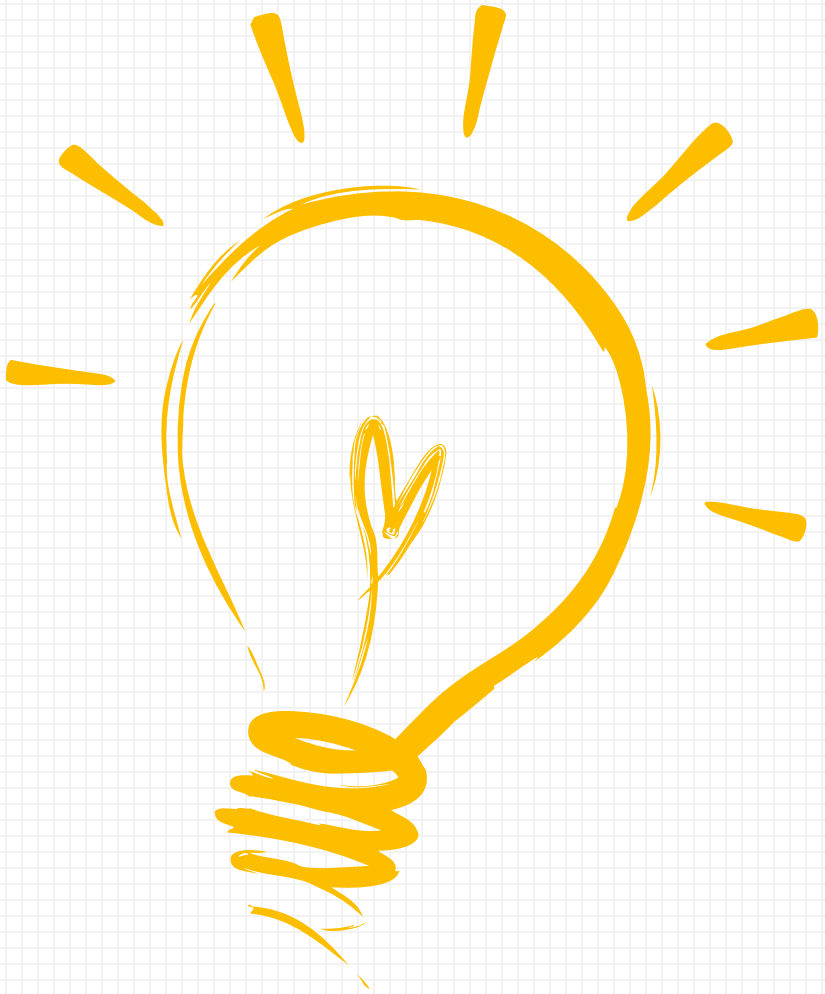
GPU缓存

CUDA并行编程系列课程

主讲：权双



CONTENTS

01 GPU缓存种类

02 GPU缓存作用

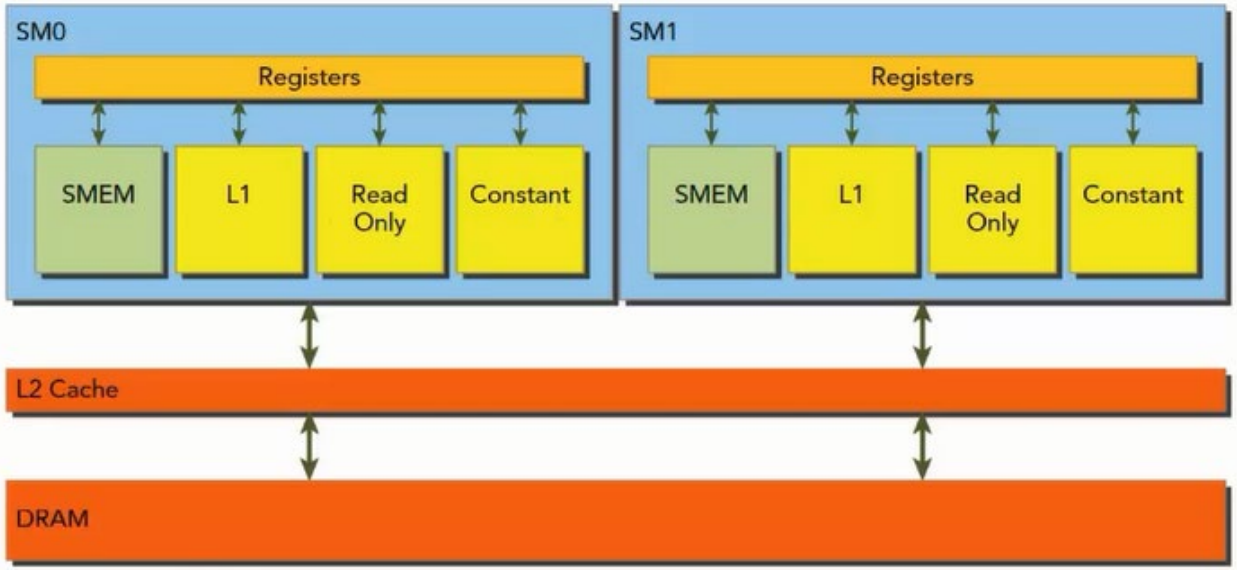
03 L1缓存查询与设置

04 L1缓存与共享内存

GPU缓存种类

- ★ 一级缓存（L1）；
- ★ 二级缓存（L2）；
- ★ 只读常量缓存；
- ★ 只读纹理缓存；

GPU缓存作用



- ★ GPU缓存是不可编程的内存；
- ★ 每个SM都有一个一级缓存，所有SM共享一个二级缓存；
- ★ L1缓存和L2缓存用来存储本地内存（local memory）和全局内存（global memory）的数据，也包括寄存器溢出的部分；
- ★ 在GPU上只有内存加载可以被缓存，内存存储操作不能被缓存；
- ★ 每个SM有一个只读常量缓存和只读纹理缓存，它们用于在设备内存中提高来自各自内存空间内的读取性能。

L1缓存查询与设置

★ GPU全局内存是否支持L1缓存查询指令：

cudaDeviceProp::globalL1CacheSupported

★ 默认情况下，数据不会缓存在统一的L1/纹理缓存中，但可以通过编译指令启用缓存：

开启：-Xptxas -dlcm=ca

除了带有禁用缓存修饰符的内联汇编修饰的数据外，所有读取都将被缓存;

开启：-Xptxas -fscm=ca

所有数据读取都将被缓存。

L1缓存与共享内存

★ 计算能力为8.9的显卡为例：

1）统一数据缓存大小为128KB，统一数据缓存包括共享内存、纹理内存和L1缓存；

2）共享内存从统一的数据缓存中分区出来，并且可以配置为各种大小，共享内存容量可设置为0，8，16，32，64和100KB，剩下的数据缓存用作L1缓存，也可由纹理单元使用；

3）L1缓存与共享内存大小是可以进行配置的，但不一定生效，GPU会自动选择最优的配置。

L1缓存与共享内存

- 伏特架构（计算能力7.0）：统一数据缓存大小为128KB，共享内存容量可以设置为0、8、16、32、64或96KB；
- 图灵架构（计算能力7.5）：统一数据缓存大小为96KB，共享内存容量可以设置为32KB或64KB；
- 安培架构（计算能力8.0）：统一数据缓存大小为192KB，共享内存容量可以设置为0、8、16、32、64、100、132或164KB.

THANKS

谢谢聆听