【CUDA 基础】6.0 流和并发

2018-06-10 | CUDA | Freshman | 0

Abstract: 本文是第六章的概述，本章也是Freshman的最后一个章节。

Keywords: 流，事件，网格级并行，同步机制，NVVP

流和并发

本文是Freshman系列的最后一篇，考虑到接下来要说的是比较高级的内容，所以把其划分到下个系列中，作为进阶内容介绍，所以本章是初级阶段的收尾。

本章内容

本章主要介绍下面内容：

- 理解流和事件的本质
- 理解网格级并发
- 重叠内核执行和数据传输
- 重叠CPU执行和GPU执行
- 理解同步机制
- 调整流的优先级
- 注册设备回调函数
- 通过NVIDIA可视化性能分析器显示应用程序执行时间轴

一般来说CUDA程序有两个几倍的并发：

1. 内核级并行
2. 网格级并行

我们前面说有都是在研究内核级别的并行，通过同一内核多线程的并行来完成并行计算，提高内核级别并行我们前面用了基本所有的篇幅介绍了以下三种途径：

1. 编程模型
2. 执行模型
3. 内存模型

这三个角度是优化内核级并行的最主要也是最基础的方法，更高级的方法虽然高级但是提升效率幅度绝没有这三种基础角度来的更有效率。
本章我们在内核之上研究并行，也就是多个内核的并行，这在一个完整应用中是很常见的，实际中的应用程序多半都不是单个内核的，多个内核最大程度的并行也就是最大限度的使用GPU设备，是提高整个应用效率的关键。

总结

本章我们考虑只在一个设备上并行内核，使用CUDA流实现网格级并发，还会使用NVVP显示内核并行执行可视化。

本文作者： 谭升