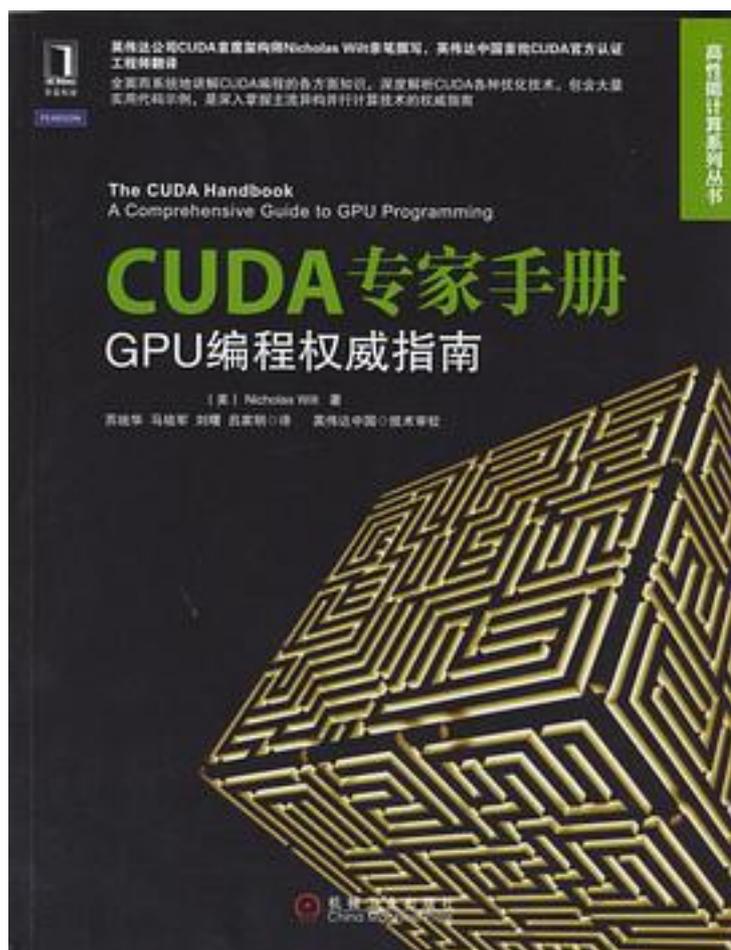


# CUDA专家手册



[CUDA专家手册\\_下载链接1](#)

著者:[美] Nicholas Wilt

出版者:机械工业出版社

出版时间:2014-8-26

装帧:平装

isbn:9787111472650

《CUDA专家手册：GPU编程权威指南》由英伟达公司CUDA首席架构师Nicholas Wilt亲笔撰写，深度解析GPU的架构、系统软件、编程环境，以及CUDA编程各方面的知识和各种优化技术，包含大量实用代码示例，是并行程序开发领域最有影响力的著作

之一。

《CUDA专家手册：GPU编程权威指南》分为三部分，共15章。第一部分(第1~4章)介绍CUDA开发的基础知识、硬件/软件架构和软件环境；第二部分(第5~10章)详细解析CUDA开发的各个方面，包括内存、流与事件、内核执行、流处理器簇、多gpu编程和纹理操作；第三部分(第11~15章)利用多个实例，深入分析流式负载、归约算法、扫描算法、N-体问题和图像处理的归一化相关系数计算，介绍如何应用各种优化技术。

作者介绍:

Nicholas

Wilt拥有逾25年底层编程经验，他的技术兴趣跨越多个领域，包括工业机器视觉、图形处理和底层多媒体软件开发等。作为英伟达公司CUDA首席架构师，他见证了CUDA从无到有的整个过程，设计并实现了多数CUDA的底层抽象机制。在加入英伟达公司之前，他曾在微软公司担任Direct3D 5.0和6.0产品的开发组组长，完成了Windows桌面管理器的原型开发，并在此期间开展了早期GPU计算的工作。目前，Wilt先生任职于亚马逊公司，从事与GPU产品相关的云计算技术。

苏统华，博士，英伟达中国首批CUDA官方认证工程师，英伟达官方认证CUDA培训师，哈尔滨工业大学英伟达教学中心负责人，主要研究领域包括大规模并行计算、模式识别、物联网智能信息处理、智能媒体交互与计算等。2013年，其所开发的CUDA识别算法，在文档分析和识别国际会议(ICDAR' 2013)上获得手写汉字识别竞赛的双料冠军。另外，他在手写汉字识别领域建立了里程碑式工作，论文他引约300次；他所建立的HIT-MW库，为全世界100多家科研院所采用；目前负责国家自然科学基金项目2项。著有英文专著《Chinese Handwriting Recognition: An Algorithmic Perspective》(德国施普林格出版社)，CUDA\*II关译作2本(机械工业出版社)。现任哈尔滨工业大学软件学院高级讲师、硕士生导师。

目录:《CUDA专家手册：GPU编程权威指南》

中文版序

推荐序

译者序

前言

第一部分基础知识

第1章简介2

1.1方法4

1.2代码4

1.2.1验证型代码5

1.2.2演示型代码5

1.2.3探究型代码5

1.3资源5

1.3.1开源代码5

1.3.2cuda专家手册库(chlib) 6

1.3.3编码风格6

1.3.4cuda sdk6

1.4结构6

第2章硬件架构8

2.1cpu配置8

2.1.1前端总线9

2.1.2对称处理器簇9

2.1.3非一致内存访问(numa) 10

- 2.1.4集成的pcie12
- 2.2集成gpu13
- 2.3多gpu14
- 2.4cuda中的地址空间17
  - 2.4.1虚拟寻址简史17
  - 2.4.2不相交的地址空间20
  - 2.4.3映射锁页内存21
  - 2.4.4可分享锁页内存21
  - 2.4.5统一寻址23
  - 2.4.6点对点映射24
- 2.5cpu/gpu交互24
  - 2.5.1锁页主机内存和命令缓冲区25
  - 2.5.2cpu/gpu并发26
  - 2.5.3主机接口和内部gpu同步29
  - 2.5.4gpu间同步31
- 2.6gpu架构31
  - 2.6.1综述31
  - 2.6.2流处理器簇34
- 2.7延伸阅读37
- 第3章软件架构39
  - 3.1软件层39
    - 3.1.1cuda运行时和驱动程序40
    - 3.1.2驱动程序模型41
    - 3.1.3nvcc、ptx和微码43
  - 3.2设备与初始化45
    - 3.2.1设备数量46
    - 3.2.2设备属性46
    - 3.2.3无cuda支持情况48
  - 3.3上下文50
    - 3.3.1生命周期与作用域51
    - 3.3.2资源预分配51
    - 3.3.3地址空间52
    - 3.3.4当前上下文栈52
    - 3.3.5上下文状态53
  - 3.4模块与函数53
  - 3.5内核（函数）55
  - 3.6设备内存56
  - 3.7流与事件57
    - 3.7.1软件流水线57
    - 3.7.2流回调57
    - 3.7.3null流57
    - 3.7.4事件58
  - 3.8主机内存59
    - 3.8.1锁页主机内存60
    - 3.8.2可分享的锁页内存60
    - 3.8.3映射锁页内存60
    - 3.8.4主机内存注册60
  - 3.9cuda数组与纹理操作61
    - 3.9.1纹理引用61
    - 3.9.2表面引用63
  - 3.10图形互操作性63
  - 3.11cuda运行时与cuda驱动程序api65
- 第4章软件环境69
  - 4.1nvcc——cuda编译器驱动程序69

- 4.2ptxas——ptx汇编工具73
- 4.3cuobjdump76
- 4.4nvidia-smi77
- 4.5亚马逊web服务79
  - 4.5.1命令行工具79
  - 4.5.2ec2和虚拟化79
  - 4.5.3密钥对80
  - 4.5.4可用区域 (az) 和地理区域81
  - 4.5.5s381
  - 4.5.6ebs81
  - 4.5.7ami82
  - 4.5.8ec2上的linux82
  - 4.5.9ec2上的windows83
- 第二部分cuda编程
- 第5章内存88
  - 5.1主机内存89
    - 5.1.1分配锁页内存89
    - 5.1.2可共享锁页内存90
    - 5.1.3映射锁页内存90
    - 5.1.4写结合锁页内存91
    - 5.1.5注册锁页内存91
    - 5.1.6锁页内存与统一虚拟寻址92
    - 5.1.7映射锁页内存用法92
    - 5.1.8numa、线程亲和性与锁页内存93
  - 5.2全局内存95
    - 5.2.1指针96
    - 5.2.2动态内存分配97
    - 5.2.3查询全局内存数量100
    - 5.2.4静态内存分配101
    - 5.2.5内存初始化api102
    - 5.2.6指针查询103
    - 5.2.7点对点内存访问104
    - 5.2.8读写全局内存105
    - 5.2.9合并限制105
    - 5.2.10验证实验：内存峰值带宽107
    - 5.2.11原子操作111
    - 5.2.12全局内存的纹理操作113
    - 5.2.13ecc (纠错码) 113
  - 5.3常量内存114
    - 5.3.1主机与设备常量内存114
    - 5.3.2访问常量内存114
  - 5.4本地内存115
  - 5.5纹理内存118
  - 5.6共享内存118
    - 5.6.1不定大小共享内存声明119
    - 5.6.2束同步编码119
    - 5.6.3共享内存的指针119
  - 5.7内存复制119
    - 5.7.1同步内存复制与异步内存复制120
    - 5.7.2统一虚拟寻址121
    - 5.7.3cuda运行时121
    - 5.7.4驱动程序api123
- 第6章流与事件125
  - 6.1cpu/gpu的并发：隐藏驱动程序开销126

- 6.2异步的内存复制129
  - 6.2.1异步的内存复制：主机端到设备端130
  - 6.2.2异步内存复制：设备端到主机端130
  - 6.2.3null流和并发中断131
- 6.3cuda事件：cpu/gpu同步133
  - 6.3.1阻塞事件135
  - 6.3.2查询135
- 6.4cuda事件：计时135
- 6.5并发复制和内核处理136
  - 6.5.1concurrencyMemcpyKernel.cu137
  - 6.5.2性能结果141
  - 6.5.3中断引擎间的并行性142
- 6.6映射锁页内存143
- 6.7并发内核处理145
- 6.8gpu/gpu同步：cudaStreamWaitEvent()146
- 6.9源代码参考147
- 第7章内核执行148
  - 7.1概况148
  - 7.2语法149
    - 7.2.1局限性150
    - 7.2.2高速缓存和一致性151
    - 7.2.3异步与错误处理151
    - 7.2.4超时152
    - 7.2.5本地内存152
    - 7.2.6共享内存153
  - 7.3线程块、线程、线程束、束内线程153
    - 7.3.1线程块网格153
    - 7.3.2执行保证156
    - 7.3.3线程块与线程id156
  - 7.4占用率159
  - 7.5动态并行160
    - 7.5.1作用域和同步161
    - 7.5.2内存模型162
    - 7.5.3流与事件163
    - 7.5.4错误处理163
    - 7.5.5编译和链接164
    - 7.5.6资源管理164
    - 7.5.7小结165
- 第8章流处理器簇167
  - 8.1内存168
    - 8.1.1寄存器168
    - 8.1.2本地内存169
    - 8.1.3全局内存170
    - 8.1.4常量内存171
    - 8.1.5共享内存171
    - 8.1.6栅栏和一致性173
  - 8.2整型支持174
    - 8.2.1乘法174
    - 8.2.2其他操作（位操作）175
    - 8.2.3漏斗移位（sm 3.5）175
  - 8.3浮点支持176
    - 8.3.1格式176
    - 8.3.2单精度（32位）180
    - 8.3.3双精度（64位）181

- 8.3.4半精度 (16位) 181
- 8.3.5案例分析: float到half的转换182
- 8.3.6数学函数库185
- 8.3.7延伸阅读190
- 8.4条件代码191
  - 8.4.1断定191
  - 8.4.2分支与汇聚191
  - 8.4.3特殊情况: 最小值、最大值和绝对值192
- 8.5纹理与表面操作193
- 8.6其他指令193
  - 8.6.1线程束级原语193
  - 8.6.2线程块级原语194
  - 8.6.3性能计数器195
  - 8.6.4视频指令195
  - 8.6.5特殊寄存器196
- 8.7指令集196
- 第9章多gpu203
  - 9.1概述203
  - 9.2点对点机制204
    - 9.2.1点对点内存复制204
    - 9.2.2点对点寻址205
  - 9.3uva: 从地址推断设备206
  - 9.4多gpu间同步207
  - 9.5单线程多gpu方案208
    - 9.5.1当前上下文栈208
    - 9.5.2n-体问题210
  - 9.6多线程多gpu方案212
- 第10章纹理操作216
  - 10.1简介216
  - 10.2纹理内存217
    - 10.2.1设备内存217
    - 10.2.2cuda数组与块的线性寻址218
    - 10.2.3设备内存与cuda数组对比222
  - 10.3一维纹理操作223
  - 10.4纹理作为数据读取方式226
    - 10.4.1增加有效地址范围226
    - 10.4.2主机内存纹理操作228
  - 10.5使用非归一化坐标的纹理操作230
  - 10.6使用归一化坐标的纹理操作237
  - 10.7一维表面内存的读写238
  - 10.8二维纹理操作240
  - 10.9二维纹理操作: 避免复制242
    - 10.9.1设备内存上的二维纹理操作242
    - 10.9.2二维表面内存的读写243
  - 10.10三维纹理操作244
  - 10.11分层纹理245
    - 10.11.1一维分层纹理246
    - 10.11.2二维分层纹理246
  - 10.12最优线程块大小选择以及性能246
  - 10.13纹理操作快速参考248
    - 10.13.1硬件能力248
    - 10.13.2cuda运行时249
    - 10.13.3驱动api250
- 第三部分实例

第11章流式负载	254
11.1设备内存	255
11.2异步内存复制	258
11.3流	259
11.4映射锁页内存	260
11.5性能评价与本章小结	261
第12章归约算法	263
12.1概述	263
12.2两遍归约	265
12.3单遍归约	269
12.4使用原子操作的归约	271
12.5任意线程块大小的归约	272
12.6适应任意数据类型的归约	273
12.7基于断定的归约	276
12.8基于洗牌指令的线程束归约	277
第13章扫描算法	278
13.1定义与变形	278
13.2概述	279
13.3扫描和电路设计	281
13.4cuda实现	284
13.4.1先扫描再扇出	284
13.4.2先归约再扫描（递归）	288
13.4.3先归约再扫描（两阶段）	291
13.5线程束扫描	294
13.5.1零填充	295
13.5.2带模板的版本	296
13.5.3线程束洗牌	297
13.5.4指令数对比	298
13.6流压缩	300
13.7参考文献（并行扫描算法）	302
13.8延伸阅读（并行前缀求和电路）	303
第14章n-体问题	304
14.1概述	305
14.2简单实现	309
14.3基于共享内存实现	312
14.4基于常量内存实现	313
14.5基于线程束洗牌实现	315
14.6多gpu及其扩展性	316
14.7cpu的优化	317
14.8小结	321
14.9参考文献与延伸阅读	323
第15章图像处理的归一化相关系数计算	324
15.1概述	324
15.2简单的纹理实现	326
15.3常量内存中的模板	329
15.4共享内存中的图像	331
15.5进一步优化	334
15.5.1基于流处理器簇的实现代码	334
15.5.2循环展开	335
15.6源代码	336
15.7性能评价	337
15.8延伸阅读	339
附录acuda专家手册库	340
术语表	347

• • • • • [\(收起\)](#)

[CUDA专家手册 下载链接1](#)

## 标签

CUDA

GPU

parallel

计算机

苏统华

programming

有电子版

## 评论

CUDA进阶。接下来两周要做Multi-GPU的DNN数据并行训练，这本书读的正是时候。  
2018 再读了一遍

-----  
只给两星，因为翻译太差，有些句子非得像对待英语的复杂句式那样，要用语法拆解句子结构，才能看明白，然后瞬间就能想象出英文原文是怎么写的。 Yet, this book is awesome for those people who want to get advanced skills in CUDA programming.  
-----

[CUDA专家手册 下载链接1](#)

书评

-----  
[CUDA专家手册 下载链接1](#)