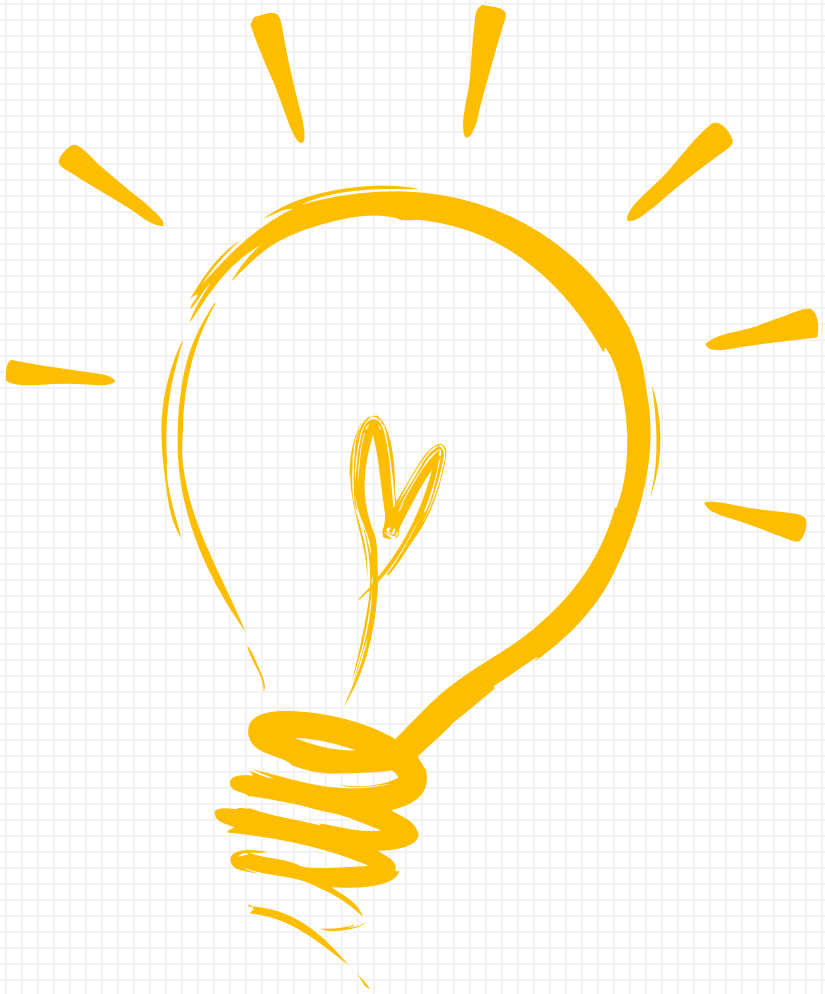
GPU硬件资源

CUDA并行编程系列课程

主讲：权双

CONTENTS

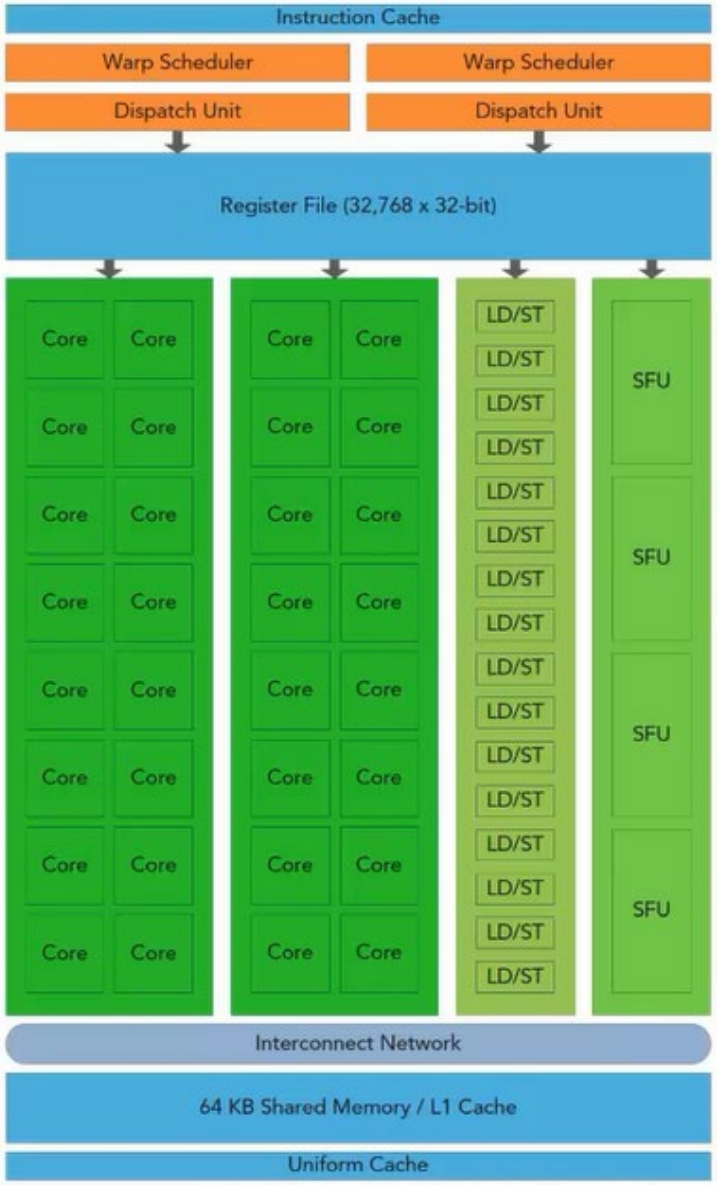


01 流多处理器--SM

02 线程模型与物理结构

03 线程束

流多处理器--SM



★ GPU并行性依靠流多处理器SM（streaming multiprocessor）来完成

★ 一个GPU是由多个SM构成的，Fermi架构SM关键资源如下：

1、CUDA核心（CUDA core）

2、共享内存/L1缓存（shared memory/L1 cache）

3、寄存器文件（RegisterFile）

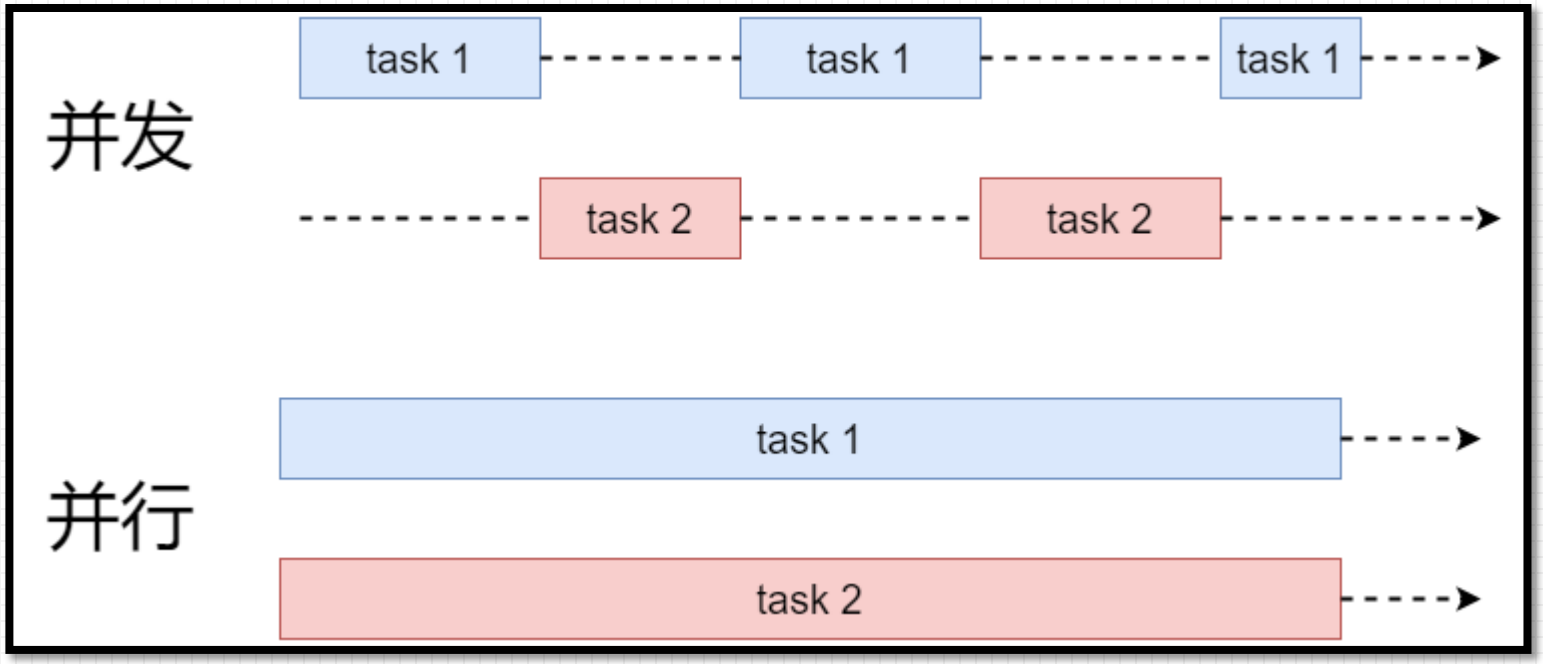
4、加载和存储单元（Load/Store Units）

5、特殊函数单元（Special Function Unit）

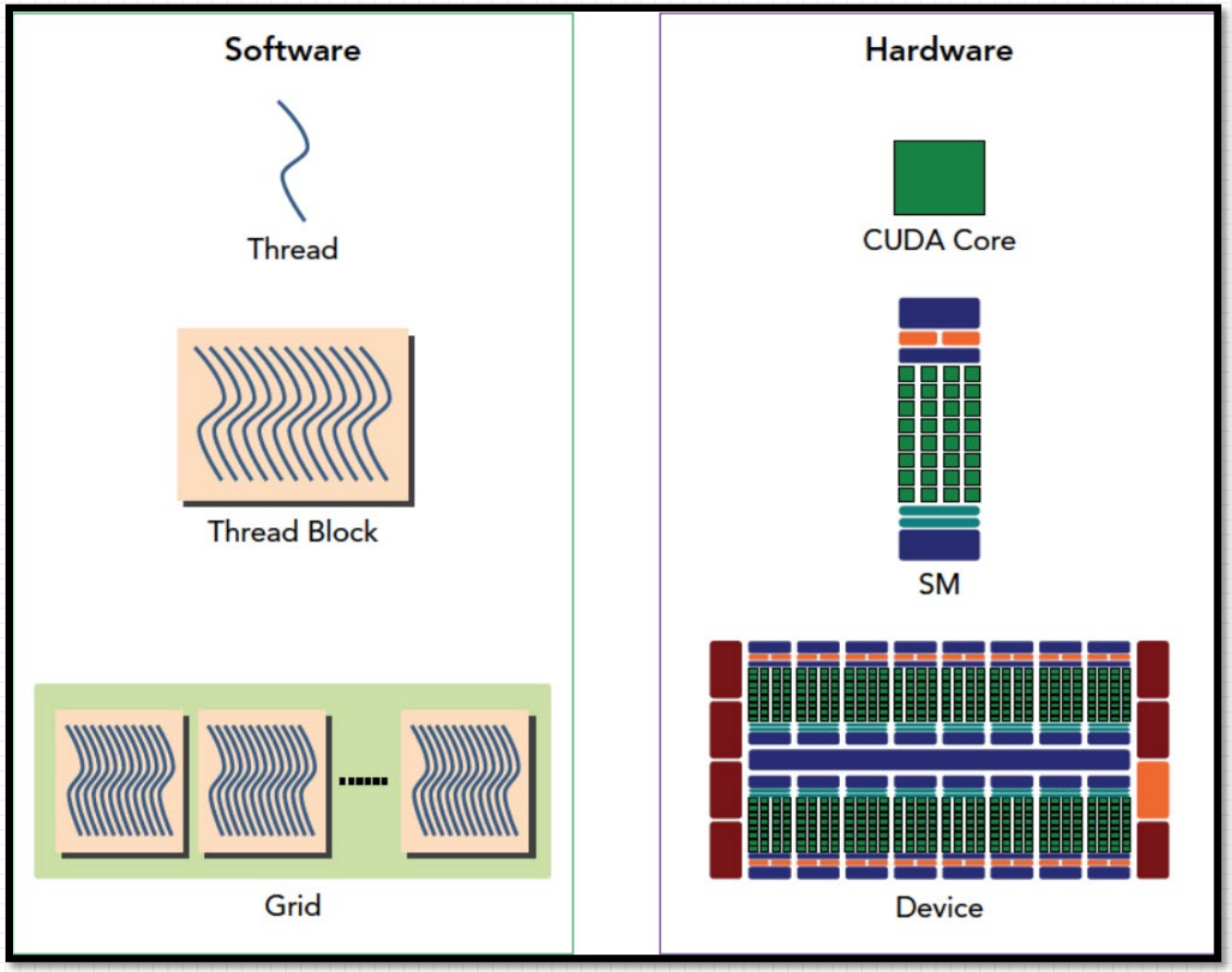
6、Warps调度（Warps Scheduler）

流多处理器--SM

- ★ GPU中每个SM都可以支持数百个线程并发执行
- ★ 以线程块block为单位，向SM分配线程块，多个线程块可被同时分配到一个可用的SM上
- ★ 当一个线程块被分配好SM后，就不可以在分配到其他SM上了



线程模型与物理结构



★ 左图线程模型，是在逻辑角度进行分析

★ 线程模型可以定义成千上万个线程

★ 网格中的所有线程块需要分配到SM上进行执行

★ 线程块内的所有线程分配到同一个SM中执行，但是每个SM上可以被分配多个线程块

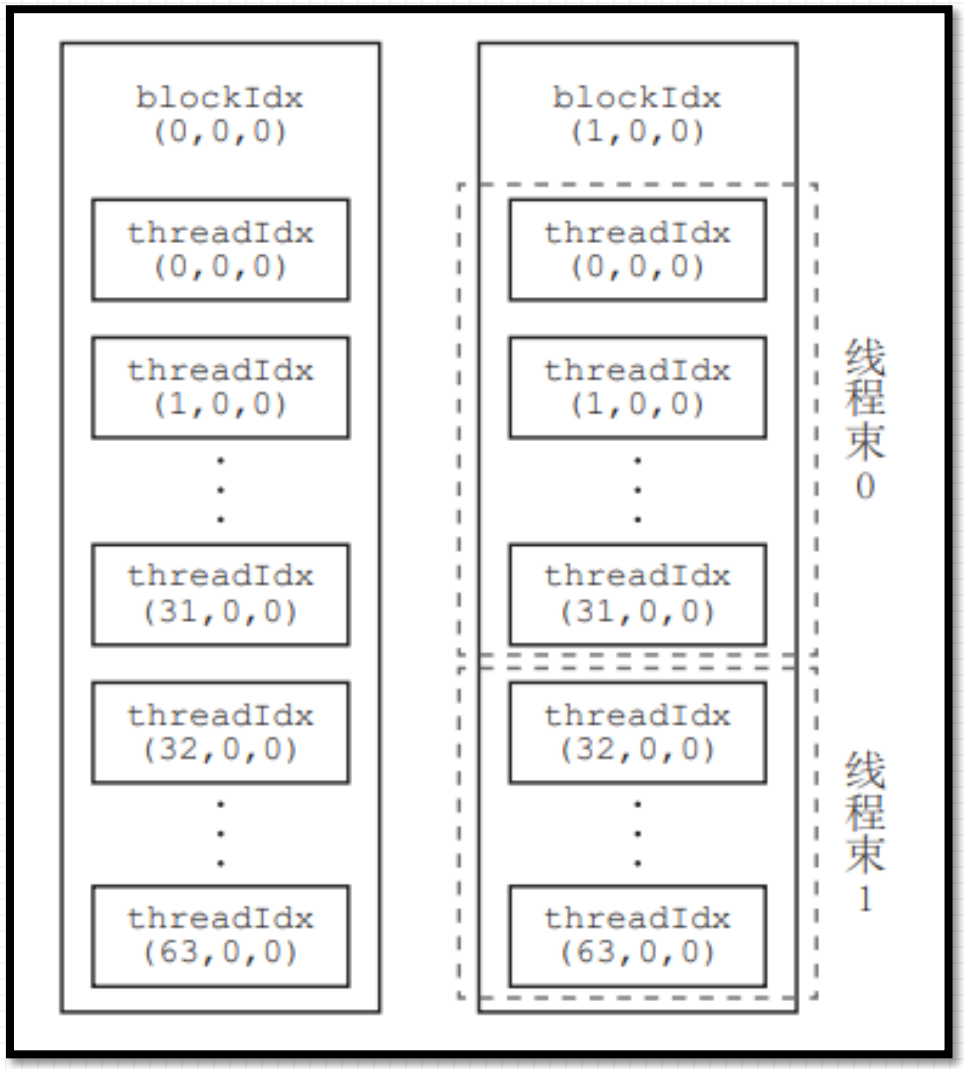
★ 线程块分配到SM中后，会以32个线程为一组进行分割，每个组成为一个wrap

★ 右图物理结构，是在硬件角度进行分析，因为硬件资源是有限的，所以活跃的线程束的数量会受到SM资源限制。

线程束

★ 什么是线程束?

CUDA 采用单指令多线程SIMT架构管理执行线程，每32个为一组，构成一个线程束。
同一个线程块中相邻的 32个线程构成一个线程束
具体地说，一个线程块中第 0 到第 31 个线程属于第 0 个线程束，第 32 到第 63 个线程 属于第 1 个线程束，依此类推。



线程束

★ 每个线程束中只能包含同一线程块中的线程

★ 每个线程束包含32个线程

★ 线程束是GPU硬件上真正的做到了并行

定义线程块（128，1，1）

定义线程块（112，1，1）

★ 线程束数量 = ceil（线程块中的线程数/32）

THANKS

谢谢聆听