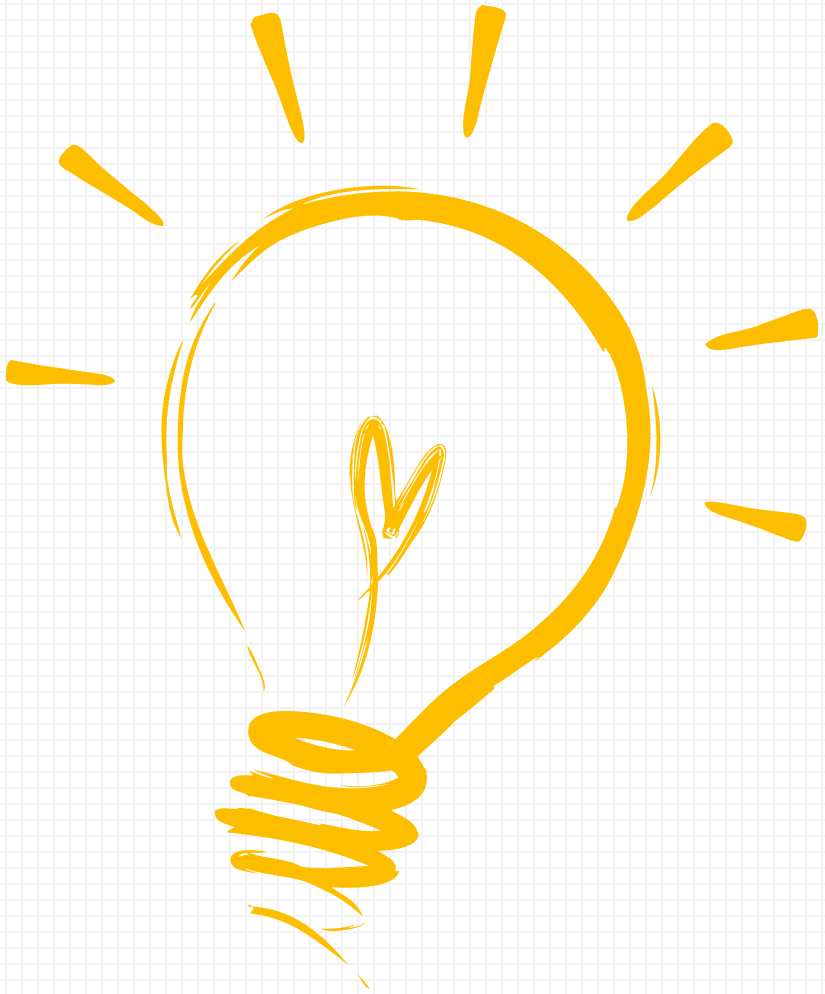
CUDA内存模型概述

CUDA并行编程系列课程

主讲：权双

CONTENTS



内存结构层次特点

★ 局部性原则：

时间局部性

空间局部性

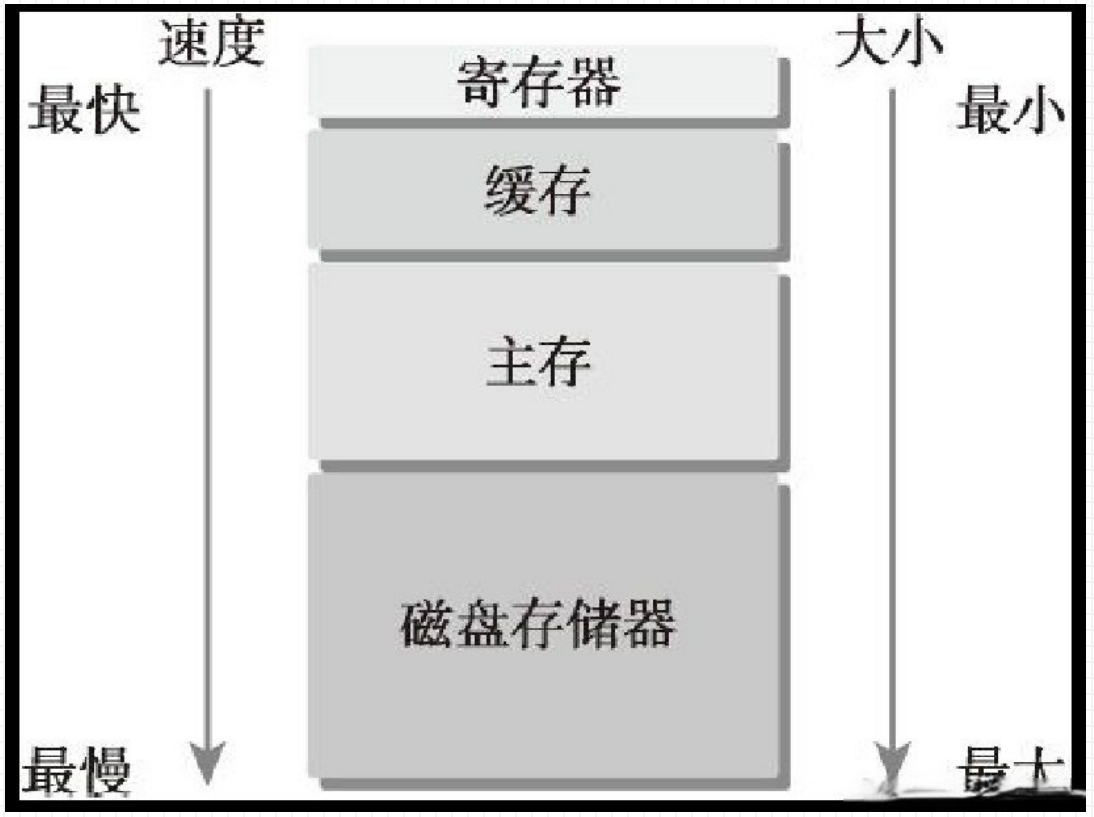
★ 如图，底部存储器特点：

1、更低的每比特位平均成本

2、更高的容量

3、更高的延迟

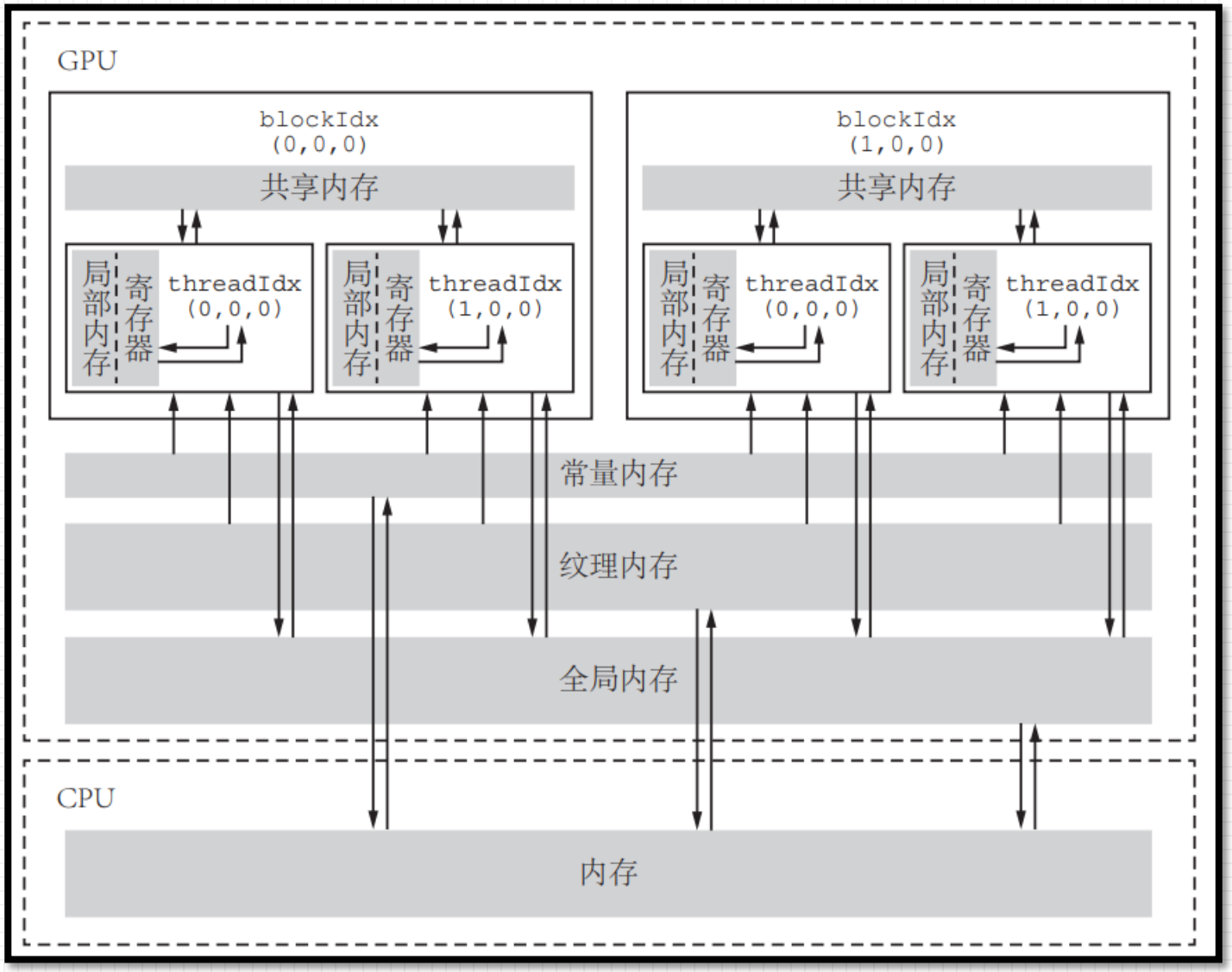
4、更低的处理器访问频率



★ CPU和GPU主存采用DRAM（动态随机存取存储器）

低延迟的内存采用SRAM（静态随机存取存储器）

CUDA内存模型



- ★ 寄存器（register）
- ★ 共享内存（shared memory）
- ★ 本地内存（local memory ）
- ★ 常量内存（constant memory ）
- ★ 纹理内存（texture memory）
- ★ 全局内存（global memory）

CUDA内存模型

CUDA内存和它们的主要特征：
1、物理位置　　2、访问权限　　3、可见范围　　4、生命周期

内存类型	物理位置	访问权限	可见范围	生命周期
全局内存	在芯片外	可读可写	所有线程和主机端	由主机分配与释放
常量内存	在芯片外	仅可读	所有线程和主机端	由主机分配与释放
纹理和表面内存	在芯片外	一般仅可读	所有线程和主机端	由主机分配与释放
寄存器内存	在芯片内	可读可写	单个线程	所在线程
局部内存	在芯片外	可读可写	单个线程	所在线程
共享内存	在芯片内	可读可写	单个线程块	所在线程块

THANKS

谢谢聆听