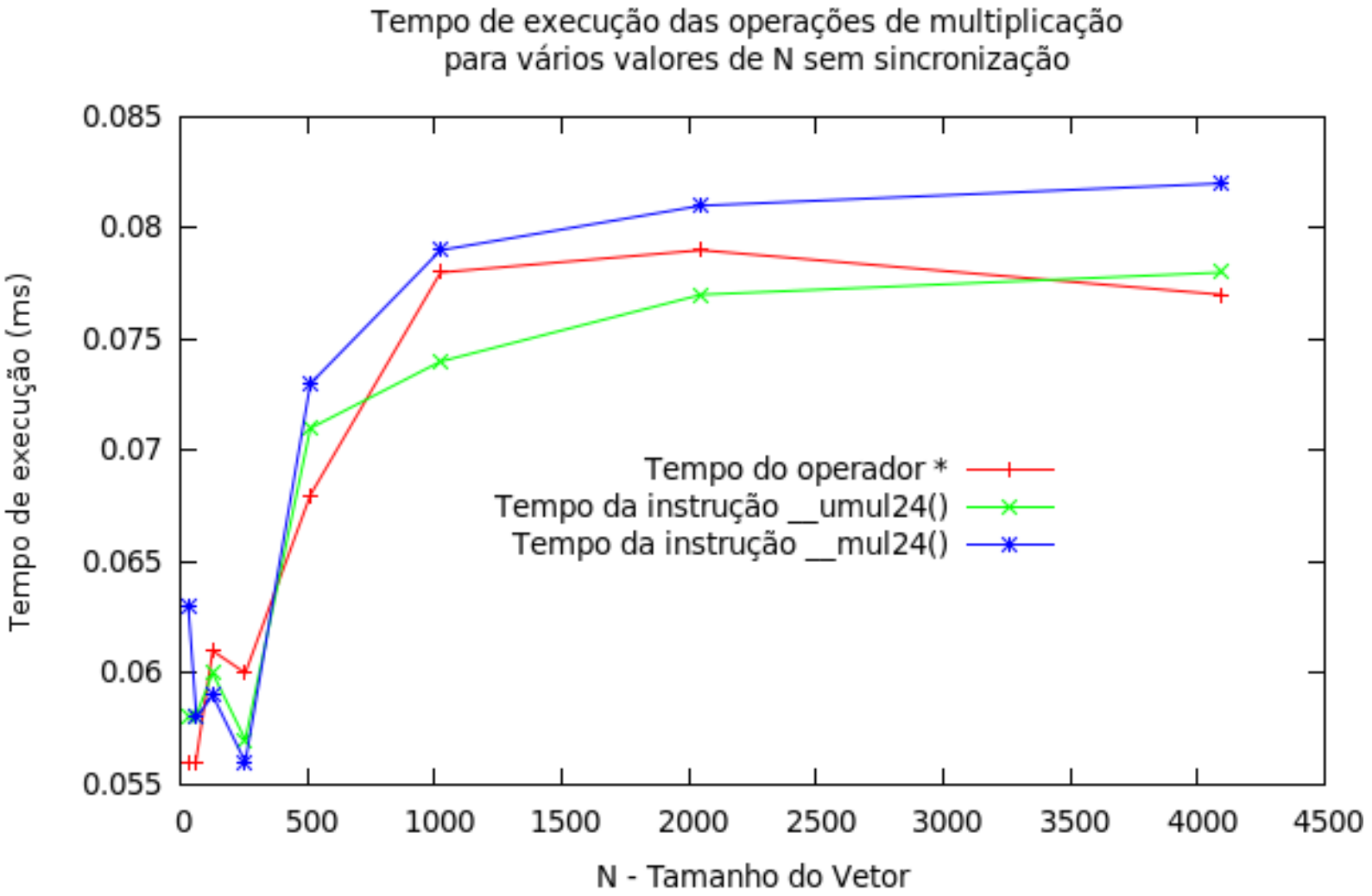

Tempo de execução da CPU para vários valores de N
Tempo CPU
Tempo de execução (ms)
N - Tamanho do Vetor
0
5
10
15
20
25
30
35
40
0
500
1000
1500
2000
2500
3000
3500
4000
4500

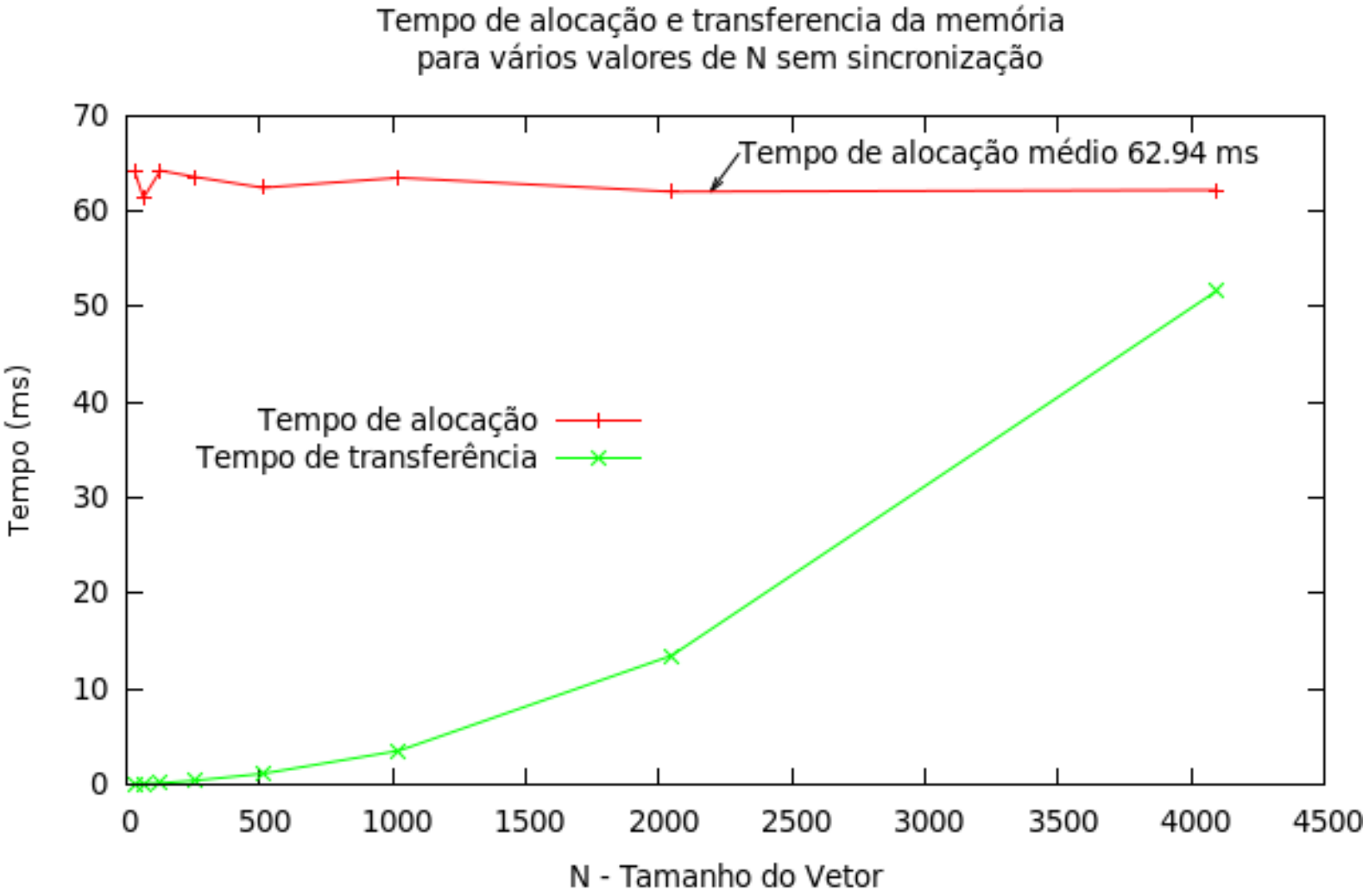

Tempo de execução da GPU para vários valores de N
Tempo GPU
Tempo de execução (ms)
N - Tamanho do Vetor
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
0
500
1000
1500
2000
2500
3000
3500
4000
4500

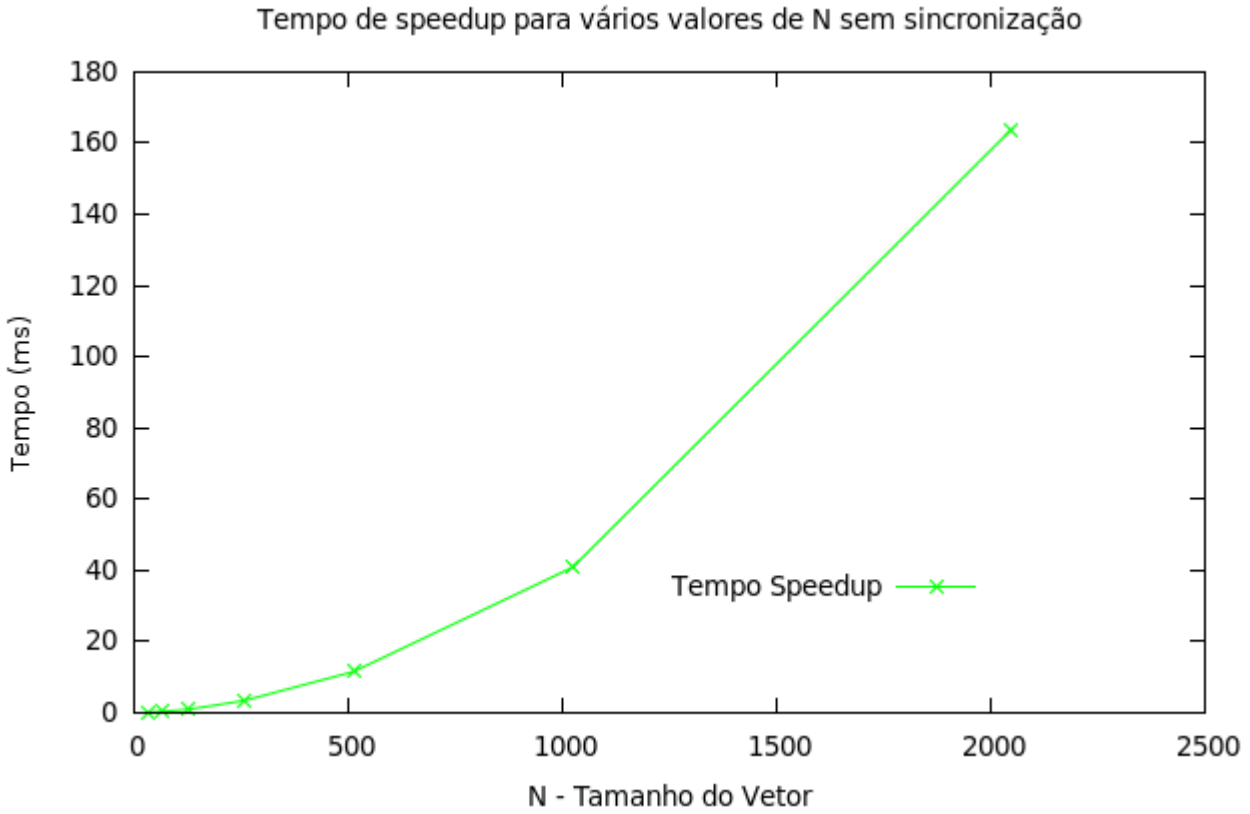


N	Tempo CPU	Tempo GPU (ms)	Tempo GPU (s) e (mim)
32	0.005000	0.249000	0.00025
64	0.015000	1.456000	0.00146
128	0.053000	20.622000	0.02062
256	0.174000	144.916000	0.14492
512	0.634000	1009.034973	> 1 segundo
1024	2.418000	7222.379883	> 7 segundos
2048	9.549000	64847.441406	> 1 mim
4096	37.462002	814199.375000	> 13 mim


Tempo de execução da CPU e GPU para vários valores de N sem sincronização
Tempo de execução (ms)
N - Tamanho do Vetor
Tempo CPU (x100)
Tempo GPU
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0
500
1000
1500
2000
2500
3000
3500
4000
4500


Tempo de execução das operações de multiplicação
para vários valores de N sem sincronização
Tempo de execução (ms)
0.085
0.08
0.075
0.07
0.065
0.06
0.055
0
500
1000
1500
2000
2500
3000
3500
4000
4500
N - Tamanho do Vetor
Tempo do operador *
Tempo da instrução __umul24()
Tempo da instrução __mul24()


Tempo de alocação e transferencia da memória
para vários valores de N sem sincronização
Tempo de alocação médio 62.94 ms
Tempo de alocação
Tempo de transferência
Tempo (ms)
N - Tamanho do Vetor
0
10
20
30
40
50
60
70
0
500
1000
1500
2000
2500
3000
3500
4000
4500



N	GPU	CPU	Speedup
32	0.041000	0.003000	0.07
64	0.042000	0.010000	0.24
128	0.042000	0.037000	0.88
256	0.046000	0.151000	3.28
512	0.052000	0.595000	11.44
1024	0.058000	2.365000	40.78
2048	0.058000	9.501000	163.81
4096	0.066000	37.444000	567.33