实验五 CUDA 编程实验

**实验目的**

1. 理解 GPU 架构下数据级并行的思想。
2. 熟悉 GPU 下 CUDA 编程框架。

**实验要求**

1. 用 C 语言编写一个程序完成两个矩阵相乘。
2. 使用 CUDA 编程框架来实现两个矩阵相乘，要求以两种方法来实现：

第一种方法是：每个thread计算C中的一个元素。A的大小为[10\*blocksize][10\*blocksize] B的大小为[10\*blocksize][20\*blocksize]。思考:

* blocksize代表了什么含义?
* 限制该程序运行速度的瓶颈是什么？

第二种方法：在该程序的基础上编写一个**分块（tiled）的矩阵乘法**。思考:

* 块的大小应该如何分配才能充分利用每个block的共享显存？
* 块过大会导致什么问题?
* 如果矩阵大小任意，应该如何修改？

1. 分别比较三个程序在不同规模下的计算性能，将结果绘制成图表并进行分析。
2. 你可以再加入其它的算法来实现矩阵相乘，并进行结果分析，属于加分项。

**实验报告**

* 算法的核心**思想**
* 算法在CUDA平台上的**实现**
  + 算法的并行性，在CUDA上并行性又如何实现
  + Grid、block的设置，shared memory如何分配
  + ......
* 算法在CUDA平台上的**优化**
* 实验结果**分析**：
  + 在某种条件(Grid、block设置等等)下，Kernel函数的执行时间。
  + 在一定条件下(Grid、block设置等等)，与其他方法比较（不同算法间、CPU和GPU、优化前优化后）。