Análise, otimização e paralelismo num algoritmo k-means

Computação Paralela

1º Semestre 1º Ano Mestrado em Engenharia Informática
Escola de Engenharia – Universidade do Minho

Trabalho realizado por:
- Duarte Lucas PG50345
- Tiago Ribeiro PG50779

01

TP1

Versão sequencial

02

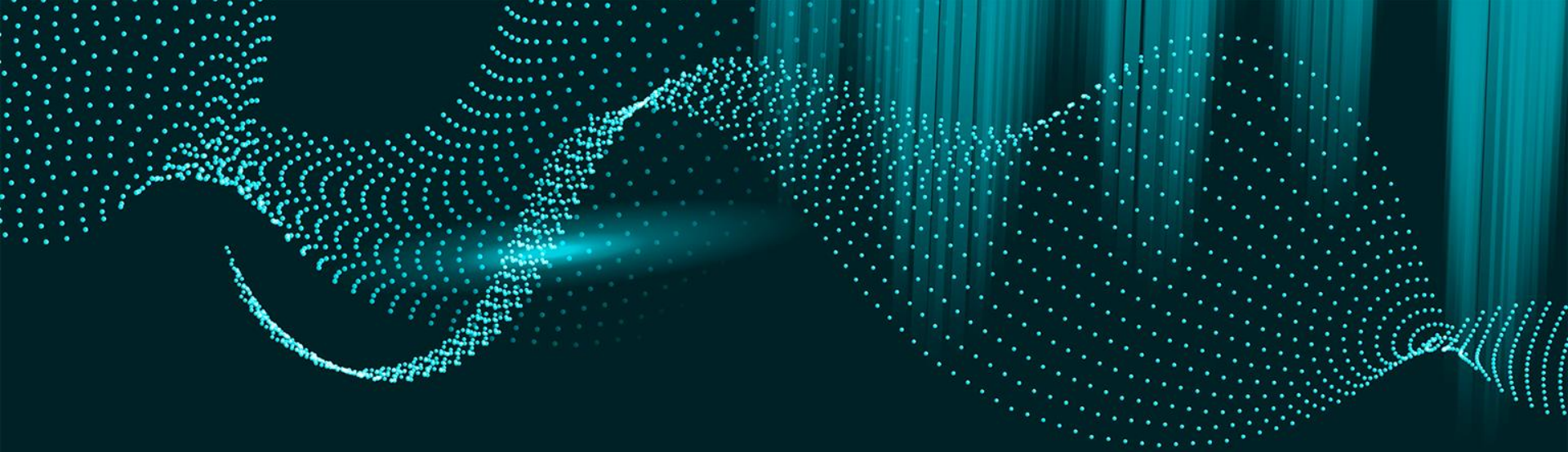
TP2

Versão paralela em OpenMP

03

TP3

Versão paralela em CUDA



01 TP1

Versão sequencial

CONCEITOS ABORDADOS

Vetorização

Realizar uma operação matemática em vários elementos, ao mesmo tempo

Loop-unrolling

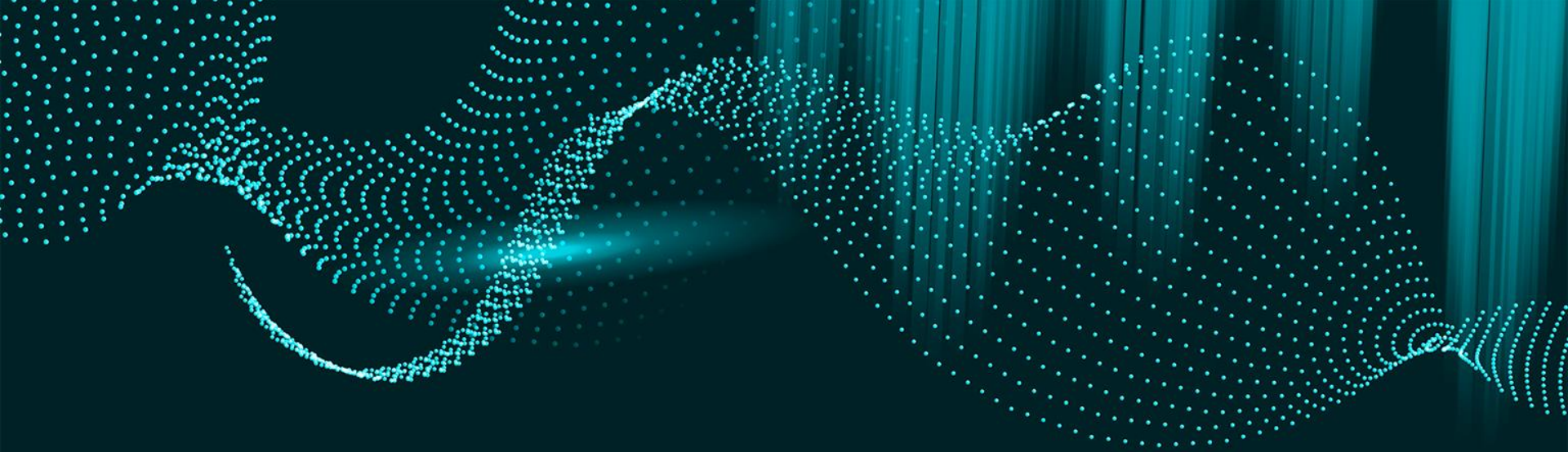
Redução de instruções que controlem o ciclo

Inlining

Remover chamadas a funções complexas

TEMPOS DE EXECUÇÃO

	#1	#2	#3	#4	#5	Média
-O2 Time	6.32	9.14	5.87	6.08	5.91	6.66
-O2 CPI	0.56	0.83	0.54	0.54	0.54	0.60
-O3 Time	7.33	9.93	7.62	9.20	10.31	8.88
-O3 CPI	0.94	1.25	0.95	1.15	1.32	1.12



02 TP2

Versão paralela em OpenMP

CONCEITOS ABORDADOS

Parelelismo

Técnica de programação que permite que várias tarefas sejam executadas simultaneamente

Data Racing

Controlo no tempo de acesso aos dados por diversas threads

OpenMP

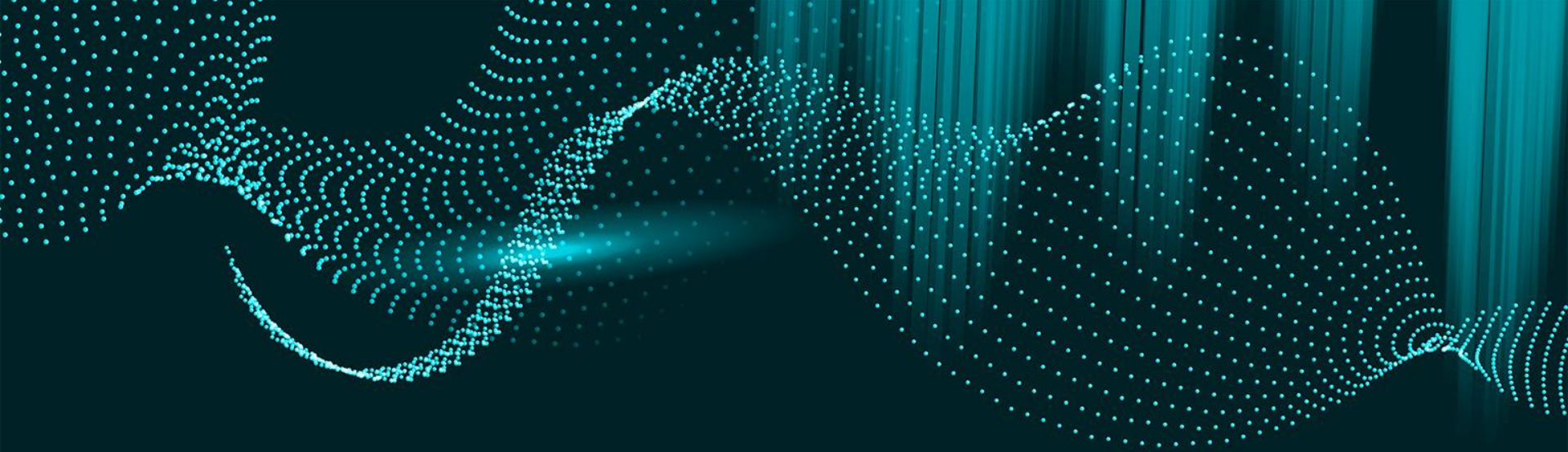
Biblioteca de programação paralela que permite o desenvolvimento de aplicações multithread

Balanceamento da Carga

Distribuição da carga pelas várias threads

TEMPOS DE EXECUÇÃO

	CC (mil milhões)	#I (mil milhões	CPI	Texec (s)
Sequencial 4 clusters	7.2	18.3	0.4	2.43
Sequencial 32 clusters	40.9	94.8	0.4	13.39
Paralelo 16T 4C	8.9	18.4	0.5	0.96
Paralelo 16T 4C	44.0	87.6	0.5	1.68



03 TP3

Versão paralela em CUDA

CONCEITOS ABORDADOS

Parelelismo

Técnica de programação que permite que várias tarefas sejam executadas simultaneamente

Data Racing

Controlo no tempo de acesso aos dados por diversas threads

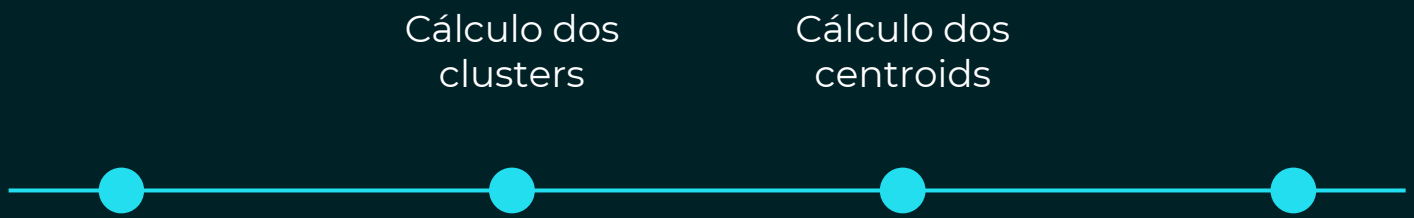
CUDA

Programação paralela que permite o desenvolvimento de aplicações multithread

Utilização da GPU

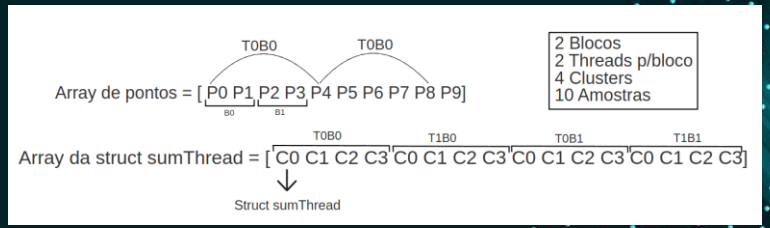
Maximizar o tempo de GPU para tirar o melhor partido

CICLO DE VIDA DO ALGORITMO



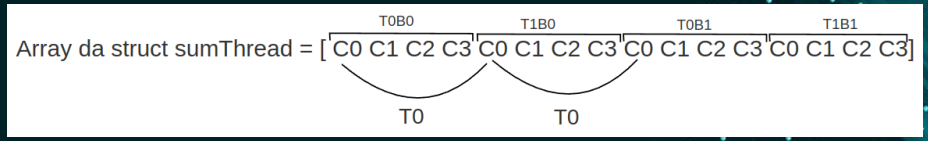
CÁLCULO DOS CLUSTERS

- Cada thread percorre o array de amostras;
- Armazena o somatório dos pontos na estrutura dos somatórios.



CÁLCULO DOS CENTROIDS

- São criadas tantas threads quantos clusters;
- Percorrem o array e coletam dados para calcular os novos centroids.



ESPECIFICAÇÕES DA GPU

	VRAM	Frequência a memória	Tamanho do Bus de memória	Largura de banda da memória	Nº de SMM's
Tamanho/Velocidade	2 GB	2446 MHz	64 bits	40.1 GB/s	3

	Tamanho
L1	64 KB p/ SMM
L2	1024 KB
VRAM	2 GB

ARQUITETURA E TESTES

ARQUITETURA GPU

- Microarquitectura Pascal
- Capacidade de computação 6.1
- Máximo de threads a correr: 6144

TESTES A REALIZAR

- 8 mil amostras (~64KB)
- 100 mil amostras (~800KB)
- 10 milhões amostras (~8MB)

RESULTADOS

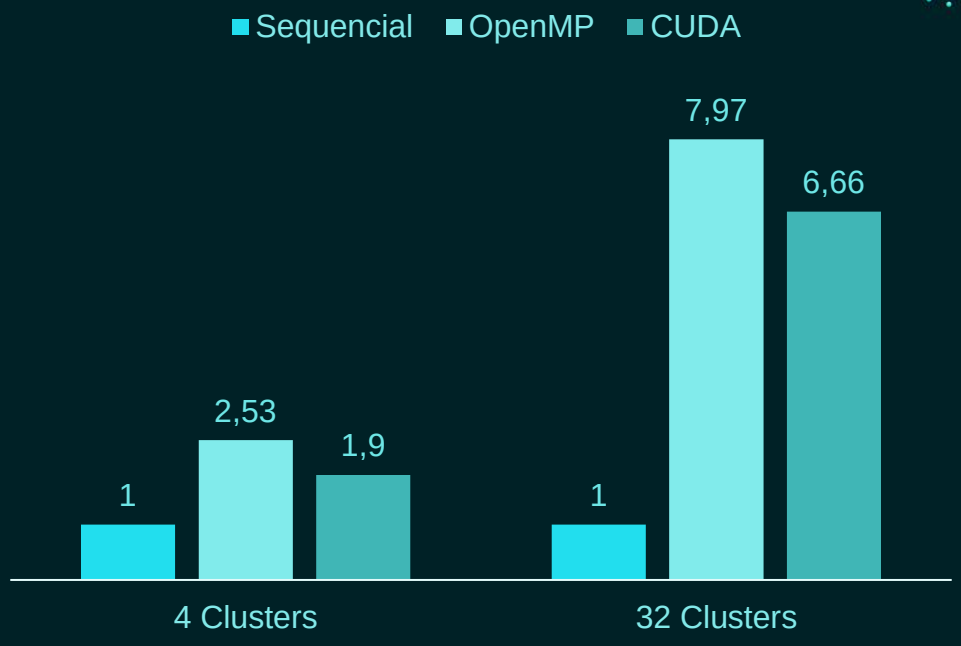
4 CLUSTERS	(1)	(2)	(3)
8 mil (~64KB)	0.098	0.175	0.240
100 mil (~800KB)	0.116	0.165	0.166
10 milhões (~8MB)	1.284	1.410	1.385

32 CLUSTERS	(1)	(2)	(3)
8 mil (~64KB)	0.106	0.175	0.242
100 mil (~800KB)	0.155	0.192	0.194
10 milhões (~8MB)	2.014	2.332	2.213

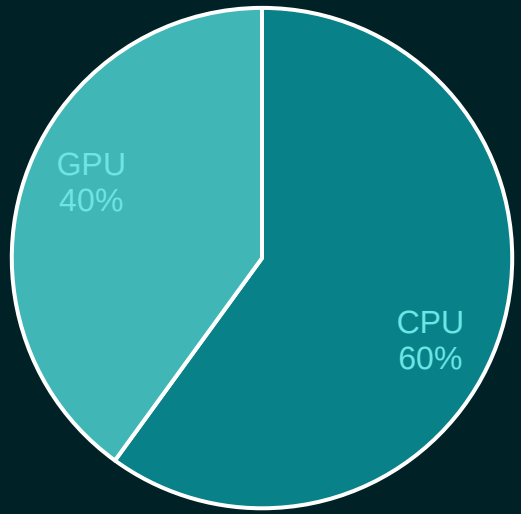
COMPARAÇÃO ENTRE VERSÕES

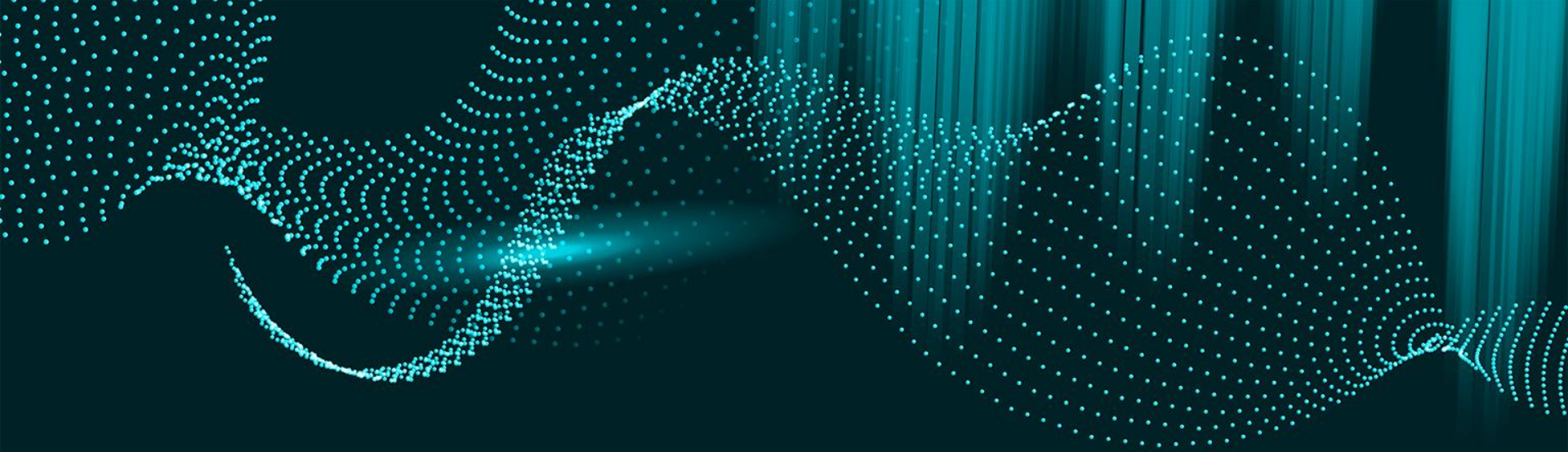
	4 Clusters	Speedup	32 Clusters	Speedup
Sequencial	2.43	1	13.39	1
OpenMP (16 threads)	0.96	2.53	1.68	7.97
Cuda Versão (1)	1.28	1.90	2.01	6.66

SPEEDUP


Sequencial
OpenMP
CUDA
1
2,53
1,9
1
7,97
6,66
4 Clusters
32 Clusters

DISTRIBUIÇÃO DO TEMPO





FIM

Conclusões
Trabalho Futuro