高性能计算系列丛书

HZ BOOKS
华章科技

ELSEVIER

CUDA Application Design and Development

高性能CUDA应用设计与开发

方法与最佳实践

（美）Rob Farber 著

于玉龙 唐堃 译 郭禾 王宇新 审校

机械工业出版社
China Machine Press

高性能计算系列丛书

CUDA Application Design and Development

高性能CUDA应用设计与开发

方法与最佳实践

[redacted]

机械工业出版社
China Machine Press

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。

CUDA APPLICATION DESIGN AND DEVELOPMENT
Rob Farber
ISBN: 978-0-12-388426-8

本书版权登记号：图字：01-2012-2655

图书在版编目（CIP）数据

高性能 CUDA 应用设计与开发：方法与最佳实践 /（美）罗布著；于玉龙，唐堃译. —北京：机械工业出版社，2013.1
（高性能计算系列丛书）
书名原文：CUDA Application Design and Development

ISBN 978-7-111-40446-0

Ⅰ. 高…　Ⅱ. ①罗…　②于…　③唐…　Ⅲ. 计算机图形学　Ⅳ. TP391.41

中国版本图书馆 CIP 数据核字（2012）第 275905 号

机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码　100037）
责任编辑：秦　健
北京市荣盛彩色印刷有限公司印刷
2013 年 1 月第 1 版第 1 次印刷
186mm × 240mm · 18.25 印张
标准书号：ISBN 978-7-111-40446-0
定价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991；88361066
购书热线：（010）68326294；88379649；68995259
投稿热线：（010）88379604
读者信箱：hzjsj@hzbook.com

译 者 序

目前，并行计算是提高算法性能的主要方式，无论是大规模集群系统还是个人电脑，甚至是智能手机，都将性能的提升寄托于并行计算。GPU 芯片曾主要用于计算机可视化和图形图像处理，其内部具有大量的并行计算单元，有着天然的并行优势。近年来，NVIDIA 和 AMD 等厂商相继推出以 GPU 为协处理器的通用并行计算解决方案，CUDA 是其中的典型代表。

CUDA 是 NVIDIA 推出的一种利用 GPU 进行通用并行计算的整套解决方案，包括硬件支持、程序语言扩展、编译器、调试器等整套开发工具链，它对矩阵数值计算、图像视频处理、机器学习、计算机可视化等问题都有良好的加速效果。CUDA 提供的 CUDA C 编程语言扩展，利用 Kernel 函数等机制，将大规模并行计算逻辑通过一种简单、有效且合理的方式进行表达，成为充分利用并行硬件实现程序性能加速的关键。同时，NVIDIA 还提供了充分的配套开发工具链，如 nvcc 编译器、cuda-gdb 调试器、gprof 和 computeprof 性能分析工具、Parallel Nsight 性能跟踪工具等。此外，CUDA 还包含了一些常用的支持库，如用于基础开发的 Driver、Runtime、Thrust API；用于线性代数数值计算的 CUBLAS；用于图像处理的 NPP 等。CUDA 甚至可以用于开发不使用 GPU 加速的应用程序。

本书是一本介绍 CUDA 程序设计的优秀图书。对于那些具有一定程序开发实践基础，并渴望学习并行计算、GPU 程序设计的读者来说，本书是快速入门 CUDA 并学会编写高效 GPU 并行计算程序的理想选择。本书注重理论与实践的结合，与其他介绍 CUDA 的图书相比，具有如下特点：

首先，对 CUDA 的介绍深入浅出、全面细致。对于基础知识的介绍可以使读者快速了解 CUDA，迅速上手编程；深入的分析又能够帮助读者了解不同的程序设计方式带来性能差异的原因，从而设计出更高效的程序，对于实践这点十分重要。

其次，在介绍 CUDA 的编程语言语法、各种资源以及工具的使用方法的基础上，立足于实践，选择了如机器学习、可视化系统等诸多有代表性的示例，将书中介绍的各种技术学以致用，不仅可以帮助读者更深入地理解 CUDA 编程思想，也有利于将 CUDA技术更好地应用于实际工作中。

此外，本书内容广泛，涵盖了 CUDA 技术的各个方面，其中不仅介绍了基本的单

GPU 系统中 CUDA 程序的设计，还涉及了单机多 GPU 系统、多机 GPU 集群系统，甚至云计算环境中 CUDA 技术的应用等。本书还介绍了 CUDA 第三方支持产品和 CUDA 的衍生产品，例如可将 CUDA 程序编译为运行在常规 CPU 系统上的 PGI 等。

通过翻译本书，我们深深体会到作者对于 CUDA 和并行计算相关领域知识的深刻理解以及为编写本书所付出的努力，从中收获颇丰。我们相信读者也能够享受本书所带来的愉悦。本书的翻译力求在保证准确和接近原文风格的前提下，专业技术词汇采用国内专业通用术语，尽可能做到易于理解。

本书的第 1、3、5、6、9、10 章由于玉龙翻译，第 2、4、7、8、11、12 章由唐堃翻译，并由郭禾、王宇新和于玉龙共同完成全文终稿定稿。

感谢机械工业出版社的各位编辑在本书翻译过程中提供的帮助；感谢沈阳药科大学王薇同学在图书译稿的校对工作中提供的大量帮助。还要感谢我们的科研团队——大连理工大学软件学院高性能计算与图像处理实验室、大连理工大学计算机科学与技术学院系统结构实验室的各位同仁为本书翻译工作提供的各种支持。另外，还要特别感谢已经毕业的程童同学，为我们实验室开展 GPU 并行计算所做出的开创性工作，以及对本书翻译工作的支持。最后，感谢家人在图书翻译工作中给予我们的支持和鼓励。

由于时间仓促，翻译过程中的瑕疵和错误在所难免，欢迎各位读者批评指正。您的批评与建议会成为我们进步的基石。

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

前　言

在技术领域以及学术、职业生涯中，把握时机是极其重要的。我们这一代程序员非常幸运，能够率先利用这种经济实惠而且普及的大规模并行计算设备。GPGPU（通用图形处理单元）技术正在世界范围内影响着计算领域的各个方面——小到手机，大到超级计算机。它还在改变着商业应用、科学计算、云计算、计算机可视化系统、游戏和机器人等领域，甚至重新定义了我们所熟知的计算机编程方式。Teraflop（每秒万亿次浮点计算）如今能以一种经济可行的方式惠及全世界大多数人。无论是青少年、学生、家长、教师、专业人士，还是小型科研机构、大型公司都能负担得起 GPGPU 设备，又能获得免费的软件开发工具（SDK）。NVIDIA 公司估计已生产并销售了超过 3 亿个可编程 GPGPU 设备。

这 3 亿个用于 CUDA（Compute Unified Device Architecture，统一计算设备架构）开发的 GPGPU 设备，在商业应用领域中展现出极大的市场潜力，并且拓展了科学计算的可行范围。最重要的是，CUDA 以及大规模并行 GPGPU 设备正在改变我们对计算的认知。CUDA 程序不再是某个时刻只进行一个或少量的操作，而是同时进行数以万计的操作！

本书将介绍 CUDA 编程思想以及如何利用数以万计的线程使得应用（无论是商业还是学术、科研方面）的执行性能获得指数级的提升。进而，本书将介绍单一应用如何在单机或集群环境中使用单个或多个 GPGPU。此外，本书还将介绍如何利用 CUDA 来开发针对多核处理器平台的应用，使 CUDA 可以用于所有的应用开发场景，甚至是没有 GPU 的环境中！

本书不仅仅关注程序语法和 API 函数，还将涉及 CUDA 的设计理念，并在体系结构方面探讨 GPGPU 硬件带来惊人性能的原因。本书也将涉及 CUDA 开发中的准则和注意事项，以利于编写简洁、易读并具有可维护性的代码。本书主要以最新的 CUDA 4. x 为准。

使用与修改代码是学习过程中的必要环节，因此，本书提供了可供编译和修改的示例代码。这些示例展示了如何在 Fermi 架构的 GPGPU 设备（NVIDIA 20 系列）中编写具有高性能的代码，旨在让读者学会编写高效的代码而不是仅仅让代码能够运行。书中的示例可以在任何支持 CUDA 的 GPGPU 上编译运行，因此使用旧版 GPGPU 的读

者也可以从本书中获益。本书在适当的位置引用了我所撰写的全面 CUDA 教程之“Doctor Dobb's Journal”（Dobb 博士的日志）专栏中的文字，来说明 CUDA 各版本之间的改进以及如何在各代 GPGPU 架构中获得高计算性能。

相关教学资料、附加示例以及读者评论在 http://gpucomputing.net 的 wiki 中可以找到，可以从以下任何一个网址中访问：

- 作者的名字：http://gpucomputing.net/RobFarber
- 本书名（连成一个单词）：http://gpucomputing.net/CUDAapplicationdesignanddevelopment
- 作者的专栏名：http://gpucomputing.net/supercomputingforthemasses

购买本书的读者还可以下载书中示例的源代码：http://booksite.mkp.com/9780123884268。

本书内容安排如下：

第 1 章介绍了 CUDA 的基本概念以及构建、调试 CUDA 应用所需要的工具。针对 Thrust C++ 和 C Runtime API，该章提供了一些简单的示例。此外，还介绍了三个有效提高程序性能的 GPU 编程准则。

第 2 章利用第 1 章所介绍的技术，提供了完整的机器学习以及相对于常规单核处理器加速了 341 倍的优化通用框架。同时，为了满足对领域知识和 GPU 程序开发能力有双重需求的读者，该章还介绍了机器学习与数值优化的核心概念。

第 3 章主要关注性能分析技术（profiling），它是高性能编程的一个重要技术。该章介绍了 CUDA 性能分析工具，并应用于第 2 章的案例中。而且将揭示一些 Thrust API 的惊人瓶颈。该章还讨论了数据挖掘技术以及主成分分析（PCA）和非线性主成分分析（NPCA）两种数据挖掘方法。无论是用户还是编程者，该章都是值得阅读的。

第 4 章主要介绍 CUDA 执行模型。该章的内容是任何希望获得 GPU 最高性能的人所必须了解的。为了帮助读者理解 GPU 的工作方式，该章还提供了示例以及性能分析的输出结果；这些资料有助于读者掌握如何使用现有工具来观察程序运行的内部细节。

第 5 章介绍并讨论了 CUDA 提供的多种 GPU 内存以及各种内存的优缺点。

第 6 章。鉴于最快与最慢的 GPU 内存之间有超过 3 个数量级的性能差距，因此合理利用 GPU 内存是获得高性能的唯一可行途径。该章将讨论使应用高效利用内存的技术，并通过观察性能分析输出结果理解这些技术。该章还给出了一个通用功能的示例，指导编写类似于关键 API 的基础通用函数。

第 7 章。GPU 提供了多种并行方式，包括多 GPU、异步核函数执行以及统一虚拟地址空间（Unified Virtual Address，UVA）。该章通过示例和对应的性能分析输出结果来介绍这些 GPU 并行方式的应用。

第 8 章。历经发展，CUDA 已适用于包括 GPU 和多核处理器在内的所有应用开发平台。该章将讨论多种 CUDA 后端设备，并给出可有效在异构多 GPU 环境中运行的示例以及响应的性能分析输出结果。该章还设计了 CUDA 支持库，以及如何在其他高级语言（如 Python、Java、R 和 FORTRAN 等）中应用 CUDA 与 GPU 计算接口。

第 9 章。在关注利用 CUDA 加速计算任务的同时，往往容易忽略 GPU 技术也是优秀的可视化平台。该章回归 GPU 的设计初衷，讨论 CUDA 如何显著地加速可视化与游戏应用。该章将给出一个完整的示例，使读者可以创建一个 3D 世界并在其中遨游。性能分析输出结果揭示了 CUDA 能够有效加速可视化应用的原因。该章的框架还可扩展并用到第 12 章的实时视频流的处理例子中。

第 10 章介绍使用 MPI（Message Passing Interface，通信传递接口）扩展第 3 章的示例来介绍可伸缩性以及性能。示例代码的另一版本证明了，在扩展至 500 个 GPGPU（单体最高具有每秒五十万亿次单精度浮点计算的能力）的过程中其性能近似于线性增长，最终获得了每秒超过三百万亿次浮点计算能力，相当于 60 000 个 x86 处理器核。

第 11 章。在 CUDA 技术浪潮中，还没有能覆盖所有方面的书籍。该章是一个综述章节，介绍了应用于不同技术的 CUDA 开源代码的项目，其中包括支持向量机（Support Vector Machine，SVM）、多维尺度分析（MDS）、交互信息、力导引图、分子模型等。对于这些项目的了解以及第 8 章讨论的其他高级语言接口，将有助于你成为一名成熟的 CUDA 开发者。

第 12 章。基于第 9 章的框架，该章给出了一个用于视频识别的实时视频流处理示例。仅需要一个廉价的摄像头或一个视频文件，你就可以完成实时的视频识别任务。本示例为学习 CUDA 而设计，因此很容易修改。经过扩展，本示例与技术可用于机器人、增强现实游戏，以及综合状态视窗（HUD）的数据融合，该章对此也进行了探讨。

学习 CUDA（以及 GPGPU）编程思想与开发技术是非常有趣的。然而，性能是使用 GPGPU 技术的终极目的，正如我的一位大学教授经常说的："要知道布丁的滋味就要去品尝一下。"图 1 是 2011 年 7 月 12 日 NVIDIA CUDA Showcase 上报道的获得性能加速比前 100 名的应用。这些报告都来自于经权威评议的学术论文和商业实体，可以看出，对于许多不同种类的应用，相比于多核处理器，GPGPU 技术都能够获得 2 个数量级（100 倍）或以上的加速。展示的这些应用值得我们花时间进行探究，而且其中大部分都可以免费下载源代码或支持库。

GPGPU 是一种颠覆性的技术，它改变了人们对计算过程的认识。正如 NVIDIA 所说，"从超智能手机到超级计算机"。该技术的到来正赶上这样一场机遇风暴，传统的多核处理器已经无法再通过增加时钟频率来实现有效的加速。传统处理器生产商唯一

能吸引消费者升级计算机的方式，就是通过双核或四核的并行来获得 2 ~ 4 倍的加速。多核并行是具有颠覆性的，要求现有软件必须重写以实现对多核心的有效利用。在计算机与科研领域转向并行硬件探究之时，加入到这前沿软件应用研发中来吧！学习 CUDA，把握这一难得的机遇。

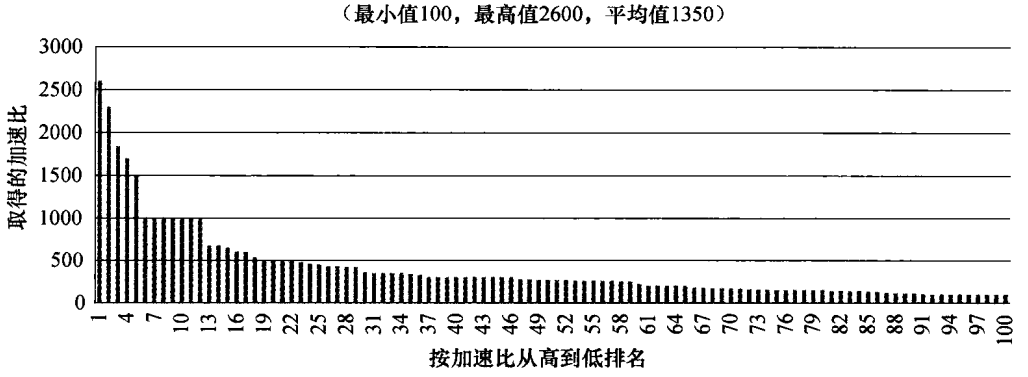

图 1　NVIDIA 展示的性能加速比前 100 名应用

致谢

仅以此书献给我的妻子 Margy 和我的儿子 Ryan，虽然他们没有直接帮助我写作，却在我创作本书的过程中一直深深地支持着我。尤其是我的儿子 Ryan 的成长，证明了长江后浪推前浪，感谢上帝让我陪他度过童年时光。

献给许多审阅过此书的朋友们，尤其是那些指出本书错误的朋友，对于他们付出的时间与帮助，我感激不尽。特别感谢爱尔兰高端计算中心的每个人，他们从我开始创作到完成本书手稿期间一直给予我支持。最后，感谢 NVIDIA 公司的朋友与同事，是他们将所有 CUDA 革命性的梦想变成了现实。

目　录