中山大學
SUN YAT-SEN UNIVERSITY

并行程序设计与算法实验

Lab10-CUDA 并行矩阵乘法

姓 名 李源卿

学 号 22336128

学 院 计算机学院

专 业 计算机科学与技术

2025 年 5 月 28 日

1 实验目的

- 理解 CUDA 编程模型 (Grid、Block、Thread) 及其在矩阵乘法中的应用。
- 学习 GPU 内存优化技术。

2 实验内容

- 实现基础矩阵乘法
- 优化矩阵乘法：共享内存，分块技术
- 测量不同实现的运行时间

3 实验结果与分析

3.1 不同实现方法的性能对比

表 1: 按行划分：性能对比 (时间单位: ms)

矩阵规模 (N)	线程块大小	朴素实现	基于共享内存优化	基于寄存器分块优化
512	8×8			
	16×16			
	32×32			
1024	8×8			
	16×16			
	32×32			
2048	8×8			
	16×16			
	32×32			

表 2: 按列划分：性能对比 (时间单位: ms)

矩阵规模 (N)	线程块大小	朴素实现	基于共享内存优化	基于寄存器分块优化
512	8×8			
	16×16			
	32×32			
1024	8×8			
	16×16			
	32×32			
2048	8×8			
	16×16			
	32×32			

表 3: 按数据块划分：性能对比 (时间单位: ms)

矩阵规模 (N)	线程块大小	朴素实现	基于共享内存优化	基于寄存器分块优化
512	8×8			
	16×16			
	32×32			
1024	8×8			
	16×16			
	32×32			
2048	8×8			
	16×16			
	32×32			

分析性能差异的原因：

- 结合 CUDA 内存模型和矩阵乘法原理，分析造成观察到的性能差异的可能原因。

  回答：

- 如何选择合适的线程块大小以提高占用率？

  回答

- 思考如果按不同的方式划分（例如，按行、列、数据块划分），可能会对性能和实现复杂度带来什么影响？

  回答：

- 何时应该优先考虑使用哪种存储？

  回答：