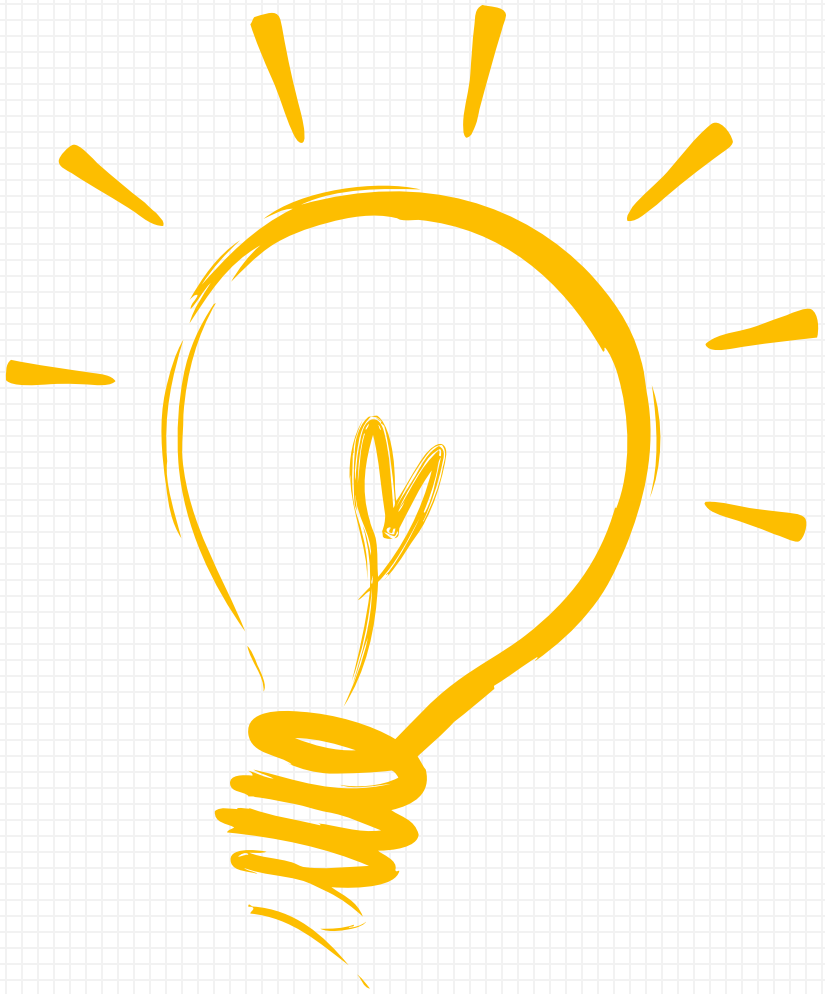
延迟隐藏

CUDA并行编程系列课程

主讲：权双



CONTENTS

延迟隐藏的概念

★ 指令延迟：在指令发出和完成之间的时钟周期被定义为指令延迟；

★ 当每个时钟周期中所有线程束调度器都有一个符合条件的线程束时，可以达到计算资源的完全利用；

★ GPU的指令延迟被其他线程束的计算隐藏称为延迟隐藏；

★ 指令可以被分为两种基本类型：

1）算数指令；

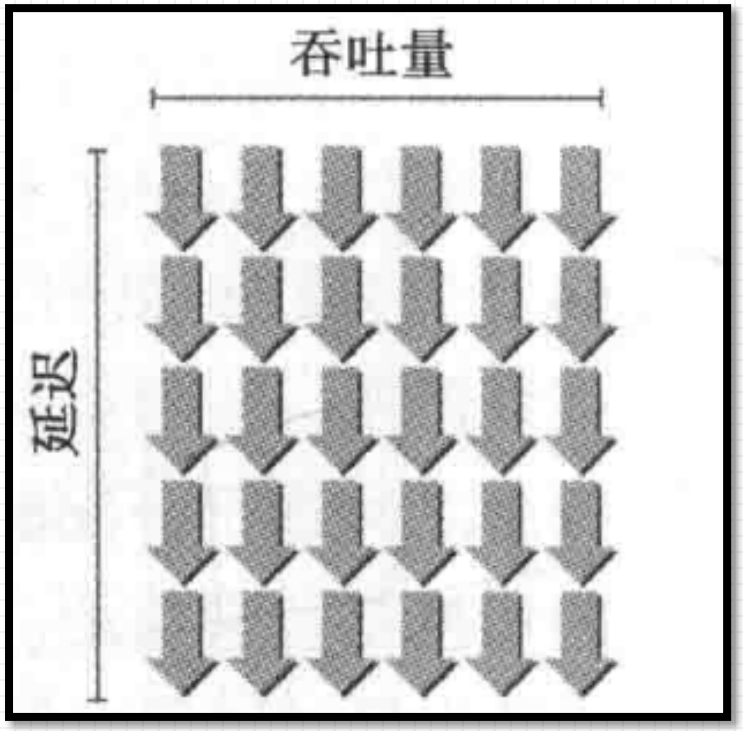
2）内存指令。

算术指令隐藏

★ 算数运算指令延迟是从开始运算到得到计算结果的时钟周期，通常为4个时钟周期；

★ 满足延迟隐藏所需的线程束数量，利用利特尔法则可以合理提供一个估计值：

所需线程束数量 = 延迟 × 吞吐量



算术指令隐藏

★ 算数运算指令延迟是从开始运算到得到计算结果的时钟周期，通常为4个时钟周期；

★ 吞吐量是SM中每个时钟周期的操作数量确定的，

★ 16-bit 所需线程束数量 = 512 / 32 = 16

★ 32-bit 所需线程束数量 = 512 / 32 = 16

★ 16-bit 所需线程束数量 = 8 / 32 = 1 （8个操作也需要1个线程束）

	指令延迟（周期）	吞吐量（操作/周期）	指令操作数量（操作）
16-bit floating-point add, multiply, multiply-add	4	128	512
16-bit floating-point add, multiply, multiply-add	4	128	512
16-bit floating-point add, multiply, multiply-add	4	2	8

计算能力8.9

★ 提升算数指令并行性方法：

1）线程中更多独立指令；

2）更多并发线程

内存指令隐藏

★ 内存访问指令延迟是从命令发出到数据到达目的地的时钟周期，通常为400~800个时钟周期；

★ 对内存操作来说，其所需的并行可以表示为在每个时钟周期内隐藏内存延迟所需的字节数；

指令延迟（周期）	BandWidt(G/S)	BandWidth(B/cycle)	GPU内存频率（GHz）	内存操作字节数量（KB）
800	504.2	50	10.0145	39

★ 504.2G/s ÷10.0145GHz ≈ 50B/cycle
800 × 50 ÷ 1024 = 39KB

★ 假设每个线程都把一个浮点数（4字节）从全局内存移动到SM中进行计算，则至少需要10000线程或者313个线程束来隐藏所有内存延迟；

39KB ÷4/线程 ≈ 10000个线程
10000个线程 ÷ 32个线程/线程束 ≈ 313个线程束

https://www.techpowerup.com/gpu-specs/geforce-rtx-4070.c3924

THANKS

谢谢聆听