中国人民大学 信息学院

并行与分布式计算 2025 年秋季学期

实验 3: CUDA 版本矩阵矩阵乘

截止日期: 2025 年 12 月 **22** 日

1. 目标

本次课程作业通过编写 CUDA 版本的矩阵矩阵乘法（SGEMM）使同学熟悉 CUDA 编程模型。鼓励大家尝试不同的优化策略。

2. 问题描述

在数学领域中，矩阵乘法将两个矩阵进行相乘，得出另一个矩阵。矩阵运算是许多科学计算问题的基础，应用广泛。GPU 和 CPU 处理器相比，具有较高的计算能力，特别适合矩阵运算等能够进行高并发的算术操作。本次作业在 GPU 上实现矩阵 - 矩阵运算。编译代码前请先在”~/.bashrc”文件下添加如下语句：

export PATH=$PATH:/usr/local/cuda/bin

指明 nvcc 编译器路径。有问题请联系助教。要求：

1. 根据内存大小测不同规模矩阵的处理速度(GFLOPS / 秒)和带宽利用情况，并给出计算公式。
2. 请计算系统的理论峰值，如果没有达到理论峰值，尝试给出原因。

3. 优化提示

1. 考虑使用 shared memory，参考“./code/ref“下的文件（降低难度的指导性文件，建议先看懂再写自己的优化版本）。
2. 考虑线程的最优访问策略。
3. 考虑 blocking 优化。

4. 分数

1. 正确性（30%）

   请保证程序执行结果是正确的，和串行运行结果一样，允许有误差。

2. 报告书写（30%）

   页数限制 2 页，打印出来即一张纸的正反面，在有限的篇幅内说明清楚即可。

3. 实验结果（40%）

   通过分析给出的实验结果性能进行评价。如果不合理将会查看代码。

5. 提交

UniCourse+交实验报告和代码，zip 格式压缩，报告中写明具体位置及如何运行。实验报告最多 2 页。写清楚姓名、学号。需包括问题描述、方法（如何解决问题/算法），实验（实验环境、结果分析、实验代码如何运行），结论等部分。