CUDA内存模型

1 概述

CUDA中可用内存被分为：

- 全局内存 global memory
- 共享内存 shared memory
- 局部内存 local memory
- 只读内存
  - 常亮内存
  - 纹理内存

学习这些内存的使用和GPU中线程模等同，本文将给出它们的介绍。

2 性能比较

使用不同内存，程序性能存在明显差异，如何度量某段CUDA程序在不同内存上的性能差异？或许你能够想到使用CPU时间，如clock_t，但是存在问题，因为整个CPU时间包含线程调度、切换等时间，在时间度量上存在问题。在CUDA中提供了专门的机制来度量时间，称为**事件**。坦白地说，CUDA中事件本质就是时间，下面将给出事件的用法。

```
cudaEvent_t start, stop;   //创建两个CUDA事件变量
cudaEventCreate(&start);   //创建start事件
cudaEventCreate(&stop);   //创建stop事件
cudaEventRecord(start,0); // 记录起始时间
/**
GPU CUDA代码
*/
cudaEventRecord(stop,0); //记录结束时间
```

```
cudaEventSynchronize(stop); // 同步所有的线程，保证在改函数之前，所有的GPU操作全部执行完毕
float elapsed;
cudaEventElapsedTime(&elapsed, start, stop); //得到执行时间，打印出执行时间即可
cudaEventDestroy(start);
cudaEventDestroy(stop);
```

3 参考资料

1. 图示和代码结合给出不同内存的使用 https://www.jianshu.com/p/7a8fe1aefd4e
2. 从架构的角度理解GPU内存模型 https://www.face2ai.com/CUDA-F-5-1-CUDA%E5%85%B1%E4%BA%AB%E5%86%85%E5%AD%98%E6%A6%82%E8%BF%B0/