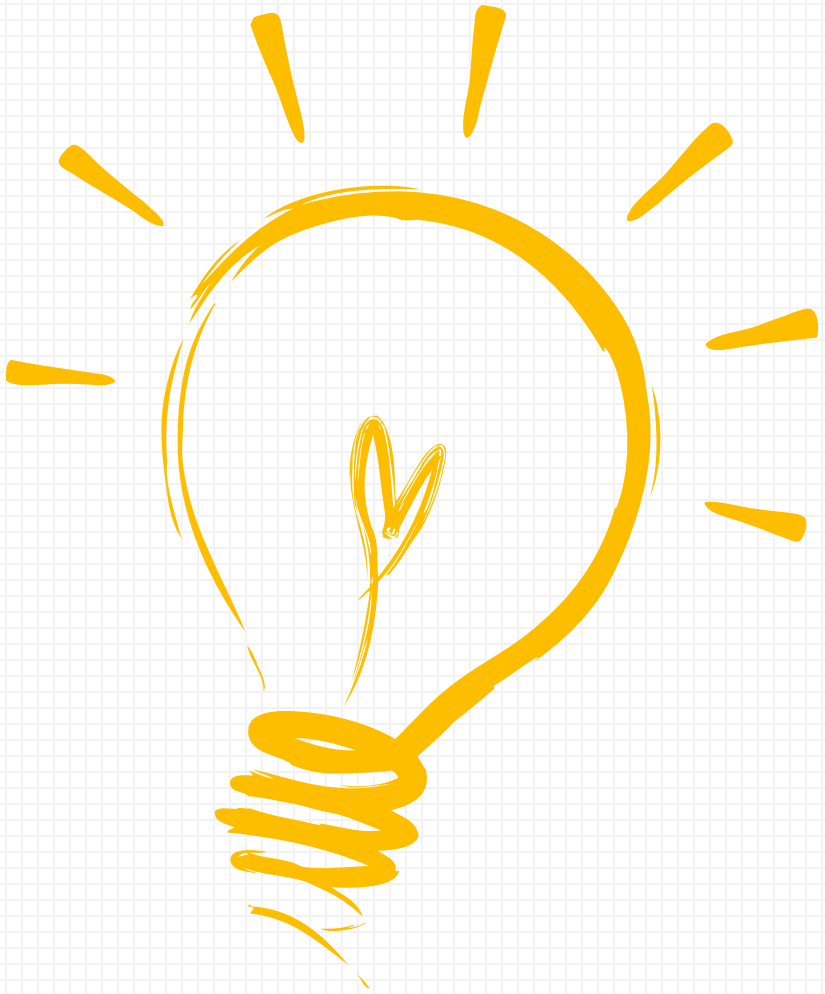
计算资源分配

CUDA并行编程系列课程

主讲：权双

CONTENTS



线程执行资源分配

★ 线程束本地执行上下文主要资源组成：

1）程序计数器；

2）寄存器；

3）共享内存；

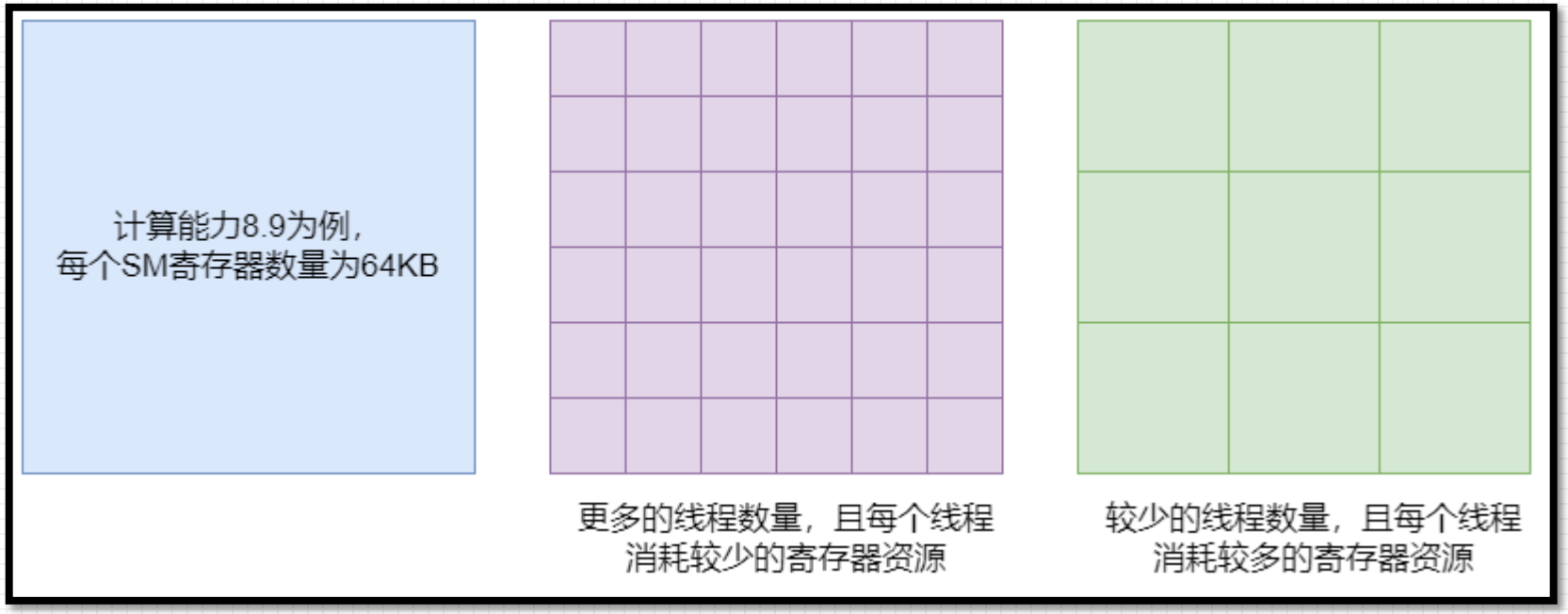
★ SM处理的每个线程束计算所需的计算资源属于片上（on-chip）资源，因此从一个执行上下文切换到另一个执行上下文是没有时间损耗的。

★ 对于一个给定的内核，同时存在于同一个SM中的线程块和线程束数量取决于在SM中可用的内核所需寄存器和共享内存数量。

寄存器对线程数目的影响

★ 每个线程消耗的寄存器越多，则可以放在一个SM中的线程束就越少；

★ 如果减少内核消耗寄存器的数量，SM便可以同时处理更多的线程束；

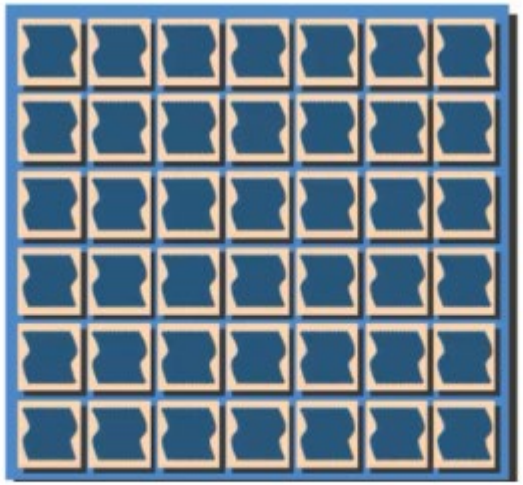


共享内存对线程块数量的影响

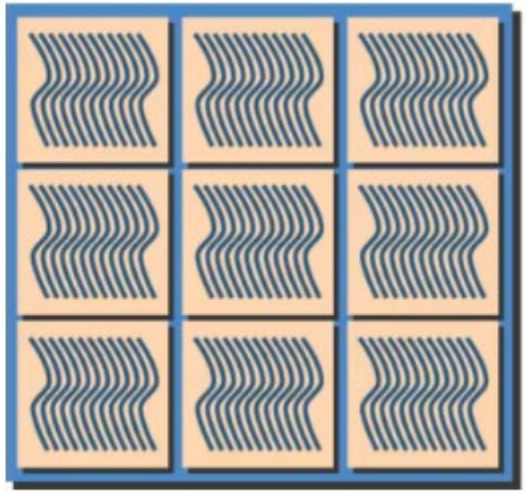
★ 一个线程块消耗的共享内存越多，则在一个SM中可以同时处理的线程块就会变少；

★ 如果每个线程块使用的共享内存数量变少，那么可以同时处理更多的线程块。

以计算能力8.9为例，
每个SM共享内存大小为100KB



更多的线程块，每个线程块
使用更少的共享内存



较少的线程块，每个线程块
使用更多的共享内存

SM占有率

★ 当计算资源（如寄存器和共享内存）已分配给线程块时，线程块被称为活跃的块，线程块所包含的线程束被称为活跃的线程束，活跃线程束可分为以下3种类型：

1）选定的线程束；

2）阻塞的线程束；

3）符合条件的线程束。

★ 占用率是每个SM中活跃的线程束占最大线程束的比值：

占用率=活跃线程束数量/最大线程束数量

计算能力	8.9
GPU型号	RTX 4070
SM数量	48
SM寄存器数量	64K
SM共享内存上限	100KB
单线程块共享内存上限	99KB
SM中最多驻留线程块数量	24
SM中最多驻留线程数量	1536

SM占有率

★ 计算能力8.9为例：

1）一个SM最多拥有的线程块个数为Nb=24；

2）一个SM最多拥有的线程个数为Nt=1536；

★ 并行性规模足够大（即核函数执行配置中定义的总线程足够多）的前提下分析SM占有率：

1）寄存器和共享内存使用很少的情况，线程块不小于64（Nt/Nb）时，可以获得100%的占有率；

2）有限寄存器对占有率的影响，当在SM上驻留最多的线程1536个，核函数中每个线程最多使用42个寄存器；

3）有限共享内存对占有率的影响，若线程块大小定义为64，每个SM需要激活24个线程块才能拥有1536个线程，达到100%的利用率每个线程块可分配4.16KB的共享内存。

★ 注意：如果一个线程块需要使用的共享内存超过了99KB，会导致核函数无法启动。

SM占有率

★ 网格和线程块大小的准则：

1）保持每个线程块中线程数量是线程束大小的倍数；
2）线程块不要设计的太小；
3）根据内核资源调整线程块的大小；
4）线程块的数量要远远大于SM的数量，保证设备有足够的并行；

THANKS

谢谢聆听