<<高性能CUDA应用设计与开发>>

图书基本信息

书名：<<高性能CUDA应用设计与开发>>

13位ISBN编号：9787111404460

10位ISBN编号：7111404467

出版时间：2013-1-1

出版时间：机械工业出版社华章公司

作者：Rob Farber

页数：271

译者：于玉龙,唐堃

前言

译者序序言作为一种革新性的技术，GPU近年来在科学计算领域备受关注。
这项技术已有大量的科学应用，并在性能与能效方面取得了显著提升。
这些应用大多由那些致力于使用GPU的先驱者开发和使用。
最近，这项技术面临一个关键问题：它能否普遍适用于科学计算领域中各种各样的算法，以及能否为更广泛的人群而不仅是那些先驱者使用。
软件开发是阻碍这项技术得以广泛应用的关键，其中包括大规模并行CUDA代码的编写与优化、新的性能与正确性分析工具的使用、CUDA支持库的使用以及对GPU硬件架构的理解。
通过书籍、教程等方式，专家们可以将他们在此领域掌握的知识和方法与其他使用者分享，这在一定程度上解决了CUDA广泛应用的难题。
本书就是这样的一本书。
在这本书中，作者详尽地解释了重要算法的实现方法，如量子化学、机器学习以及计算机视觉等。
本书不仅讲述了GPU编程的基本方法，还介绍了如何改写算法以便从GPU架构中最大化获益。
此外，这本书提供了许多案例研究用以解释与充实GPU的重要概念，如CUDA线程、GPU存储层次结构以及多GPU的扩展性（书中使用一个MPI示例例证了其扩展性可以近似线性地增至使用500个GPU）。
最后，任何一种编程语言都不能独立存在。
可以说，任何一个成功的编程语言，都伴随由强大的编译器、性能与正确性分析工具以及优化的支持库组成的整套系统环境。
这些软件开发中的实用工具是快速开发应用程序的关键。
本书的诸章节描述了如何使用CUDA编译器、调试器、性能分析工具、库以及与其他语言的互操作，本书在这些方面没有令人失望。
我享受从这本书中学到的一切，我确信您也会。
Jeffrey S.Vetter美国橡树岭国家实验室杰出研究员佐治亚理工学院教授

<<高性能CUDA应用设计与开发>>

内容概要

本书是广受推崇的系统学习高性能CUDA应用开发与设计的经典著作，是美国国家安全实验室资深高性能编程专家多年工作经验结晶，橡树岭国家实验室资深专家鼎力推荐！
本书不仅从硬件角度深入解读了CUDA的设计理念和GPGPU硬件的体系结构，而且从软件角度系统讲解了CUDA应用设计与开发的思想、方法、技巧、准则、注意事项和最佳实践。

第1章首先介绍了CUDA的核心概念和编程思想，以及构建与调试CUDA应用所需的工具和方法，然后讲解了有效提高程序性能的CPU编程准则；第2章讲解了CUDA在机器学习与优化中的核心概念与应用，并给出了完整的通用框架；第3章介绍了CUDA的性能分析工具套件以及性能分析的方法，同时讨论了PCA和NLPCA两种数据挖掘方法；第4章讲解了CUDA的执行模型，深刻揭示了GPU的工作方式和原理；第5章介绍了CUDA提供的多种GPU内存，以及各种内存的优缺点；第6章讲解了高效利用内存的技术；第7章介绍了GPU提供的多种并行方式及其应用；第8章首先讨论了多种CUDA后端设备，以及CUDA如何与Python、Java、R等高级语言交互；第9章讲解了CUDA与图形渲染混合编程；第10章讲解了在云计算和集群环境中使用CUDA的方法和技术细节；第11章介绍了CUDA在高维数据处理、力导向图、交互式工作流、量子化学等现实问题中的应用；第12章为学习CUDA设计了一个综合性的针对实时视频流的应用案例。

作者简介

Rob Farber，资深高性能编程专家，是Irish高端计算中心和美国国家实验室等权威机构的高性能编程技术顾问，同时为多家《财富》世界500强企业提供咨询服务，经验十分丰富，在该领域颇具权威和影响力。他还是一位技术作家，任职于Santa Fe学院，在《Dr. Dobb ' s Journal》、《Scientific Computing》等媒体上发表了多篇关于高性能编程的经典技术文章，深受读者喜爱。此外，他还是《财富》排名前100名的两家公司的合伙创始人。

<<高性能CUDA应用设计与开发>>

书籍目录

章节摘录

版权页： 插图： 计算能力为2.0的设备为每个流多处理器增加了L1缓存，并且在流多处理器和全局内存之间增加了一套所有流多处理器都共享的L2缓存，如图5.2所示。
5.3 L2缓存 GPU上所有流多处理器都可访问的L2缓存（二级缓存）是一种高速数据存储设备，可极大地降低设备内存子系统的工作压力。
由于L2缓存的存在，许多程序在计算能力为2.0的设备上的运行速度会变得很快。
这可归结为两个原因：在不需要CUDA编程者做任何调整的情况下，L2缓存可通过LRU方式缓存数据，这使得很多CUDA Kernel避免了全局内存的带宽瓶颈。
非规则内存访问会表现出极差的性能，而L2缓存对非规则内存访问具有明显的加速效果。
对于很多算法，程序自身的特性决定了其能否有效地部署在不具备L2缓存的计算能力为1.x的设备上。
Fermi架构的GPU提供了768KB共享的L2缓存，并保证了在各流多处理器间的一致性。
也就是说，任何线程都可以对L2缓存中的内容进行修改。
然后，任何其他的GPU线程都可以读取该地址，得到更新后的正确数据。
当然，必须使用原子操作才能保证当前数据写操作完成后再允许其他线程的读取操作。
因为早期GPGPU架构采用了不同的处理路径来处理数据的读操作和写操作，所以即使最普通的读写操作也存在执行上的风险。
特别是只读的纹理加载路径和只写的像素数据路径。
为了保证数据正确性，在早期GPU架构中，如果任何线程修改了一个被缓存的内存区域的数据，都将导致读操作路径上该内存区域的缓存数据无效，并清空已存在的缓存数据。
在Fermi架构中，流多处理器间共享的L2缓存消除了在早期GPU架构中纹理和光栅缓存中存在的这一性能瓶颈。
所有的数据加载与存储操作都经过L2缓存（包括CPU与GPU之间的数据拷贝）。
这就解释了为什么GPU和主机间的数据传输居然会影响缓存命中率和程序性能。
显然，异步的Kernel执行也会造成缓存污染。

媒体关注与评论

本书是所有希望系统学习和使用CUDA编写用于科学计算或可视化程序等并行应用的程序员的必读书之一。
本书通俗易懂，并且提供了便于读者动手实践的详细操作案例。
——Jack Dongrra　田纳西大学计算机科学分布式处理器研究院

编辑推荐

《高性能CUDA应用设计与开发:方法与最佳实践》编辑推荐：美国国家安全实验室资深专家撰写，硬件角度解读CUDA设计理念与体系结构，软件角度讲解CUDA应用开发思想、方法、技巧与准则。

名人推荐

本书是所有希望系统学习和使用CUDA编写用于科学计算或可视化程序等并行应用的程序员的必读书之一。
本书通俗易懂，并且提供了便于读者动手实践的详细操作案例。
——Jack Dongrra 田纳西大学计算机科学分布式处理器研究院

<<高性能CUDA应用设计与开发>>

版权说明