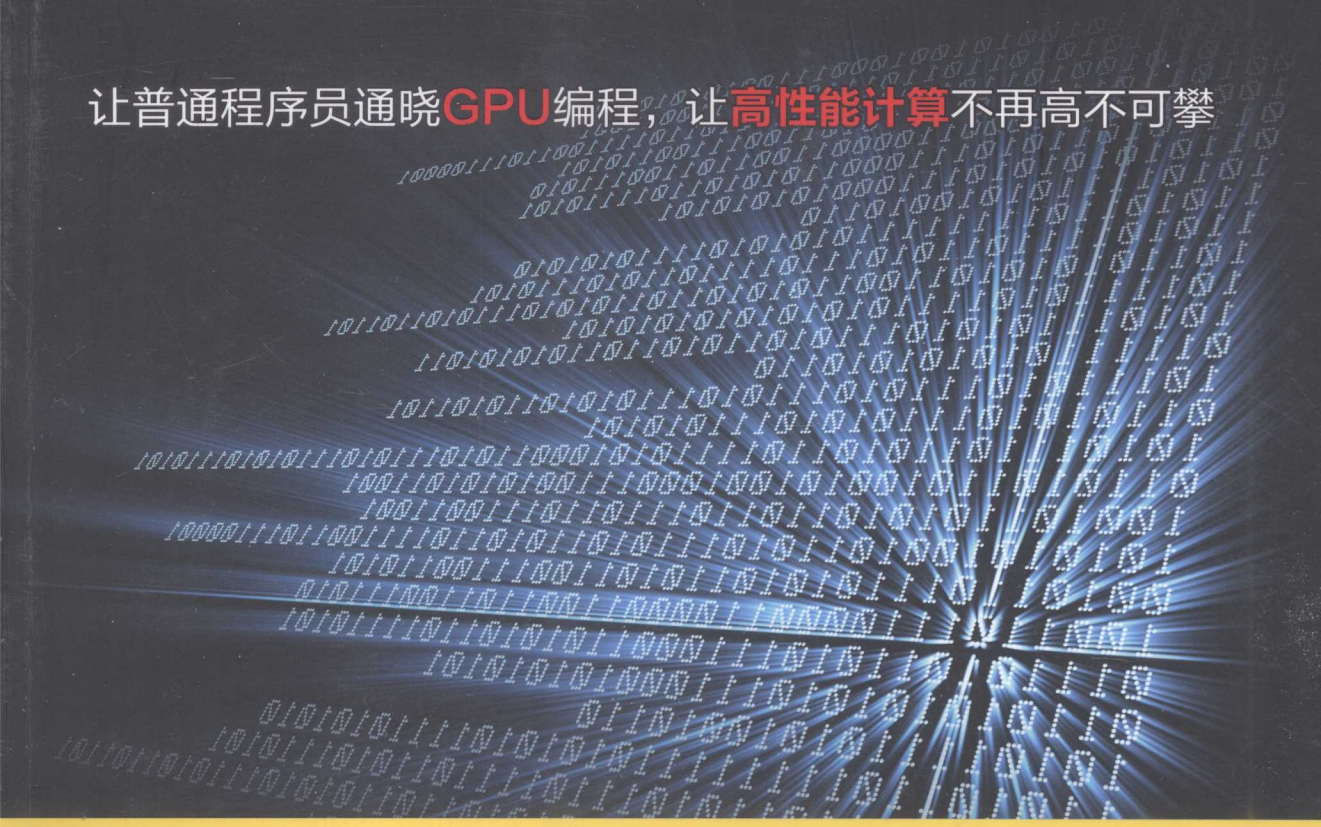
让普通程序员通晓GPU编程，让高性能计算不再高不可攀



GPU Programming and Code Optimization
High Performance Computing for the Masses

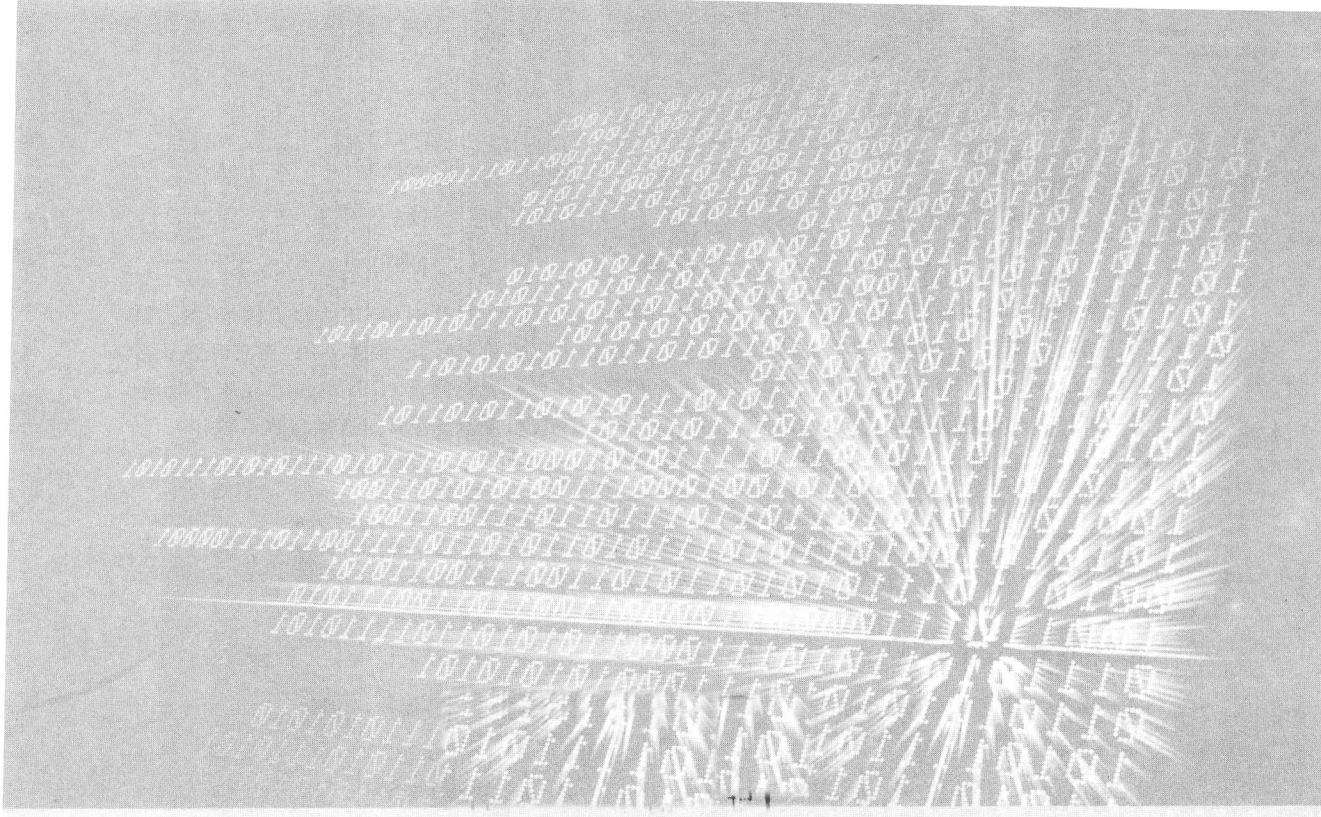
GPU编程与优化

——大众高性能计算

方民权　张卫民
方建滨　周海芳　著
高　畅

- ◎ 系统全面的GPU知识体系
- ◎ 实践引导的有效优化方法
- ◎ 切中要害的性能影响因素
- ◎ 通俗翔实的异构协同并行
- ◎ 循序渐进的优化实例解析
- ◎ 精炼真实的GPU性能测评

清華大學出版社



GPU Programming and Code Optimization

High Perform ng for the Masses

GPU编程与优化

——大众高性能计算

方民权　张卫民
方建滨　周海芳　著
高　畅

清華大學出版社
北京

内 容 简 介

本书第一篇系统地介绍 GPU 编程的相关知识，帮助读者快速入门，并构建 GPU 知识体系；第二篇和第三篇给出大量实例，对每个实例进行循序渐进的并行和优化实践，为读者提供 GPU 编程和优化参考；第四篇总结影响 GPU 性能的关键要素（执行核心和存储体系），量化测评 GPU 的核心性能参数，给出 CPU/GPU 异构系统上覆盖完全的各种混合并行模式及其实践，帮助读者真正透彻理解 GPU。

本书适合作为计算机及相关专业的教材，也可作为 GPU 程序开发人员和科研人员的参考书。

图书在版编目(CIP)数据

GPU 编程与优化：大众高性能计算/方民权等著. --北京：清华大学出版社，2016
ISBN 978-7-302-44642-2

Ⅰ. ①G…　Ⅱ. ①方…　Ⅲ. ①图像处理－程序设计　Ⅳ. ①TP391.41

中国版本图书馆 CIP 数据核字(2016)第 179437 号

责任编辑：白立军
封面设计：杨玉兰
责任校对：李建庄
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954
印 刷 者：清华大学印刷厂
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：27.25　　**字　　数**：660 千字
版　　次：2016 年 9 月第 1 版　　**印　　次**：2016 年 9 月第1次印刷
印　　数：1～2000
定　　价：59.00 元

产品编号：070097-01

前　言

多核与众核异构平台因其超强的浮点运算能力而成为当前高性能计算领域的新贵。2010年以来，已有3台异构超级计算机夺魁TOP 500，分别是搭载CPU/GPU异构系统的天河1A和泰坦超级计算机、搭载CPU/MIC异构系统的天河2号超级计算机。在这两类主流的多核与众核异构平台中，CPU/GPU异构平台在性价比、能耗比等方面表现尤为突出，例如，在Green500前10中有9台采用了这种架构。对于高性能计算用户而言，CPU/GPU异构系统无疑是一个良好的选择。

另一方面，当前PC已普遍装备GPU（独立显卡），使得这种CPU/GPU异构系统的硬件平台随处可见。尽管这类平台的GPU工作重心是游戏娱乐而非浮点计算，但在一些精度要求不高的领域仍然优势显著。此外，基于GPU编程的工具包是免费的，因此可用较低的成本构建合适的CPU/GPU异构并行平台。由于显卡的普及，CPU/GPU异构并行程序也能在几乎所有的PC中广泛应用。

然而，仅有硬件是没有应用价值的，异构系统上的程序开发是实现异构系统价值的直接且唯一的步骤。但是，异构并行软件开发面临着巨大挑战，主要包括异构数据通信、基于GPU体系结构的编程与优化、多编译器的联合编译等，具体到实践则难度更大。编写本书的目的就是辅助用户解决这些GPU异构并行软件开发的难题。

目前市面上已有很多GPU编程书籍，其中一些已经论述相当全面，为什么还要撰写本书呢？作为一名有多年开发经验的GPU程序员，阅读这些书籍总感觉有些不足。

首先，对于刚接触GPU的开发者，由于所要认识的GPU体系结构与常用的CPU体系结构差异巨大，相关的理论知识较难理解，而已有的一些论著为了增强理论性和学术性还对相关理论知识进行了抽象提升，因而不够通俗易懂，即使是GPU编程老手也未必能完全读懂。而本书将从GPU程序员的角度出发，通俗易懂地阐述GPU编程与优化相关的理论知识。

其次，从GPU理论到编程优化实践的过渡是非常关键的，但目前市面上的书籍重理论而轻实践。仅有理论而缺乏实践和直观实践效果，难以对程序员读者产生直接价值。况且很多没有实践论证的理论知识未必正确。本书将紧密结合理论和实践，并试图从实践中总结理论知识，从而帮助读者更好地理解。

此外，程序性能优化是GPU编程的重中之重。当前GPU书籍中提及了大量关于优化方法的理论知识，但很少针对每种优化方法和策略给出应用实例（即使有也可能只有一个，借鉴范围不够广），更没有针对某一个实例进行系统性循序渐进的优化。事实上，对于新手而言，优化时最大的困惑就是知道优化方法却不知道用到哪儿、怎么用。本书试图针对大量经典实例进行循序渐进式的优化，为读者提供详尽的优化参考。

最后，GPU编程实践时不仅仅只是编程，还涉及编译器、运行环境等相关配套知识，若是没有这些配套知识，即使看懂了、写出了相关代码又有什么意义呢？这无异于纸上谈兵。本书将涉及代码编写、编译，运行时需要涉及的所有配套知识，包括系统环境、Linux命令、编译选项、性能分析、并行计算相关常识等。

综上所述，本书的定位是帮助GPU程序员从“零知识”入门到精通的书籍，书中内容包含通俗易懂的GPU理论知识，配套的知识体系，大量代码实例及其循序渐进的优化过程、详细的性能分析和知识点总结，与性能直接相关的GPU核心特征获取、分析和论述。对于GPU程序的开发人员，本书具有较为全面的参考价值。

本书的结构和阅读建议

本书共计四篇：第一篇共5章，主要是GPU的理论知识，包括GPU的领域背景（高性能计算概述）、GPU概述、GPU硬件架构、GPU软件体系和CUDA C编程；第二篇共4章，基于4个入门级的GPU实例展示了其详细的并行和优化过程，分别是向量加法、向量内积、矩阵乘法和矩阵转置；第三篇共5章，分别描述了5类不同应用的GPU编程和优化过程，包括卷积、曼德博罗特集、前缀求和、排序和简单图像处理；第四篇共4章，阐述了影响GPU程序性能的核心因素，分别从GPU执行核心、GPU存储体系、影响GPU性能的关键因素、CPUs和GPUs的协同运算4个角度展开探讨。

首遍阅读时，建议按行文顺序阅读，本书已按知识难易程度做了梳理；接着动手实践入门篇和提高篇；若要进一步深入阅读则需结合核心篇章的GPU核心知识与入门篇和提高篇的实践，逐步理解提升；最终在本书基础上进一步优化各类应用，进而开发出自己的优化方法。

对于急于从实例运行入手的读者，可根据4.3节内容安装环境，然后跳读入门篇和提高篇的实例章节，在获得一定成功经验后再返回阅读理论篇和核心篇。

本书的特点和优势

本书内容丰富，涵盖了系统全面的GPU知识体系、循序渐进的实例优化、从实践导出的真实有效的优化方法、影响GPU性能的核心因素、GPU性能测评和CPU/GPU异构协同优化等内容。本书语言朴实，通俗易懂，对不易理解的概念定义，通过笔者的理解重新进行阐述。本书还提供简单易读的实例代码，详尽的编译命令和清晰的运行结果数据。

GPU发展迅猛，GPU架构几乎每两年就更新换代（就在本书第3轮修订期间，Pascal架构发布，笔者又重新修订了相应章节），目前市面上的书籍暂时没有提供完整的GPU架构知识，本书总结了所有的GPU架构，阐述了更加系统完善的GPU知识体系。

GPU优化是关键，市面上同类书籍中阐述了很多优化方法，但经历了循序渐进优化的具体实例相对缺乏，甚至有些优化方法存在问题（本书有相应的实验佐证）。

致谢

本书大纲由方民权和方建滨共同商定；方民权完成本书代码的编写、实验数据测试与

分析、本书初稿、首轮修订、重要修订(如错误修正、结构调整和章节内容增加等)以及后续修订内容的权衡和更新;张卫民、方建滨、周海芳等完成后续数轮修订,正是有他们的修订才使本书真正可读;高畅提供了 Linux 图形界面 CUDA 安装的文档和验证,验证了本书所有代码的结果正确性和性能准确性,以及进行一些关键的查漏补缺。正是所有作者共同努力,本书才能真正成稿。

本书支撑课题包括国家自然科学基金项目(41375113 和 61272146)、湖南省研究生创新资助项目(CX2015B030)、国防科学技术大学计算机学院联合博导组项目(面向异构平台的海洋预报软件可移植性技术研究)。

文中涉及许多笔者的个人学术观点和经验总结,由于笔者水平有限,所有成果仅供参考,如有不准确甚至错误之处,望读者谅解并批评指正。另因编写时间仓促,尽管已进行 6 轮修订,但书中纰漏和瑕疵在所难免,若发现笔误或有其他意见建议,请致信 admin@hpc6.com,万分感谢!

编　者

2016 年 5 月

笔者的话

2016年1月20日是一个值得纪念的日子，这本书终于完成初稿，尽管有的章节还有待大幅修改。自2015年3月有了初步构思，我就义无反顾地开始撰写本书，到最终完稿，期间也多次调整论述结构。作为一名在校博士研究生，顶着博士毕业的压力，写一本对博士毕业“无用”的书，确实挺不容易的，不过我认为是值得的。

本科，我学的是机械专业，硕士转学计算机，2012年接触GPU编程，当时MIC已然兴起，却还能感受到GPU的“余热”。说实话，刚接触GPU时整个人是懵的，师兄已经毕业，师门就我一人研究GPU，当时只有两本GPU书——《GPU高性能运算之CUDA》和*CUDA by example*供参考，书中的知识还有些“过时”（Tesla架构和Fermi架构的区别）、“不系统”、甚至“错误”（见13.3.2节）（当然不可否认这两本书对我的帮助），网上资料就更少了。也许是我个人水平有限，仅Windows的GPU开发环境安装就耗费了半个多月时间，Linux的GPU环境更是失败了不知道多少次（当时网上资料很少，且很多方法尝试后均不可行）；首个GPU程序开发完全不知道如何下手，头大了近20天才在某天灵光一闪想通，对GPU优化更是无从着手。其实上述过程只需要一位“老师”带着成功一遍，就能少走很多弯路。说这么多，其实想表达的意思是，新手时的我非常期盼一本系统的、全面的、循序渐进的GPU书籍，这也是我写本书的动因。

2015年，在学术圈GPU早已“过时”，为什么还致力于撰写本书呢？首先，GPU确实稳定好用：2014年也曾研究过MIC，“不可捉摸”的ICC编译器耗光了我的耐心，往往逻辑正确的代码编译器就是报错，明明简单向量化的代码编译器就是认为不可向量化；而CUDA代码只要编写正确，均能正常编译和执行，且性能取决于CUDA代码本身。其次，我认为GPU并行是一个大趋势大市场：GPU有3个层次可扩展性（见3.1节），能满足各种市场需求；GPU（显卡）已在PC普遍装备，性价比高，GPU平台已相当普及；GPU的性能不易受其他程序影响，相比之下CPU程序性能受其他程序影响程度较大；GPU发展迅猛，在未来量子计算机或生物计算机出现前（出现后也不可能立刻取代现有的计算机），CPU/GPU异构系统将会长期存在。

本书的定位是辅助“零知识”GPU程序开发人员从入门到精通的书籍，试图做到“知行合一”（最完美的想法是所有的理论知识均有实验结果佐证，但显然还有差距）、系统性强、知识全面、优化循序渐进。希望本书能够帮助每一位GPU程序开发人员。

方民权

目 录

第一篇 理 论 篇

第二篇 入 门 篇

第三篇 提 高 篇

第四篇 核 心 篇

附　录

第一篇　理论篇

理论篇主要阐述系统全面的 GPU 知识体系，包括 GPU 并行计算的领域背景（高性能计算概述）、GPU 概述、GPU 硬件架构、GPU 软件体系和 GPU 编程方法（CUDA C 编程）。通过阅读本篇，读者能够较为系统全面地掌握 GPU 相关的基础理论和技术知识，建立必要的知识结构。

本篇内容丰富，笔者推荐几项值得重点关注的内容：①向量机和阵列机的辨别，深入浅出地对比了向量机（CPU）和阵列机（GPU）的区别（1.5 节）；②GPU 购买注意事项（2.4 节）；③全面的 GPU 架构知识，包括 Tesla、Fermi、Kepler、Maxwell 和 Pascal 架构（3.1 节）；④实测可行的 CUDA 环境安装指南，包括 Windows、Linux 命令行和 Linux 图形界面等[①]（4.3 节）；⑤CUDA 编程七步曲，简述了 CUDA 编程的基本步骤（5.2 节）。

① 同类 CUDA 书籍也有环境安装内容，但笔者亲测大都不可用，也许是笔者水平有限吧。