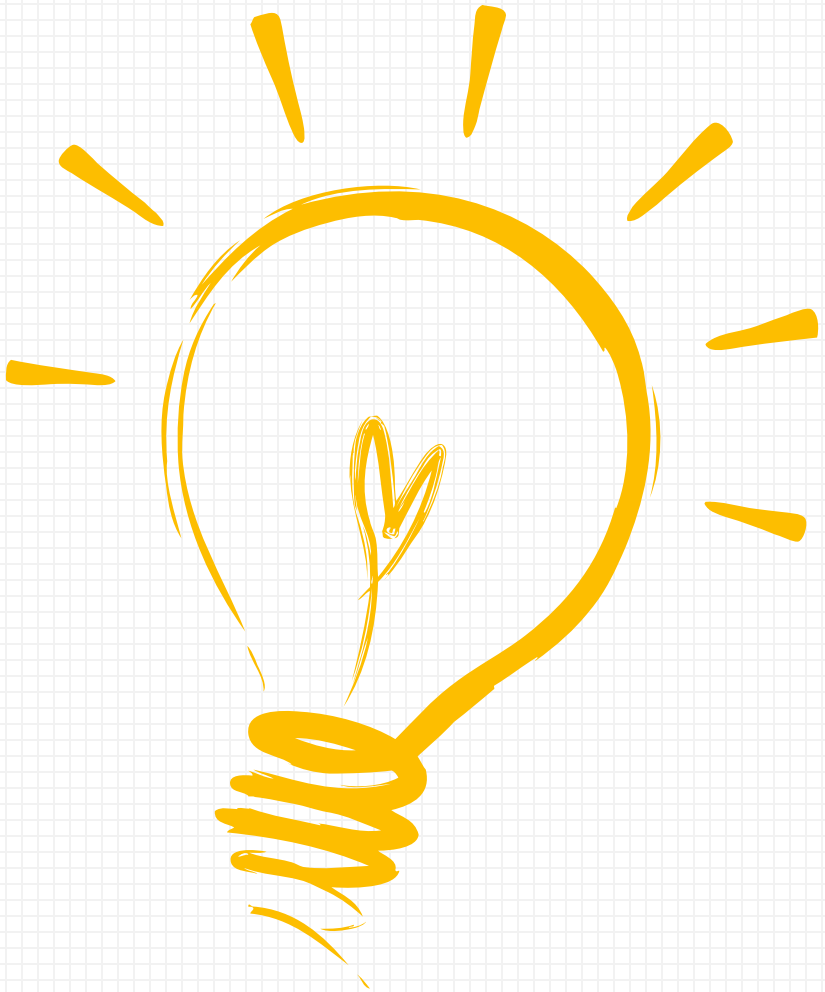
共享内存

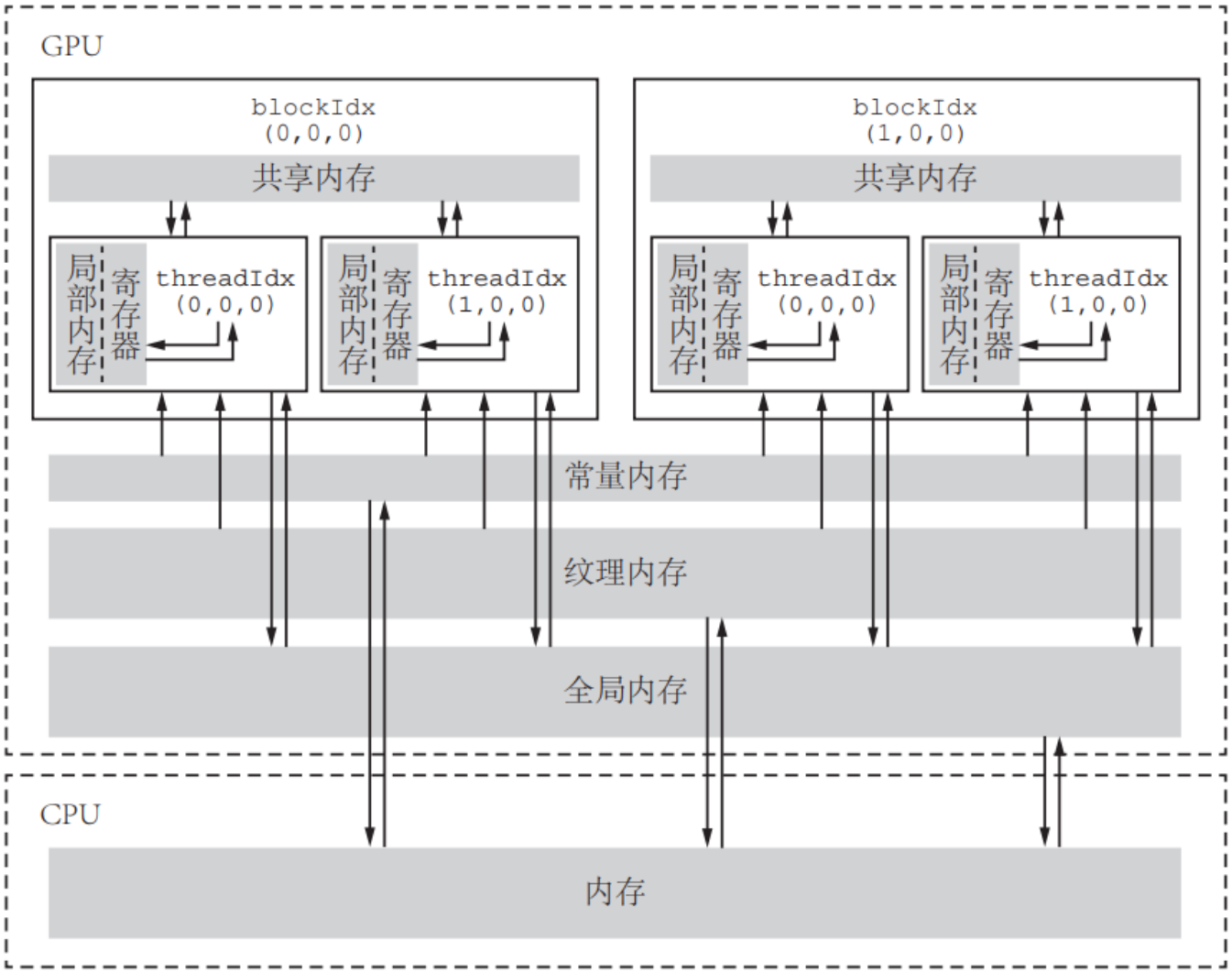
CUDA并行编程系列课程

主讲：权双

CONTENTS



共享内存作用



- 共享内存在片上（on-chip），与本地内存和全局内存相比具有更高的带宽和更低的延迟；
- 共享内存中的数据在线程块内所有线程可见，可用线程间通信，共享内存的生命周期也与所属线程块一致；
- 使用__shared__修饰的变量存放于共享内存中，共享内存可定义动态与静态两种；
- 每个SM的共享内存数量是一定的，也就是说，如果在单个线程块中分配过度的共享内存，将会限制活跃线程束的数量；
- 访问共享内存必须加入同步机制：
  线程块内同步　void __syncthreads();

共享内存作用

	Compute Capability													
Technical Specifications	5.0	5.2	5.3	6.0	6.1	6.2	7.0	7.2	7.5	8.0	8.6	8.7	8.9	9.0
Maximum amount of shared memory per SM	64 KB	96 KB	64 KB		96 KB	64 KB	96 KB		64 KB	164 KB	100 KB	164 KB	100 KB	228 KB
Maximum amount of shared memory per thread block[32]	48 KB						96 KB	96 KB	64 KB	163 KB	99 KB	163 KB	99 KB	227 KB

★ 不同计算能力的架构，每个SM中拥有的共享内存大小是不同的；

★ 每个线程块使用的最大数量不同架构是不同的，计算能力8.9是100K；

共享内存作用

★ 经常访问的数据由全局内存（global memory）搬移到共享内存（shared memory），提高访问效率；

★ 改变全局内存访问内存的内存事务方式，提高数据访问的带宽。

静态共享内存

★ 共享内存变量修饰符：__shared__

★ 静态共享内存声明： __shared__ float tile[size, size];

★ 静态共享内存作用域：

1、核函数中声明，静态共享内存作用域局限在这个核函数中；

2、文件核函数外声明，静态共享内存作用域对所有核函数有效。

★ 静态共享内存在编译时就要确定内存大小

共享内存和一级缓存划分

__host__cudaError_t cudaFuncSetCacheConfig (const void *func, cudaFuncCache cacheConfig)

Sets the preferred cache configuration for a device function.

Parameters

func
- Device function symbol

cacheConfig
- Requested cache configuration

Returns

cudaSuccess, cudaErrorInvalidDeviceFunction

The supported cache configurations are:

- cudaFuncCachePreferNone: no preference for shared memory or L1 (default)
- cudaFuncCachePreferShared: prefer larger shared memory and smaller L1 cache
- cudaFuncCachePreferL1: prefer larger L1 cache and smaller shared memory
- cudaFuncCachePreferEqual: prefer equal size L1 cache and shared memory

★ 在L1缓存和共享内存使用相同硬件资源的设备上，可通过cudaFuncSetCacheConfig运行时API指定设置首选缓存配置;

★ func必须是声明为__global__的函数;

★ 在L1缓存和共享内存大小固定的设备上，此设置不起任何作用;

THANKS

谢谢聆听