

# Projeto 3 — DGEMM Sequencial e Paralela para processamento em GPGPUs com CUDA

## **Autores:**

Italo Santana Seara

Wilson Santos Silva Filho

**Disciplina:** DEC107 — Processamento Paralelo

**Professor:** Esbel Tomas Valero Orellana

**Data:** 26 de novembro de 2025

# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>2</b>
<b>2</b>	<b>Metodologia</b>	<b>3</b>
2.1	Implementação Sequencial . . . . .	3
2.2	Implementação Paralela com OpenMP . . . . .	3
2.3	Implementação Paralela com MPI . . . . .	3
2.4	Implementação Paralela para GPGPUs com CUDA . . . . .	3
2.4.1	Organização de Threads e Blocos . . . . .	4
2.4.2	Uso de Memória Compartilhada <code>__shared__</code> e Hierarquia de Memória	4
2.4.3	Sobreposição de Computação e Comunicação . . . . .	4
2.5	Diferença relativa máxima . . . . .	5
2.6	Descrição do hardware utilizado nos testes . . . . .	5
2.7	Métricas utilizadas para avaliação . . . . .	6
2.7.1	Tempo de Execução . . . . .	6
2.7.2	Speedup . . . . .	6
2.7.3	Eficiência . . . . .	7
<b>3</b>	<b>Resultados</b>	<b>8</b>
3.1	Tempo de execução . . . . .	8
3.2	Speedup . . . . .	8
3.3	Eficiência . . . . .	9
<b>4</b>	<b>Discussão</b>	<b>10</b>
4.1	Análise de Desempenho . . . . .	10
4.2	Escalabilidade . . . . .	10
4.3	Limitações . . . . .	10
<b>5</b>	<b>Conclusão</b>	<b>11</b>

# 1 Introdução

A multiplicação de matrizes é uma operação fundamental em diversas áreas da computação científica e engenharia, presente em aplicações que vão desde álgebra linear numérica até aprendizagem de máquina e simulações físicas. Neste projeto, focamos na versão de precisão dupla da multiplicação geral de matrizes (DGEMM), cujo problema pode ser formalizado como: dado  $A \in \mathbb{R}^{m \times k}$  e  $B \in \mathbb{R}^{k \times n}$ , calcular  $C = AB$ , onde

$$C_{ij} = \sum_{t=1}^k A_{it} B_{tj}, \quad 1 \leq i \leq m, 1 \leq j \leq n.$$

A complexidade aritmética desta operação cresce como  $\mathcal{O}(mkn)$  e, para matrizes quadradas de ordem  $N$ , como  $\mathcal{O}(N^3)$ , o que torna essencial a exploração de paralelismo para resolver instâncias de grande porte de forma eficiente.

No Projeto 1 foram implementadas versões sequenciais e paralelas (por memória compartilhada, com OpenMP) da DGEMM. No Projeto 2, foi implementado a versão com o MPI, onde o estudo do paralelismo migrando o algoritmo para modelo de memória distribuída. No presente projeto 3, daremos continuidade ao estudo de processamento paralelo implementando a DGEMM em GPGPUs (General-Purpose computing on Graphics Processing Units) utilizando o alto poder de processamento desses *devices*.

Este projeto tem como objetivos específicos:

- Implementar uma versão paralela de DGEMM utilizando a programação de GPGPUs para Nvidia (CUDA) e utilizar algumas rotinas (por exemplo, `cudaMemoryCopy`, implementar estrutura hierárquica de memória (`__global__`, `__device__`, `__host__` e `__shared__`, )).
- Comparar o resultado da versão CUDA com o resultado da versão sequencial, calcular a diferença relativa máxima.
- Medir e analisar desempenho (tempo total, *speedup* e eficiência) em várias configurações de blocos e threads e tamanhos de matrizes, e comparar os resultados com os obtidos em ambiente de memória compartilhada (OpenMP, MPI).
- Comparar os resultados com os projetos 1 e 2, apresentando discussões sobre o ganho obtido com o uso da GPU e a diferença de escalabilidade entre os modelos de paralelismo

## 2 Metodologia

Nesta seção, detalhamos a implementação da multiplicação de matrizes DGEMM em suas versões sequencial e paralela utilizando programação CUDA, bem como a metodologia adotada para medir desempenho e validar diferença relativa máxima.

### 2.1 Implementação Sequencial

A implementação sequencial da DGEMM utiliza a versão otimizada do Projeto 1, que emprega blocos para melhorar a localidade de dados. As matrizes são armazenadas em vetores unidimensionais (*row-major order*) para garantir um acesso eficiente à memória. A função principal responsável pela multiplicação é apresentada no repositório do projeto.<sup>1</sup>

Esta implementação serve como referência para comparação de desempenho com a versão paralela da GPU com CUDA.

### 2.2 Implementação Paralela com OpenMP

A versão paralela com OpenMP, desenvolvida no Projeto 1, utiliza diretivas de paralelismo para distribuir o trabalho entre múltiplas threads em um ambiente de memória compartilhada. A decomposição por blocos é mantida, e a paralelização é aplicada principalmente nos loops externos da multiplicação de matrizes.

Esta implementação também está disponível no repositório do projeto.<sup>1</sup>

### 2.3 Implementação Paralela com MPI

A versão paralela da DGEMM desenvolvida utilizando o padrão MPI no Projeto 2, que permite a comunicação entre processos em um ambiente de memória distribuída, utilizou-se de estratégias que envolve e adota a decomposição das matrizes por blocos, onde cada processo é responsável por calcular uma parte da matriz resultado  $C$ . Esta implementação também está disponível no repositório do projeto.

### 2.4 Implementação Paralela para GPGPUs com CUDA

A versão paralela da DGEMM foi desenvolvida utilizando a programação para GPGPUs da Nvidia, utilizando CUDA. A GPU possui milhares de núcleos, organizar quem faz o quê é crucial. As GPUS usam um sistema de coordenadas onde a Thread: A menor unidade de execução. Cada thread roda o Kernel. O Block (Bloco): Um grupo de threads que podem compartilhar memória rápida (Shared Memory) e se sincronizar e a Grid (Grade): O conjunto total de blocos que executam o kernel. A implementação completa está disponível no repositório do projeto.<sup>1</sup>

### 2.4.1 Organização de Threads e Blocos

O código utiliza uma estratégia de decomposição de domínio em 2D, mapeando threads diretamente para as coordenadas da matriz de resultado  $C$ . Estrutura de Bloco (dim3 block): O kernel é configurado com blocos quadrados de tamanho  $\text{tile} \times \text{tile}$  (por padrão,  $32 \times 32$ , totalizando 1024 threads por bloco, que é o máximo permitido na maioria das arquiteturas CUDA). Estrutura de Grade (dim3 grid): A grade é calculada para cobrir as dimensões da matriz  $C$  ( $M$  linhas por  $N$  colunas). O número de blocos é arredondado para cima  $((N + \text{tile} - 1) / \text{tile})$  para garantir que matrizes cujas dimensões não sejam múltiplos de 32 sejam processadas corretamente. Mapeamento:  $\text{threadIdx.x}$  mapeia para a coluna dentro do bloco (e do tile).  $\text{threadIdx.y}$  mapeia para a linha dentro do bloco. A coordenada global  $(\text{row}, \text{col})$  é calculada como:

$$\text{row} = \text{blockIdx.y} \times \text{blockDim.y} + \text{threadIdx.y}$$

$$\text{col} = \text{blockIdx.x} \times \text{blockDim.x} + \text{threadIdx.x}$$

### 2.4.2 Uso de Memória Compartilhada `__shared__` e Hierarquia de Memória

A diferença crítica entre as funções `dgemm_kernel_basic` e `dgemm_kernel` é o uso da hierarquia de memória para contornar o gargalo de largura de banda da memória global (DRAM). A Otimização via Tiling: O `dgemm_kernel` implementa a técnica de Tiling (ladrilhamento). A Declaração: `extern __shared__ double smem[];` aloca memória rápida on-chip (L1/Shared), acessível por todas as threads do mesmo bloco. Utilizamos o uso de carregamento cooperativo: O código divide as matrizes  $A$  e  $B$  em sub-blocos (tiles) quadrados de tamanho  $T \times T$ . Todas as threads do bloco colaboram para carregar um tile de  $A$  e um tile de  $B$  da memória global (lenta) para a memória compartilhada (rápida). O comando `__syncthreads()` garante que todo o tile esteja carregado na memória compartilhada antes que qualquer thread comece a calcular. Do Cálculo: Uma vez que os dados estão na smem, as threads calculam o produto escalar parcial lendo apenas da memória compartilhada.

Isso causa um impacto: Em vez de fazer  $2 \times K$  leituras da memória global por thread (como no kernel básico), cada elemento é lido da memória global apenas uma vez por tile e reutilizado  $T$  vezes. Isso reduz o tráfego de memória global por um fator de  $T$  (neste caso, 32 vezes menos acesso à RAM da GPU).

### 2.4.3 Sobreposição de Computação e Comunicação

Se analisarmos algumas funções como a `dgemm_parallel_cuda` podemos observar um fluxo:

`cudaMemcpy(..., cudaMemcpyHostToDevice)` onde realiza-se a Cópia H->D

`dgemm_kernel <<< ... >>>` (Execução do Kernel) `cudaMemcpy(..., cudaMemcpyDeviceToHost)` (Cópia D->H)

Percebemos que não há sobreposição de computação e comunicação pois: As chamadas `cudaMemcpy` por padrão são bloqueantes (ou seja, síncronas em relação ao host ou serializadas nos stream padrão ) Então, a GPU fica ociosa durante a transferência de dados via PCIe, e o PCIe fica ocioso durante o cálculo no Kernel.

Para implementações futuras: Para matrizes muito grandes, o tempo de transferência pode ser significativo, o que deveria ser usado com alicerce e as funções de CUDA Streams e `AsyncMemcpy` para que a GPU processe um pedaço da matriz enquanto o outro está sendo transferido.

## 2.5 Diferença relativa máxima

Vamos validar a versão sequencial comparando os resultados obtidos com a função . . . . A implementação sequencial (rodando na CPU) serve como a referência (ground truth). Ela é, por definição, a maneira mais simples e verificável de calcular o resultado (neste caso, a multiplicação de matrizes).

REVISAR A corretude da implementação paralela com CUDA foi realizada comparando o resultado obtido com o da versão sequencial, calculando a diferença relativa máxima entre os elementos das matrizes resultantes, utilizando uma abordagem que evita divisão por zero. A fórmula utilizada foi:

$$\Delta = \max_{i,j} \frac{|C_{\text{seq}}(i,j) - C_{\text{cuda}}(i,j)|}{|C_{\text{seq}}(i,j)| + \epsilon} \quad (1)$$

onde  $\epsilon = 10^{-12}$  evita divisões por zero.

Considerar corretas as implementações cuja diferença máxima seja menor que um limite aceitável (por exemplo,  $10^{-8}$ ).

## 2.6 Descrição do hardware utilizado nos testes

Os testes foram realizados em uma máquina com as seguintes especificações:

- **Processador:** Ryzen 7 5700G (8 núcleos)
- **Memória RAM:** 32 GB
- **Sistema Operacional:** Windows 11 + WSL 2.4.10.0 (Ubuntu 24.04.2 LTS)
- **Compilador:** GCC 13.3.0
- **GPU:**

---

<sup>1</sup>Link para as implementações: [https://github.com/italoseara/DEC107/blob/main/Projeto3/src/dgemm\\_cuda.cu](https://github.com/italoseara/DEC107/blob/main/Projeto3/src/dgemm_cuda.cu)

## 2.7 Métricas utilizadas para avaliação

Usaremos as seguintes métricas para definir e comparar as implementações:

### 2.7.1 Tempo de Execução

Medido em segundos, é o tempo total gasto para completar a multiplicação das matrizes. Não inclui o tempo de inicialização ou finalização do programa, apenas o tempo gasto na execução da função de multiplicação.

$$T = T_{end} - T_{start}$$

Onde:

- $T$  é o tempo de execução.
- $T_{start}$  é o tempo registrado no início da execução da função.
- $T_{end}$  é o tempo registrado ao final da execução da função.

O tempo total de execução será medido utilizando a função `DESCREVER FUNCAO AQUI...` para a versão MPI, e `DESCREVER FUNCAO AQUI...` para a versão CUDA e para a versão sequencial.

Foi utilizado um script em python para automatizar a execução dos testes e coletar os tempos. O script executa cada versão múltiplas vezes com diferentes entradas para obter uma média dos tempos, calcula as métricas de desempenho e gera um relatório com os resultados, que foram utilizados para criar os gráficos apresentados na seção de resultados.

### 2.7.2 Speedup

Mede o ganho de desempenho da versão paralela em relação à sequencial.

$$S_p = \frac{T_s}{T_p}$$

Onde:

- $S_p$  é o speedup com  $p$  processadores/threads.
- $T_s$  é o tempo de execução da versão sequencial.
- $T_p$  é o tempo de execução da versão paralela com  $p$  processadores/threads.

### 2.7.3 Eficiência

Mede quão bem os recursos de processamento estão sendo utilizados pela versão paralela. Onde 1 é a eficiência ideal (100%).

$$E_p = \frac{S_p}{p} = \frac{T_s}{p \cdot T_p}$$

Onde:

- $E_p$  é a eficiência com  $p$  processadores/threads. Varia entre 0 e 1.
- $S_p$  é o speedup com  $p$  processadores/threads.
- $p$  é o número de processadores/threads.
- $T_s$  é o tempo de execução da versão sequencial.
- $T_p$  é o tempo de execução da versão paralela com  $p$  processadores/threads.



## 3 Resultados

Nesta seção, apresentamos os resultados obtidos a partir dos testes realizados nas implementações sequencial, paralela com CUDA da multiplicação de matrizes DGEMM. O resultado vai ser comparado com versões paralelizadas utilizando OpenMP e MPI. Os resultados incluem a validação de corretude, tabelas e gráficos de desempenho, bem como cálculos de *speedup* e eficiência.

### 3.1 Tempo de execução

Figura 1: Tempo de execução (OpenMP)

Figura 2: Tempo de execução (MPI)

Figura 3: Tempo de execução (CUDA)

REFAZER: As Figuras 1 e 2 mostram o tempo de execução das implementações com OpenMP e MPI, respectivamente, para diferentes tamanhos de matrizes e números de threads/processos. Observa-se que o tempo de execução diminui com o aumento do número de threads/processos, evidenciando o benefício do paralelismo. Contudo, é possível perceber que a implementação com OpenMP apresenta tempos de execução menores em comparação com a implementação com MPI para o mesmo número de threads/processos, o que pode ser atribuído à sobrecarga de comunicação inerente ao modelo de memória distribuída utilizado pelo MPI.

### 3.2 Speedup

Figura 4: Speedup (OpenMP)

Figura 5: Speedup (MPI)

Figura 6: Speedup (CUDA)

REFAZER: As Figuras 4 e 5 apresentam o *speedup* obtido pelas implementações com OpenMP e MPI, respectivamente. Observa-se que ambas as implementações conseguem alcançar um *speedup* significativo com o aumento do número de threads/processos.

### 3.3 Eficiência

Figura 7: Eficiência (OpenMP)

Figura 8: Eficiência (MPI)

Figura 9: Eficiência (CUDA)

REFAZER: As Figuras 7 e 8 mostram a eficiência das implementações com OpenMP e MPI, respectivamente. A eficiência tende a diminuir com o aumento do número de threads/processos, o que é esperado devido à sobrecarga de comunicação e sincronização. Não há uma diferença significativa na eficiência entre as duas implementações, embora a implementação com OpenMP apresente uma leve vantagem em alguns casos.

## 4 Discussão

Nesta seção, discutimos os resultados obtidos nas implementações sequencial, paralela com cuda da multiplicação de matrizes DGEMM. Analisamos o desempenho, a escalabilidade e as limitações de cada abordagem.

### 4.1 Análise de Desempenho

A implementação com OpenMP apresentou um desempenho superior em comparação com a implementação com MPI para o mesmo número de threads/processos. Isso pode ser atribuído à menor sobrecarga de comunicação no modelo de memória compartilhada utilizado pelo OpenMP, em contraste com o modelo de memória distribuída do MPI, que requer troca de mensagens entre processos.

### 4.2 Escalabilidade

Ambas as implementações demonstraram boa escalabilidade com o aumento do número de threads/processos. No entanto, a eficiência diminuiu à medida que o número de threads/processos aumentou, indicando que a sobrecarga de comunicação e sincronização começa a impactar negativamente o desempenho em configurações com muitos processos.

Porém, a versão com MPI possui uma vantagem clara: sua capacidade de escalar para um número maior de processos em ambientes distribuídos, como clusters de computadores, onde o OpenMP não é aplicável. Isso torna a implementação com MPI mais adequada para aplicações que exigem alta escalabilidade e distribuição de carga em múltiplos nós.

### 4.3 Limitações

Uma limitação observada na implementação com MPI é a complexidade adicional na gestão da comunicação entre processos, o que levou a diversos erros e dificuldades no debug. Além disso, a sobrecarga de comunicação pode se tornar um gargalo em sistemas com alta latência ou largura de banda limitada.

## 5 Conclusão

Neste relatório, apresentamos a implementação e análise de desempenho da multiplicação de matrizes DGEMM em suas versões sequencial, paralela com OpenMP e paralela com MPI. A implementação com OpenMP demonstrou um desempenho superior em ambientes de memória compartilhada, enquanto a implementação com MPI destacou-se pela sua capacidade de escalar em ambientes distribuídos. Ambas as implementações apresentaram boa escalabilidade, embora a eficiência tenha diminuído com o aumento do número de threads/processos devido à sobrecarga de comunicação.

Em trabalhos futuros, sugerimos explorar otimizações adicionais na implementação com MPI, como o uso de técnicas avançadas de comunicação e balanceamento de carga. Além disso, seria interessante investigar a combinação de OpenMP e MPI para aproveitar os benefícios de ambos os modelos de paralelismo em sistemas híbridos.

## Referências

- [1] Orellana, E., Materiais de slides vistos em aula
- [2] OpenMP, Disponível em: <https://www.openmp.org/wp-content/uploads/OpenMP-RefGuide-6.0-OMP60SC24-web.pdf>. Acesso em: 22 de Setembro de 2025.
- [3] Open MPI, Disponível em: <https://www.open-mpi.org/doc/v4.1/>. Acesso em: 11 de Novembro de 2025.
- [4] Brasil Escola, Disponível em: <https://brasilescola.uol.com.br/matematica/multiplicacao-matrizes.htm>. Acesso em: 22 de Setembro de 2025.
- [5] VSP-BERLIN, Disponível em: <https://svn.vsp.tu-berlin.de/repos/public-svn/publications/kn-old/strc/html/node9.html>. Acesso em: 23 de Setembro de 2025.
- [6] Wikipedia, Disponível em: [https://en.wikipedia.org/wiki/Loop\\_nest\\_optimization](https://en.wikipedia.org/wiki/Loop_nest_optimization). Acesso em: 23 de Setembro de 2025.