HZ BOOKS
华章科技

PEARSON

英伟达公司CUDA首席架构师Nicholas Wilt亲笔撰写，英伟达中国首批CUDA官方认证工程师翻译

全面而系统地讲解CUDA编程的各方面知识，深度解析CUDA各种优化技术，包含大量实用代码示例，是深入掌握主流异构并行计算技术的权威指南

高性能计算系列丛书

The CUDA Handbook
A Comprehensive Guide to GPU Programming

CUDA专家手册
GPU编程权威指南

（美）Nicholas Wilt 著

苏统华 马培军 刘曙 吕家明◎译　英伟达中国◎技术审校

机械工业出版社
China Machine Press

高性能计算系列丛书

The CUDA Handbook
A Comprehensive Guide to GPU Programming

CUDA专家手册
GPU编程权威指南

（美）Nicholas Wilt 著
苏统华 马培军 刘曙 吕家明◎译 英伟达中国◎技术审校

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

CUDA 专家手册：GPU 编程权威指南 /（美）威尔特（Wilt. N.）著；苏统华等译．—北京：机械工业出版社，2014.8
（高性能计算系列丛书）
书名原文：The CUDA Handbook: A Comprehensive Guide to GPU Programming

ISBN 978-7-111-47265-0

I. C… II. ①威… ②苏… III. 图象处理－程序设计－手册 IV. TP391.41-62

中国版本图书馆 CIP 数据核字（2014）第 161732 号

本书版权登记号：图字：01-2013-5981

CUDA 专家手册：GPU 编程权威指南

[美] Nicholas Wilt 著

出版发行：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：关 敏　　责任校对：董纪丽
印　刷：三河市宏图印务有限公司　　版　次：2014 年 8 月第 1 版第 1 次印刷
开　本：186mm×240mm 1/16　　印　张：23.25
书　号：ISBN 978-7-111-47265-0　　定　价：85.00 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　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

Forword 中文版序

得知本书正在被翻译成瑰丽的汉字，我难抑内心的激动。众核计算对中国读者来说并不陌生。中国已经建造了多个最快的超级计算机，其中很多采用了 CUDA 技术。我们耳熟能详的天河 -1A，建成于 2010 年 10 月，配备了 7168 片特斯拉（Tesla）M2050 型号的 GPU。本书第 4 章介绍的亚马逊 cg1.4xlarge 实例，也采用了同种型号的 GPU。天河 -1A 曾雄霸世界超级计算机 500 强榜首逾半年之久，即使现在，它仍排在世界最快计算机的前 12 位。它惊人的高速度源于中国自主研发的互联技术把来自英特尔和英伟达的硬件高效地集成起来了。

当然，CUDA 并非只能应用于超级计算机。GPU 的出货量巨大，相较其他众核计算技术，它们具有价格优势。因此，支持 CUDA 的 GPU 很容易用来搭建经济的计算平台，这对于并行程序设计的教育和教学意义重大。

CUDA 技术仍在迅猛发展着。我敢断定，它将继续在并行计算的大潮中扮演重要角色。本书向大家展示了 CUDA 的内部工作机制，此中译本将推动 CUDA 在中国的普及和发展。

最后，向承担本书翻译工作的苏博士及其团队，致以诚挚谢意！感谢他们孜孜不倦地致力于 GPU 计算的推广，特别是把本书带给广大中国读者。

Nicholas Wilt

推荐序 *Foreword*

自 CPU 单核性能在 2004 年左右停止提升后，功耗的限制及大量并行应用本身的特点决定了多核（CPU）加众核（GPU）的异构架构成为从最快的超级计算机到手机等移动设备上计算芯片的主流架构。所以如果想在现在和未来的任何计算设备上写出性能具有竞争力的程序，所有的开发者都需要学习异构架构下的并行计算技术。CUDA 是目前最成熟、使用者最多的异构并行计算技术。本书对开发者更深入地掌握主流异构并行计算技术会有很大帮助。

程序优化的目的是提高整体性能，所以优化时需要做的第一件事就是找出程序的哪个部分需要提高整体性能，即瓶颈分析。这是至关重要而又经常被刚接触优化的开发者忽略的一件事。如果不先进行定量的瓶颈分析，而是立刻优化自己认为是瓶颈的地方，经常会事倍功半。瓶颈分析的第一步是热点分析，即找出哪些部分占程序的大部分运行时间。热点分析可以通过在源代码里加时间测量代码或使用性能分析工具（profiler）进行。当确认整体瓶颈是 GPU 的函数或数据传输后，读者就可参阅本书对应的 GPU 优化知识进行优化。

本书第一部分共 4 章，介绍 GPU 的架构、系统软件及编程环境。性能优化本质上就是通过理解架构及相应系统软件的特点，并在程序中利用这些特点更高效地完成计算。所以优化的第一步就应该是理解目标平台的架构及系统软件基础。这部分讨论的很多内容是第一次在公开的 CUDA 书籍中出现。

第二部分共 6 章，介绍 CUDA 编程各个方面的具体知识，特别是如何利用 GPU 特点的很多优化技术。

如果性能分析告诉我们 CPU-GPU 数据传输是瓶颈，那么可以参考第 6 章使用流来计算隐藏数据传输；另一方面，如果 GPU 函数的执行是瓶颈，那么下一步就是分析内核函数的架构瓶颈。此时，定量分析也是至关重要的。如果不先做定量分析找出瓶颈在哪里，而是立刻随机地尝试各种优化技术，往往难以收到成效。这样随机尝试，可能有些优化恰好有效

果，有些优化则没有效果，会浪费大量时间在尝试各种技术上。另外，我们也无法明白为什么一种优化有效果而其他的却没有效果。除此之外，如果不做分析，我们也无从得知一个内核函数到底优化到何种程度就可以结束，从而有可能花费大量时间在一个已没有太大优化空间的内核函数上。这些都是优化技术的初学者很容易觉得优化是一件复杂而低效的工作的原因。但其实如果能采用先做定量分析再采用相应的优化技术的方法，优化完全可以变成一种高效、有据可依的技术。

对内核函数做分析，本质就是找出内核函数的性能瓶颈在架构的什么地方。一般情况下，性能瓶颈有三种可能类型。

1. 内存密集型。这种情况是内核函数的大部分时间都花在数据的读写指令上。从硬件的角度看，在这个内核函数运行期间，大部分时间都花在数据在内存总线的传输上并且内存总线的带宽已充分利用。对这种情况，如何优化可以参考第 5 章及第 10 章的内容。

2. 指令密集型。这种情况是内核函数的大部分时间都花在各种指令的执行上。从硬件的角度看，在这个内核函数运行期间，大部分时间都花在处理其指令单元的执行上。对这种情况，如何优化可以参考第 8 章的内容。

3. 延迟密集型。这种情况是内核函数的大部分时间都花在等待高延迟的指令（如内存读写）上。从硬件的角度看，在这个内核函数运行期间，大部分时间都花在数据在内存总线的传输上但内存总线的带宽没有被充分利用。对这种情况，如何优化可以参考第 7 章占有率的内容。

本书的第三部分共 5 章，给出了多个领域的具体例子，介绍如何应用各种优化技术。本部分的内容有助于读者学以致用。

虽然本书具体讨论 CUDA，但并行优化的技术及想法对几乎任何众核架构都是适用的。所以读者如果在阅读本书了解具体 CUDA 优化技术之外，也能有意识地体会蕴含其中的并行优化的基本方法，会有更大收获。

王鹏

英伟达高性能计算开发技术经理

译者序 *Words of the Translator's*

CUDA 入门相对比较容易。也许只需寥寥几个小时的学习，你就可以开始编写支持 CUDA 的程序，初窥 GPU 加速的魅力。但是要修炼成为榨取 GPU 处理能力的高级 CUDA 工程师，却并非易事。你需要了解并行计算的基本理论，需要研究 CPU/GPU 的硬件架构，需要熟悉 CUDA 的软件抽象架构和工具，需要掌握必要的领域知识，等等。目前，可以助你入门的 CUDA 书籍和教程为数不少，但在利用 CUDA 来深度榨取 GPU 潜力方面，仍几乎空白。

本书为深度优化 CUDA 程序性能提供了全方位的指导。乍看起来，本书只是一本参考手册。不错，它确实基于英伟达官方 CUDA 参考手册的很多内容。然而它又不单单是一本传统意义上的手册，它比较全面地介绍了 CUDA 的高级编程特性，是该领域难得的权威参考书。全书高屋建瓴而又简明扼要。作者从繁芜的文献资源中提炼并予以升华形成独到的总结。对很多复杂的主题，能够举重若轻；在很多关键的问题上，给大家一种豁然开朗之感。很多精辟论述是市面上任何一本 GPU 书籍所缺失的。译者在阅读本书时，亦为作者在技术和语言上的双重驾驭能力所折服。

本书作者 Wilt 从事 GPU 底层开发逾 20 年，功力深厚。在 CUDA 诞生之前，从事了 8 年 Direct3D 的开发；在英伟达供职的 8 年间，参与了 CUDA 的开发，实现了多数 CUDA 的抽象机制；目前转战亚马逊，从事 GPU 云产品的开发。作者根据自身多年的经验积淀，对 CUDA 编程模型进行了深入浅出的介绍，并结合四个典型应用场合，完美诠释了它们的具体用法和优化过程。得益于本书的启发，我们研究组基于 CUDA 的文字识别算法得到 194X 的加速，使我们能够抢在其他竞争者之前，取得 2013 年手写汉字识别竞赛的两项冠军。很多时候，速度意味着胜利！

本书的技术深度和欲覆盖的读者层面，决定了本书翻译工作的难度。本书的翻译持续

了一年之久，期间得到过很多人的帮助。除了本书列出的四位译者外，还有 5 位译者参与了本书的初稿翻译：韦振良参与了第 2 章和第 7 章，李硕参与了第 4 章和第 6 章，李松泽参与了第 10 章，孙黎参与了第 14 章，胡光龙参与了第 15 章。其中，与光龙是在参加第一届英伟达认证工程师的考试途中相遇，我们一起迷了路、一起考试过关、一起分享 CUDA 技术点滴，又一起在本书的翻译期间进行合作，谢谢光龙的友谊和无私帮助。在本书译文统稿之后，我们努力想找到合适的 CUDA 专家为我们把关。如果没有英伟达公司的侯宇涛经理，这可能会变成无法完成的任务。感谢侯经理为我们推荐了最优秀的技术审校专家，帮我们邀请了高性能计算开发技术中国区经理王鹏博士为本译作写序并选定本译作在全国推广。感谢英伟达高性能计算部门的专家王泽寰和赖俊杰，即使在春节期间，他们也在为改进本书的译稿质量而绞尽脑汁。如果没有泽寰和赖经理的审校，原书的内容恐难以尽可能原汁原味地展现。王鹏经理特意为本书作了推荐序，其中浓缩了他多年的 CUDA 性能优化经验；非常感谢，您真的做到了“读者最需要什么，您就写点什么”的宗旨。最后，还要特别感谢机械工业出版社的编辑团队，他们做了大量卓有成效的工作，是本书的幕后英雄。

本书的翻译，还得到了多项项目的资助，在此一并致谢。国家自然科学基金（资助号：61203260）、黑龙江省科研启动基金（资助号：LBH-Q13066）、哈尔滨工业大学科研创新基金（资助号：HITNSRIF2015083）对本书的翻译提供了部分资助。另外，哈尔滨工业大学创新实验课《CUDA 高性能并行程序设计》、黑龙江省教育厅高等教育教学改革项目（资助号：JG2013010224）、哈尔滨工业大学研究生教育教学改革研究项目（资助号：JCJS-201309）也对本书的翻译提供了大力支持。

很多时候，翻译技术书籍是件吃力不讨好的事情。即使是最有资格的译者，也一定无法免受被读者指责的境遇。对读者负责，对 CUDA 教学推广负责，是我和我的团队始终铭记的准则。我们很享受翻译本书的过程，也很愿意承担因我们的水平和疏忽所招致的批评。文中可能存在的任何翻译问题，都是我们的责任。真诚地希望读者朋友不吝赐教，欢迎大家的任何意见：cudabook@gmail.com。

并行计算是计算的未来。希望本书的翻译可以帮助你让并行计算与大数据成功联姻，并享用由此带来的前所未有的新鲜体验。愿你的 CUDA 探索之旅愉快！

苏统华

哈尔滨工业大学软件学院

前　言 *Preface*

如果你正在读本书，意味着我无须再向你兜售 CUDA。你应该已经对 CUDA 有所了解，可能使用过英伟达的 SDK 和文档，抑或参加过并行程序设计的课程，再或者阅读过类似《CUDA by Example》（Jason Sanders 和 Edward Kandrot 著，2011 年由 Addison-Wesley 出版）这样的入门书籍。

我在审校《CUDA by Example》时，惊诧于该书内容的浅显。它假设读者是零基础，试图描述从内存类型及其实际应用到图形互操作性，甚至到原子操作等方方面面的内容。它是很优秀的 CUDA 入门书籍，但也只能做到泛泛而谈。对于更深层次的知识，例如，平台的工作机理、GPU 的硬件结构、编译器驱动程序 nvcc 以及以前缀求和（“扫描”）为代表的基本并行算法，则鲜有涉及。

本书考虑到不同基础的读者，旨在帮助 CUDA 新手进阶中级水平，同时帮助中级程序员继续提升他们的水平到一个新高度。对于入门性质的书籍，最好从头到尾阅读，而对本书则可以根据需要选读。如果你正准备建立支持 CUDA 的新平台并在上面进行编程，建议你精读第 2 章。如果你正纳闷你的应用程序是否将受益于 CUDA 流带来的额外并发性，你应该查看第 6 章。其他的章节分别提供了软件架构、GPU 的纹理操作和流处理器簇等 GPU 子系统的详细描述，还有依据不同数据存取模式和在并行算法领域的重要程度而精心挑选的应用案例。尽管不同章之间难免存在交互引用，但每一章的内容都是相对独立的。

本书将披露包括 CUDA 5.0 在内的最新技术。最近几年，CUDA 和它的目标平台得到了长足发展。在《CUDA by Example》出版之际，GeForce GTX 280（GT200）才刚刚面世。迄今，CUDA 硬件已历经两代演变。因此，本书除了在如映射锁页内存（mapped pinned memory）的现有特性上不吝笔墨外，还特别关注 CUDA 支持的新特性，像费米架构的 ballot、开普勒架构的 shuffle、64 位和统一虚拟寻址特性以及动态并行（dynamic parallelism）

等。此外，本书还将讨论最近的平台进展，例如英特尔沙桥（Sandy Bridge）CPU 上集成的 PCIe 总线控制器。

尊敬的读者，你可以全篇通读本书，也可以把它放于电脑边随时查阅。但不管怎样，我真诚地希冀你能从中得到乐趣，如同我执笔分享时一样快意。

致谢

借此机会，感谢英伟达公司的朋友们。他们耐心地为我释疑、检视我的作品并反馈建设性意见。特别的谢意送给 Mark Harris、Norbert Juffa 和 Lars Nyland。

本书的审校者在成稿之前审阅了本书，他们付出了大量时间，提供的宝贵意见提高了本书的质量和清晰性，保证了技术的正确性。在此，特别感谢 Andre Brodtkorb、Scott Le Grand、Allan MacKinnon、Romelia Salomon-Ferrer 和 Patrik Tennberg 的反馈意见。

本书的写作过程历经了重重困难，如果没有编辑 Peter Gordon 的超凡耐心和支持，很难最终呈现给大家。Peter 的助手 Kim Boedigheimer 为本项目的顺利完成做了大量工作，帮助我设定了项目的各种专业标准。对她在征求、协调评审意见以及在本书成稿之后把本书上传到 Safari 网站的整个过程中付出的努力，表示特别感谢。

在本书的写作过程中，我的妻子 Robin 和我的儿子 Benjamin、Samuel 和 Gregory 一直对我支持有加，谢谢他们。

目　录 *Contents*

第二部分 CUDA 编程