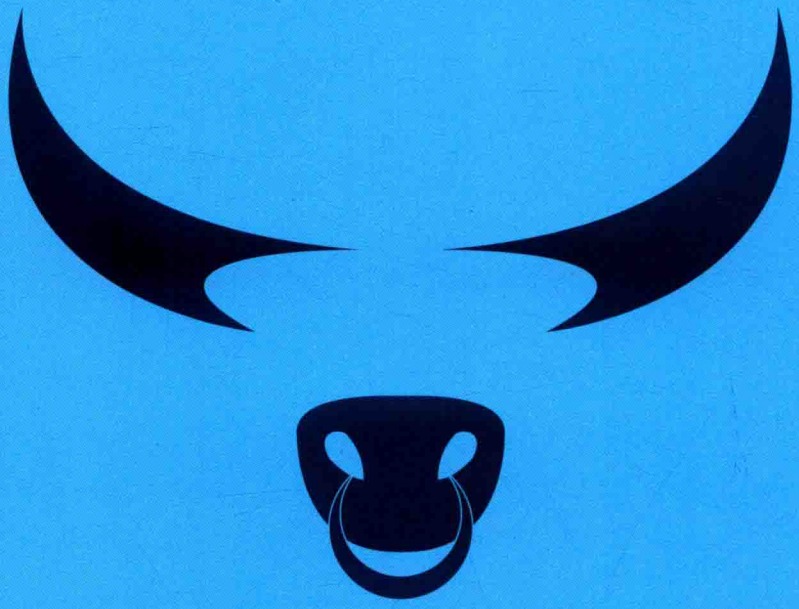
异步图书
www.epubit.com

基于本书内容的在线视频课程——“Linux 操作系统分析”
被教育部评为 “国家精品在线开放课程”
被教育部高等学校计算机类专业教学指导委员会评为
“2018 年度中国高校计算机教育慕课联盟优秀在线开放课程”



庖丁解牛 Linux 内核分析

孟 宁 娄嘉鹏 刘宇栋◎编著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

关注公众号，获取本书配套资源

庖丁解牛 Linux内核分析

孟 宁 娄嘉鹏 刘宇栋◎编著

人民邮电出版社
北京

图书在版编目（CIP）数据

庖丁解牛Linux内核分析 / 孟宁，娄嘉鹏，刘宇栋编著. -- 北京 : 人民邮电出版社，2018.10
ISBN 978-7-115-49186-2

Ⅰ. ①庖… Ⅱ. ①孟… ②娄… ③刘… Ⅲ. ①Linux操作系统 Ⅳ. ①TP316.85

中国版本图书馆CIP数据核字(2018)第209270号

内 容 提 要

本书从理解计算机硬件的核心工作机制（存储程序计算机和函数调用堆栈）和用户态程序如何通过系统调用陷入内核（中断异常）入手，通过上下两个方向双向夹击的策略，并利用实际可运行程序的汇编代码从实践的角度理解操作系统内核，分析 Linux 内核源代码，从系统调用陷入内核、进程调度与进程切换开始，最后返回到用户态进程。

本书配有丰富的实验指导材料和练习，适合作为高等院校计算机相关专业的指导用书，也适合 Linux 操作系统开发人员自学。

◆ 编　　著　孟　宁　娄嘉鹏　刘宇栋
责任编辑　张　爽
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市君旺印务有限公司印刷

◆ 开本：800×1000　1/16
印张：12.25
字数：256 千字　　2018 年 10 月第 1 版
印数：1 – 3 000 册　　2018 年 10 月河北第 1 次印刷

定价：49.00 元

读者服务热线：(010) 81055410　印装质量热线：(010) 81055316
反盗版热线：(010) 81055315
广告经营许可证：京东工商广登字 20170147 号

编写顾问委员会

陈莉君　西安邮电大学

李　曦　中国科学技术大学

黄敬群　台湾成功大学

代　栋　北卡罗来纳大学夏洛特分校

孙志岗　网易教育事业部

石　磊　实验楼在线教育

序

大大小小、可见与不可见的计算机已成为现代人日常工作、学习和生活中必不可少的工具。操作系统是计算机之魂，作为用户使用计算机的接口，它负责调度执行各个用户程序，使计算机完成特定的任务；作为计算机硬件资源的管理者，它负责协调计算机中各类设备高效地工作。操作系统的重要性不言而喻。

对于软件工程师，理解操作系统的工作原理和关键机制是设计高质量应用程序的前提，但要做到这一点是十分困难的。一方面，操作系统设计涉及计算机科学与工程学科的方方面面，包括数据结构与算法、计算机组成与系统结构、计算机网络，甚至程序设计语言与编译系统等核心知识，以及并发、同步和通信等核心概念。另一方面，作为一个复杂庞大的软件产品，理解操作系统更需要理论与实践深度结合。

操作系统的相关学习资料十分丰富。有阐述基本原理者，有剖析典型系统者，还有构造示例系统者；有面向专业理论者，亦有面向应用实践者。角度多种多样，内容简繁不一。

本书的最大特点在于作者结合其多年的 Linux 操作系统实际教学经验编撰而成。作为一位经验丰富的高级软件工程师和专业教师，本书作者基于自己学习和研究 Linux 的心得，创新性地以一个 mykernel 和 MenuOS 为基础实验平台进行教学和实验组织，实现了理论学习与工程实践的自然融合，达到了事半功倍的效果。同时，书中设计了丰富的单元测试题和实验，引导读者循序渐进地掌握所学知识，并有效地促进读者深入思考和实践所学内容。作者基于本书开设的操作系统课程，其教学形式涉及面对面的课堂教学和在线慕课教学，选课对象既包括软件工程硕士，又包括一般工程实践者，学习人数已数以万计。本书的出版体现了作者认真吸收大量的学员反馈，不断优化课程的教学内容和过程组织的成果。

易读性是本书的另一特色。作者采用二维码这一新媒体时代的代表性技术组织全书的内容，达到了兼顾完整性和简洁性的目标。

作为一名多年从事计算机系统结构研究和教学的教育工作者，我认为本书的出版对于提升国内操作系统教学和实践水平非常有益，相信它必将受到读者的喜爱！

李曦

此为试读，需要完整PDF请访问：www.ertongbook.com

前　言

作者于2000年左右开始接触计算机，一直对计算机系统的工作机制抱有浓厚的兴趣，阅读了很多相关书籍，包括关于分析 Linux 源代码的书籍，但一直不得要领，没能准确把握计算机系统工作的核心机制。2009年，我与中国科学技术大学软件学院结缘，从软件工程师转行成为教师。在学校里，我非常幸运地与陈香兰老师一起教授“Linux 操作系统分析”课程，可是面对2000万行的 Linux 内核代码和厚厚的《深入理解 Linux 内核》这本教材，我发现自己依然无法从全局和本质上把握 Linux 系统。

直到 2013 年暑假，我替另一位老师代课，教授“操作系统原理”课程（见二维码 1），凭借近 10 年使用 Linux 系统和学习 Linux 内核的经验，我为课程实验定下了一个“小目标”：学习“操作系统原理”就要动手编写一个小型操作系统。教学中的作业、实验和考试就像各种比赛一样，看似是在考学生，实际是在考验教师的水平和能力。当时学生的编程经验和动手能力普遍不足，很难独自完成编写一个哪怕非常微小的操作系统的任务，这时就需要教师给予启发和指导，帮助学生一步步完成预定的目标。正是在这次教学过程中，我在 Linux 内核繁杂的 CPU 初始化工作的基础上完成了一个简单、虚拟、可编程的计算机硬件模拟环境 mykernel（见二维码 2），在这个仅支持时钟中断的虚拟 CPU 中就可以建立属于自己的内核了。有了 mykernel，稍有编程能力的学生就可以编写一个简单的时间片轮转调度的小型内核，并且能读懂代码，深刻理解如何在 CPU 的一个指令执行流上实现多个进程。

二维码1

二维码2

正是有了实现 mykernel 的经验，我在之后的“Linux 操作系统分析”课程教学中有了清晰的思路。其中一位同学关于 mykernel 的总结也体现了我的感受：

mykernel 这样一个短小精悍的模拟内核，时常会给我提供看问题的角度和思路。当被庞杂的 Linux 内核代码弄得一头雾水时，我就去看看 mykernel，很多复杂的问题就可以用简单的机制解释了。

mykernel 为 Linux 内核初学者提供了一个很好的平台，目前有很多的 Linux 内核学习者在使用。台湾成功大学的黄敬群创建的 kernel-in-kernel 项目（见二维码 3）是一个 mykernel 的衍生项目，黄敬群

二维码3

还专门发邮件以取得我的授权。

在我看来，mykernel 是深入理解 Linux 的一个不错的工具，也是“Linux 操作系统分析”慕课课程及本书的一个重要实验。除了 mykernel 这一个实验外，本书还有哪些内容？一位慕课课程学员的总结非常到位，远远超过了我自己来介绍这门慕课课程及本书的文字水平，这里也分享给读者：

这门课没讲什么？

在学习操作系统时，我们知道了操作系统将 CPU 抽象为进程，将内存抽象为虚拟内存，学习了进程的调度算法、内存页面的置换算法，这门课并没有关注这些算法。操作系统的主要功能就是为用户屏蔽硬件的操作细节，帮助用户管理计算机系统的各种资源。同步机制是我们处理并发任务和进行资源管理的重要手段。关于原子操作、信号量和自旋锁等内容，该课程中没有讲解。在操作系统原理课程中，没有着重讲解的各种设备驱动程序实际上占了 Linux 内核代码的大部分比例，这门课并没有这部分内容。没有讲解文件系统的结构与实现，以及 VFS 等。

这门课讲了什么？

对于要研究 Linux 内核的人来说，x86 汇编语言是你必须要面对的第一关。因为操作系统需要大量对寄存器的操作，这是与体系结构相关的操作，所以必须用汇编语言来解决。这门课在一开始就讲解了 x86 汇编语言，并在后面的课程中不断巩固，这一点对于阅读内核源码非常有用。该课程用一个简单的演示内核 mykernel 来说明 Linux 是如何启动的，包括一个进程是怎样描述的（PCB 信息）、0 号进程（idle）的创建与演化、1 号进程 init 的创建与加载、2 号进程 kthreadd 的创建等。这可以使我们从顶层对 Linux 内核有一个大概的认识。并且课中手把手地进行源码教学，可以让人减少对结构复杂的内核代码的恐惧。我们日常使用内核，其实大部分功能都是使用它的系统调用，如从创建一个新的进程 fork、装载程序 execve，到输入/输出、时间查询等。因此，我们研究内核，很大一部分都是在研究如何实现这些系统调用。这门课花了两周时间来讲解系统调用在内核中是如何进行的。如果把进程创建和可执行程序的装载也当成系统调用的讲解，那么实际上占了课程的一半。因此，课程的设置正体现了这些系统调用在内核构成中的重要性。课中提供了一个试验环境 MenuOS，该系统实现了一个命令行菜单系统，我们只需要添加我们希望执行的功能函数到菜单就好了。同时，利用 Qemu 和 gdb，我们跟踪了各种系统调用的执行过程。虽然这门课没有讲具体的调度算法，如 Linux 内核中著名的完全公平队列 CFS，但对于进程调度来说，除了调度算法，还有两个重要的问题，那就是进程的调度时机与切换过程，该课程花了一节课的时间来讲解 schedule()函数的实现。我们不仅需要学习 Linux 内核的相关知识，

更需要学习正确的人生观和世界观，这门课的精髓在于不仅教会你如何分析 Linux 内核，而且教你做事的方法论："天下难事必做于易，天下大事必做于细"。对于代码规模庞大无从下手的内核，我们从小处入手，步步为营，最终掌控全局。

由于慕课课程的受众比较多元化，课程的容量和看视频做实验的时间都需要严格控制，因此 8 周的慕课课程及本书内容主要聚焦在 Linux 系统工作的核心机制上，算是基础核心篇，相对来讲比较短小精悍。我所讲授的中国科学技术大学软件学院的研究生课程"Linux 操作系统分析"涉及的内容要比上述内容更多。不少学员提出问题：学完慕课课程之后想继续深入学习，需要学习哪些内容？我个人认为，深入理解 Linux 系统除了理解 Linux 系统工作的核心机制之外，文件的概念和实现也非常重要。类 UNIX 系统非常成功的抽象就是"一切都是文件"，深入理解文件的概念和内核实现对于理解 Linux 内核尤为重要。如果有机会继续做后续课程，我来选择的话首先要做的就是文件抽象篇。

无论是基础核心篇，还是上述提及的文件抽象篇，都要注重理解，而非应用。从应用的角度来学习和研究 Linux 内核，其实还可以分为 API 接口篇、网络协议篇和驱动程序篇，分别对应的阅读人群大致为底层应用软件或系统软件的开发人员、网络相关的工作人员和硬件驱动程序开发人员。可能有读者会疑惑为什么没有内存管理，内存管理的底层实现基本上固化到了 CPU 芯片内部，它对于理解 Linux 系统工作的核心机制和系统架构都相对单纯独立，已经通过进程的地址空间在逻辑上做了清晰的隔离。而从应用的角度来看，垃圾回收（Garbage Collection，GC）成为语言的标配已经是大势所趋。除非专业从事存储器产品研发、芯片内部存储管理模块或内核内存管理模块开发等细分领域，我个人认为操作系统原理中涉及的内存管理相关的知识已经足够了。

致谢

感谢中国科学技术大学软件学院曾一起合作教授"Linux 操作系统分析"课程的陈香兰和李春杰两位老师，他们为"Linux 操作系统分析"慕课课程及本书做出了前期基础性的贡献。

感谢网易云课堂的孙志岗，没有他的鼓励和支持，我多年获得的教学成果恐怕至今也不会以慕课课程的方式在互联网上与学习者见面，本书更是无从谈起。

感谢实验楼的石磊在开发和配置"Linux 操作系统分析"慕课课程的实验环境过程中提供了很多支持和帮助。

感谢电子工业出版社章海涛老师提出了很好的意见和建议，以及为本书前期的筹备工作所做的贡献。

感谢本书的两位合作者，分别是北京电子科技学院的娄嘉鹏和刘宇栋，没有你们的鼓励和鼎力支持，本书出版恐怕遥遥无期。

感谢“Linux 操作系统分析”慕课课程建设之前的中国科学技术大学软件学院的几届学生，感谢你们在学习过程中撰写了很多高质量的博客，为慕课课程和本书做出了贡献。

感谢人民邮电出版社陈冀康、张涛、张爽 3 位编辑为本书顺利出版所做的工作和努力。

由于写作时间仓促及作者的能力有限，本书难免会有不足之处，敬请各位读者批评指正，我的电子邮件地址为 mengning@ustc.edu.cn。

孟宁

2018 年春

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，点击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“**异步社区**”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“**异步图书**”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

此为试读，需要完整PDF请访问：www.ertongbook.com

吾生也有涯，而知也无涯。以有涯随无涯，殆已；已而为知者，殆而已矣。为善无近名，为恶无近刑。缘督以为经，可以保身，可以全生，可以养亲，可以尽年。

庖丁为文惠君解牛，手之所触，肩之所倚，足之所履，膝之所踦，砉然响然，奏刀騞然，莫不中音。合于《桑林》之舞，乃中《经首》之会。

文惠君曰："嘻！善哉！技盖至此乎？"庖丁释刀对曰："臣之所好者道也，进乎技矣。始臣之解牛之时，所见无非牛者。三年之后，未尝见全牛也。方今之时，臣以神遇，而不以目视，官知止而神欲行。依乎天理，批大郤，道大窾，因其固然。技经肯綮之未尝，而况大軱乎！良庖岁更刀，割也；族庖月更刀，折也。今臣之刀十九年矣，所解数千牛矣，而刀刃若新发于硎。彼节者有间，而刀刃者无厚，以无厚入有间，恢恢乎其于游刃必有余地矣。是以十九年而刀刃若新发于硎。虽然，每至于族，吾见其难为，怵然为戒，视为止，行为迟。动刀甚微，謋然已解，如土委地。提刀而立，为之四顾，为之踌躇满志，善刀而藏之。"文惠君曰："善哉！吾闻庖丁之言，得养生焉。"

摘自《庄子·养生主》

目　录

第 1 章　计算机工作原理 …… 1

1.1　存储程序计算机工作模型 …… 1
1.2　x86-32 汇编基础 …… 3
1.2.1　x86-32 CPU 的寄存器 …… 4
1.2.2　数据格式 …… 6
1.2.3　寻址方式和常用汇编指令 …… 7
1.2.4　汇编代码范例解析 …… 11
1.3　汇编一个简单的 C 语言程序并分析其汇编指令执行过程 …… 13
1.4　单元测试题 …… 26
1.5　实验 …… 27

第 2 章　操作系统是如何工作的 …… 29

2.1　函数调用堆栈 …… 29
2.2　借助 Linux 内核部分源代码模拟存储程序计算机工作模型及时钟中断 …… 32
2.2.1　内嵌汇编 …… 32
2.2.2　虚拟一个 x86 的 CPU 硬件平台 …… 34
2.3　在 mykernel 基础上构造一个简单的操作系统内核 …… 36
2.3.1　代码范例 …… 36
2.3.2　代码分析 …… 42
2.4　单元测试题 …… 48
2.5　实验 …… 48

第 3 章　MenuOS 的构造 …… 50

3.1　Linux 内核源代码简介 …… 50
3.2　构造一个简单的 Linux 内核 …… 56
3.3　跟踪调试 Linux 内核的启动过程 …… 60
3.4　单元测试题 …… 65
3.5　实验 …… 66

第 4 章　系统调用的三层机制（上）……67
4.1　用户态、内核态和中断……67
4.2　系统调用概述……70
4.2.1　操作系统提供的 API 和系统调用的关系……70
4.2.2　触发系统调用及参数传递方式……71
4.3　使用库函数 API 和 C 代码中嵌入汇编代码触发同一个系统调用……72
4.3.1　使用库函数 API 触发一个系统调用……72
4.3.2　内嵌汇编语法简介……73
4.3.3　C 代码中嵌入汇编代码触发一个系统调用……75
4.3.4　含两个参数的系统调用范例……76
4.3.5　通用的触发系统调用的库函数 syscall……78
4.4　单元测试题……79
4.5　实验……80
第 5 章　系统调用的三层机制（下）……81
5.1　给 MenuOS 增加命令……81
5.2　使用 gdb 跟踪系统调用内核函数 sys_time……83
5.3　系统调用在内核代码中的处理过程……85
5.3.1　中断向量 0x80 和 system_call 中断服务程序入口的关系……86
5.3.2　在 system_call 汇编代码中的系统调用内核处理函数……87
5.3.3　整体上理解系统调用的内核处理过程……88
5.4　单元测试题……91
5.5　实验……92
第 6 章　进程的描述和进程的创建……93
6.1　进程的描述……93
6.2　进程的创建……97
6.2.1　0 号进程的初始化……98
6.2.2　内存管理相关代码……99
6.2.3　进程之间的父子、兄弟关系……100
6.2.4　保存进程上下文中 CPU 相关的一些状态信息的数据结构……101
6.2.5　进程的创建过程分析……103
6.3　单元测试题……120
第 7 章　可执行程序工作原理……122
7.1　ELF 目标文件格式……122

7.1.1 ELF 概述 …… 122
7.1.2 ELF 格式简介 …… 123
7.1.3 相关操作指令 …… 128
7.2 程序编译 …… 129
7.2.1 预处理 …… 129
7.2.2 编译 …… 130
7.2.3 汇编 …… 131
7.2.4 链接 …… 133
7.3 链接与库 …… 134
7.3.1 符号与符号解析 …… 134
7.3.2 重定位 …… 137
7.3.3 静态链接与动态链接 …… 139
7.4 程序装载 …… 143
7.4.1 程序装载概要 …… 143
7.4.2 fork 与 execve 内核处理过程 …… 148
7.4.3 庄周梦蝶 …… 153
7.4.4 小结 …… 154
7.5 单元测试题 …… 155
7.6 实验 …… 156
第 8 章 进程的切换和系统的一般执行过程 …… 158
8.1 进程调度的时机 …… 158
8.1.1 硬中断与软中断 …… 158
8.1.2 进程调度时机 …… 159
8.2 调度策略与算法 …… 161
8.2.1 进程的分类 …… 161
8.2.2 调度策略 …… 162
8.2.3 CFS 调度算法 …… 164
8.3 进程上下文切换 …… 165
8.3.1 进程执行环境的切换 …… 165
8.3.2 核心代码分析 …… 167
8.4 Linux 系统的运行过程 …… 172
8.5 Linux 系统构架与执行过程概览 …… 174
8.5.1 Linux 操作系统的构架 …… 174

8.5.2　ls 命令执行过程即涉及操作系统相关概念……175
8.6　进程调度相关源代码跟踪和分析……176
8.6.1　配置运行 MenuOS 系统……176
8.6.2　配置 gdb 远程调试和设置断点……177
8.6.3　使用 gdb 跟踪分析 schedule()函数……177
8.7　单元测试题……179

第1章

计算机工作原理

本章重点介绍计算机的工作原理，具体涉及存储程序计算机工作模型、基本的汇编语言，以及 C 语言程序汇编出来的汇编代码如何在存储程序计算机工作模型上一步步地执行。其中重点分析了函数调用堆栈相关汇编指令，如 call/ret 和 pushl/popl。

1.1 存储程序计算机工作模型

存储程序计算机的概念虽然简单，但在计算机发展史上具有革命性的意义，至今为止仍是计算机发展史上非常有意义的发明。一台硬件有限的计算机或智能手机能安装各种各样的软件，执行各种各样的程序，这在人们看起来都理所当然，其实背后是存储程序计算机的功劳。

存储程序计算机的主要思想是将程序存放在计算机存储器中，然后按存储器中的存储程序的首地址执行程序的第一条指令，以后就按照该程序中编写好的指令执行，直至程序执行结束。

相信很多人特别是学习计算机专业的人都听说过图灵机和冯·诺依曼机。图灵机关注计算的哲学定义，是一种虚拟的抽象机器，是对现代计算机的首次描述。只要提供合适的程序，图灵机就可以做任何运算。基于图灵机建造的计算机都是在存储器中存储数据，程序的逻辑都是嵌入在硬件中的。

与图灵机不同，冯·诺依曼机是一个实际的体系结构，我们称作冯·诺依曼体系结构，它至今仍是几乎所有计算机平台的基础。我们都知道“庖丁解牛”这个成语，比喻经过反复实践，掌握了事物的客观规律，做事得心应手，运用自如。冯·诺依曼体系结构就是各种计算机体系结构需要遵从的一个“客观规律”，了解它对于理解计算机和操作系统非常重

此为试读，需要完整PDF请访问：www.ertongbook.com