HZ BOOKS
华章教育

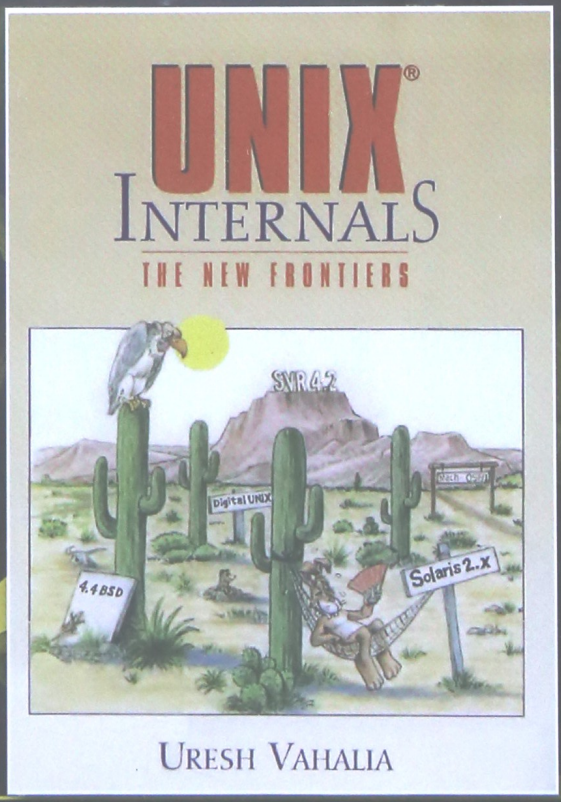
计 算 机 科 学 丛 书

PEARSON

深入理解UNIX系统内核

[美] Uresh Vahalia 著　李雨 薛磊 黄庆新 译

UNIX Internals
The New Frontiers

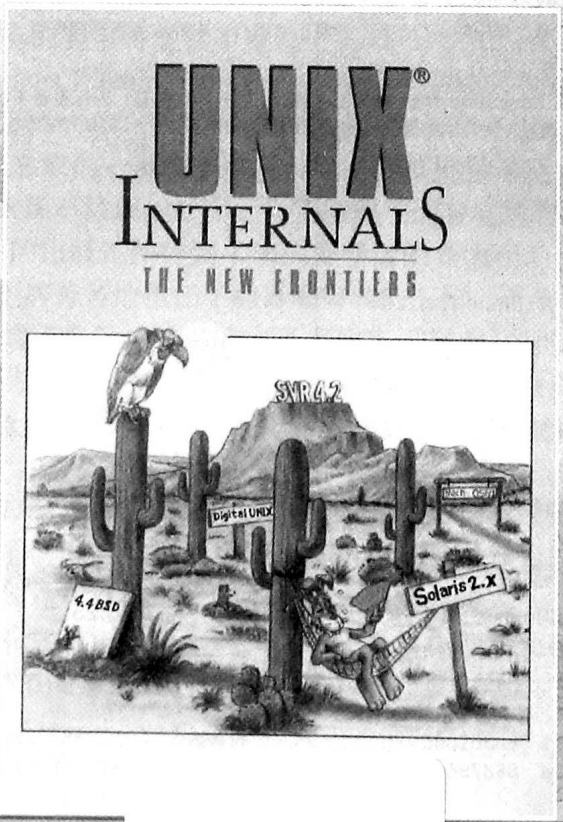


机械工业出版社
China Machine Press

计 算 机 科 学 丛

深入理解UNIX系统内核

[美] Uresh Vahalia 著 李雨 薛磊 黄庆新 译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

深入理解 UNIX 系统内核 /（美）瓦哈利亚（Uresh Vahalia）著；李雨，薛磊，黄庆新译．
—北京：机械工业出版社，2015.2
（计算机科学丛书）
书名原文：UNIX Internals：The New Frontiers

ISBN 978-7-111-49145-3

I. 深… II. ① 瓦… ② 李… ③ 薛… ④ 黄… III. UNIX 操作系统 IV. TP316.81

中国版本图书馆 CIP 数据核字（2015）第 012268 号

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：吴　怡　　责任校对：董纪丽
印　　刷：北京瑞德印刷有限公司　　版　　次：2015 年 5 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：30.25
书　　号：ISBN 978-7-111-49145-3　　定　　价：119.00 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

文艺复兴以来，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的优势，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力相助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专门为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www.hzbook.com
电子邮件：hzjsj@hzbook.com
联系电话：（010）88379604
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

华章教育
华章科技图书出版中心

译者序

UNIX Internals: The New Frontiers

我与这本书结缘还是2000年年初在学校前书店的一次偶遇。当时刚刚从《电脑报》对电影《泰坦尼克号》的介绍中得知有一个开源操作系统叫Linux，便产生了一股莫名的兴奋，开始千方百计地搜寻相关材料。可惜当时几乎没有机会访问互联网，国内又没有像样的Linux专业书籍。遇到这本书并买回家的原因只是因为书名包括“UNIX”，而且是讲“内幕”的，但只记得当时是看得N头雾水，小白完全理解不了什么是LWP。

时光如白驹过隙，一晃过了10年，非科班出身的小白已经做了码农6、7年，做Linux内核工作也3年了，整理书架时再次偶然发现了这本略带灰尘的奇书，勾起了无数回忆。再度翻起后，往往每看几十页竟有打通任督二脉的感觉。去年，偶然得知机械工业出版社要再次翻译这本书，不禁心头为之一热，虽有珠玉在前，仍然挡不住跃跃欲试的冲动。

原著的序言中提到，本书刚刚出版时在操作系统领域中就独树一帜，时至今日依然如此，鲜有其他著作能出其右。我想一定有读者会觉得对于这样一本出版于上个世纪的“UNIX史记”，今天再拎出来，除了考古之外还有意义吗？对于Linux/Unix的铁杆粉丝们，本书的书名就是意义！对于其他读者，我也试试举出如下几个理由。

首先，Linux目前的流行是毋庸置疑的，但是如果你非常质疑Linux为什么是现在这个模样，那么本书应该能解答你的部分疑惑。Linux与UNIX的亲密关系决定了两者在设计和实现上有多方面的相似性。假如你看完了第14章，而没有联想到Linux内核中的VMA、page cache等概念，那一定是有什么地方不对:) 本书以追根溯源的方式探讨了现代Linux系统设计的WHY，市面上太多的书更加关注在WHAT和HOW上。在读本书时，你会有许多这样的感悟：“原来那样做是有问题的！”

其次，二三十年前设计实现本书中这些软件大师们所面对的问题，今天的无数工程师仍然在其他场合不断地再次遇到，尤其是在那些性能攸关的领域，比如在C100K甚至C10M的场景下。例如，内存池是一种通过改进空间局部性提高性能的常见手段。在第12章中，读者可以看到UNIX内核在不同的时空下，是如何从简单的资源映射分配器演化出完善Slab分配器的，其间的设计权衡经验和对多处理器环境的思考，在许多现在流行软件的设计和实现上仍然闪闪发光。

其实这本书的亮点远不只以上两个。再举个例子，读者一定了解什么是面向对象编程，并且听说过设计模式云云。在本书里你可以看到不用C++的OOP，而公元前的设计模式遍布在UNIX内核的很多核心子系统中，这两点在现代Linux内核中依然如故。类似本书这样的史籍还有《链接器和加载器》、《现代体系结构上的UNIX系统：内核程序员的SMP和Caching技术》等。

在以微博、Twitter为代表的互联网快餐阅读大潮下，许多人可以每天看几十分钟的肥皂文章，却难以坚持读完一份110页的干货“What every programmer should know about the memory”，即使后者的营养足以秒杀80%以上的互联网内容。这样，选择一本如此厚的历史书确实需要挑战自己的耐力，那么，你，敢不敢呢？

本书由三名内核开发人员共同翻译，黄庆新负责翻译第1～7章，薛磊负责翻译第

8、9、11、16、17章，李雨负责翻译第10、12、13、14、15章。非常感谢机械工业出版社的吴怡编辑对我们的信任，使得我们三个有幸翻译这本慕名已久的奇书。在本书成书期间，吴怡编辑一次又一次地容忍了译稿的延期，还不断地在解答和解决我们三个在初次翻译中所遇到的问题和困难。我在阿里云飞天团队的同事黄江伟认真检查了内存管理章节，在正确性、严谨性等方面提出了大量有份量的改进意见，对译文质量的提升有很大贡献。在翻译期间，我们三个人要么步入了婚姻的殿堂，要么即将升级为父亲，也都切身体会到翻译是一件非常消耗时间和精力的工作，这期间如果没有家人的支持是难以想象的，感谢你们的理解和耐心！

最后，经过十年IT领域的耕耘，译者深知自己水平有限，虽然原书中的一些错误已经在译文中纠正，但本书中的错误和疏漏仍然在所难免，敬请读者指正，我们的微博是：

李　雨：http://weibo.com/clocktick

薛　磊：http://weibo.com/douzhr

黄庆新：http://weibo.com/huangqingxin

李　雨

2014年11月于北京

序 言

UNIX Internals: The New Frontiers

Peter H. Salus
《*Computer Systems*》杂志主编

这个世界上，UNIX 的种类比大多数品牌冰淇淋的口味还要多。尽管整个行业推动了 X/Open 及其成员的发展，但是单一 UNIX 规范与当初宣称的目标渐行渐远。事实上，这个目标可能并不是那么重要，因为自从 Interactive System 提供了第一个商业版 UNIX 系统、Whitesmiths 提供了第一个 UNIX 克隆之后，用户社区就一直在面对各种平台上的不同实现。

UNIX 在 1969 年问世，还未满 10 岁时，各种版本就开始泛滥了。在 UNIX 20 岁之前，形成了两个阵营：自由软件基金会（Open Software Foundation）和 UNIX 国际联盟（UNIX International），以及大量的其他版本。技术上的两个主要流派分别为 AT&T（现为 Novell）和加州大学伯克利分校。Maurice Bach [Bach 86]、Sam Leffler、Kirk McKusick、Mike Karels 和 John Quarterman [Leff 89] 对这些 UNIX 进行了很好的介绍。

然而，没有一本书可以提供关于各种 UNIX 操作系统实现的总览式介绍，而本书作者 Uresh Vahalia 填补了这个空白。他对 SVR4、4.4BSD 和 Mach 技术内幕的阐述是前所未有的，书中甚至还呈现了对 Solaris、SunOS、Digital UNIX 和 HP-UX 的详尽探讨。

他的讲解非常清晰，而且不像其他某些作者那样对某种 UNIX 有所偏爱。最近，像 Linux 这种相对较新的 UNIX 克隆还在不断衍生新的 UNIX 变种，甚至伯克利 UNIX 的衍生版本也开始了新的分裂，像这样一本揭示 UNIX 内部技术并以推动 UNIX 发展为动机的著作，正是生逢其时。

1972 年 6 月，Ken Thompson 和 Denis Ritchie 出版了《UNIX Programmer's Manual》第 2 版。两位作者在这本书的引言中提到："UNIX 的安装数量增长到了 10 个，已经超出预期。"恐怕他们根本没有预料到现在所发生的情况。

我曾经在别的地方看到过对 UNIX 系统历史的介绍 [Salu94]，但是 Vahalia 在本书中对各种 UNIX 进行了真正独特和全面的对比剖析。

参考文献

[Bach 86] Bach, M.J., *The Design of the UNIX Operating System*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[Leff 89] Leffler, S.J., McKusick, M.K., Karels, M.J., and Quarterman, J.S., *The Design and Implementation of the 4.3 BSD UNIX Operating System*, Addison-Wesley, Reading, MA, 1989.

[Salu 94] Salus, P.H., *A Quarter Century of UNIX*, Addison-Wesley, Reading, MA, 1994.

[Thom 72] Thompson, K., and Ritchie, D.M., *UNIX Programmer's Manual*, Second Edition, Bell Telephone Laboratories, Murray Hill, NJ, 1972.

从20世纪70年代早期开始，UNIX系统经历了相当大的变化。它从一个小的、实验室性质的由贝尔实验室几乎免费分发的操作系统开始，逐渐成长，直到当前拥有着一群不断增长的忠实的拥护者。多年来，它从众多的学术界和行业成员中吸收了很多贡献，忍受着在其所有权以及标准化上的争执，并且进化为稳定的、成熟的操作系统。现在，有许多商业的和研究用的UNIX变种系统，每种变种系统虽然都不同，但是又足够相似，可以作为UNIX系统家族中的一种。UNIX程序员对一种UNIX变种的使用经验可以用在很多硬件平台以及其他UNIX变种上。

很多图书已经介绍过UNIX系统的不同特性，但是大多数都只是介绍诸如命令行或编程接口这些用户可见的方面，很少能介绍UNIX系统的内部实现。UNIX系统的内部实现主要是指其核心组成的部分，也就是UNIX内核研究。迄今为止，每本UNIX系统内部实现书籍只针对某一种UNIX的发行版本。比如，Bach的《*The Design of the UNIX Operating System*》[Bach 86]是介绍System V Release 2（SVR2）版本内核的标志性书籍。Leffler等著的《*The Design and Implementation of 4.3BSD UNIX Operating System*》[Leff 88]由一些主要的操作系统设计者来详尽介绍4.3BSD版本内核。Goodheart和Cox的《*The Magic Garden Explained*》著作详细介绍了System V Release 4.0（SVR4）版本的内核。

设计视角

本书从系统设计角度来审视UNIX内核。介绍了一些主流的商业性和研究性的UNIX变种。针对内核中的每个模块，本书探索其结构和设计，解释主流UNIX系统如何选择具体模块的实现方法，以及每种方法的优缺点。通过这样的对比使本书有着独特的视角，并允许读者从批判性的角度来检查系统的设计。在研究一个操作系统时，关注其优缺点是很重要的一部分。本书将通过分析一些可选的方法来介绍每种系统的优缺点。

UNIX变种

虽然本书最关注SVR4.2的内核实现，但是也详细探索了4.4BSD、Solaris 2.x、Mach以及Digital Unix系统。而且本书介绍了一些其他变种的有趣特性，这主要包括一些没有进入商业发行版的研究性变种。本书分析了从20世纪80年代中期到90年代中期UNIX系统的主要发展。为了完整性，本书同时包含了对传统UNIX功能和实现的简要介绍。在必要时，本书还提供了一个发展记录，以传统方案为起始，分析其缺点和局限并介绍现代的解决方案。

阅读对象

本书适合作为大学教材或作为专业参考书。作为大学教材，本书适用于高年级本科

生或研究生操作系统课程。本书并非一本介绍性书籍，而且不会说明诸如内核、线程以及虚拟内存这样的概念性知识。每章都包含一系列精心设计的练习题，用于进一步地思考和研究系统设计。许多练习题都是开放性的，并且需要学生进行一些额外的阅读研究。每章也有一个详尽的参考文献清单，通过该清单可以进一步阅读有关内容。

本书也可用作操作系统开发人员、应用程序开发人员以及系统管理员的专业参考书籍。操作系统设计者和架构师可以研究相近系统的内核结构、评估不同设计的优缺点以及使用本书中的一些想法来开发下一代操作系统。应用程序开发人员可以使用操作系统内核的知识，更好地利用其特征来开发更加高效的程序。系统管理员可以在理解不同参数和使用模式对系统行为的影响之后，更加有效地配置和调整系统。

组织结构

第 1 章追溯了 UNIX 系统的演变并分析了影响系统主要变化的因素。第 2 章到第 7 章介绍了进程子系统。其中第 2 章介绍在传统 UNIX 系统中（SVR3、4.3BSD 和早期变种）的进程和内核架构。第 3 章到第 7 章介绍现代系统（SVR4、4.4BSD、Solaris 2.x 以及 Digital Unix）的特性。第 3 章讨论线程及其在内核和用户库中如何实现。第 4 章介绍信号、作业控制以及登录会话管理。第 5 章介绍 UNIX 调度器和对实时应用程序的支持。第 6 章研究进程间通信的技术（IPC），包括 System V IPS 的特性集。其中还介绍了 Mach 系统架构，该系统使用 IPC 作为最基本的原语来构造内核。第 7 章讨论在现代单处理器和多处理器系统中应用的同步框架。

接下来的四章介绍了文件系统。第 8 章介绍用户可见的文件系统接口，以及定义了内核和文件系统交互的 vnode/vfs 接口。第 9 章提供一些具体的文件系统实现细节，包含原始的 System V 文件系统（s5fs）、伯克利快速文件系统（FFS）以及一些小的、利用 vnode/vfs 接口提供服务的特殊用途的文件系统。第 10 章介绍一些分布式文件系统，比如太阳微系统公司的网络文件系统（NFS）、AT&T 的远程文件共享（RFS）、卡内基梅隆大学的 Andrew 文件系统（AFS）以及 Transarc 公司的分布式文件系统（DFS）。第 11 章介绍一些使用日志提供更高可靠性和性能的高级文件系统，同时介绍一种基于可堆叠 vnode 层的新文件系统框架。

第 12 章～第 15 章介绍内存管理。第 12 章讨论内核内存分配并介绍一些有趣的分配算法。第 13 章介绍虚拟内存的概念并使用 4.3BSD 的实现来说明一些问题。第 14 章介绍 SVR4 和 Solaris 的虚拟内存架构。第 15 章介绍 Mach 和 4.4BSD 内存模型，同时分析诸如旁路转换缓冲区和虚拟地址缓存等硬件特性的影响。

最后两章主要介绍 I/O 子系统。第 16 章介绍设备驱动程序框架、内核与 I/O 子系统的交互、SVR4 设备驱动程序接口以及内核与驱动程序交互接口规范。第 17 章讨论 STREAMS 框架，用于开发网络协议、网络驱动程序和终端驱动程序。

排版约定

我在本书中遵循了一些印刷规范。首次出现的术语或概念用楷体。内部内核函数和变量名字以及代码示例都是等宽字体，比如 `ufs_lookup()`。当指定调用语法时，系统调用名是斜体，而参数是等宽字体。在图中，实线箭头代表直接指针，虚线箭头表明箭

头的源和目标对象是间接推论的。

尽管我做了最大的努力，但错误和疏忽仍然在所难免。如果你发现有什么错误，或者有任何建议，请通过电子邮件（vahalia@acm.org）发送给我。

致谢

许多人为本书的出版做出了贡献。首先，我想感谢我的儿子 Rohan 以及我的妻子 Archana，他们的耐心、爱以及无私的奉献使我得以完成本书。我利用了本属于他们的周末和夜晚时间来进行写作，而他们给予我的一向都是微笑并一直鼓励我。我也要感谢父母对我的爱和支持。

接下来，我要感谢我的朋友 Subodh Bapat，他鼓励我承担这本书的写作工作，并花费大量时间给出了专业性建议和持续性激励，非常感谢他允许我使用他的书（《Object-Oriented Networks》[Bapa 94]）中提到的工具、模块和宏，还要感谢他对我的草稿一丝不苟的审核以及对我在写作风格上清晰的指导。

许多审稿人贡献了非常多的时间和专业知识并多次审稿，提供了非常有价值的意见和建议，使得本书的质量有了提升。非常感谢 Peter Salus 的鼓励和支持以及 Benson Marguiles、Terry Lambert、Mark Ellis 和 William Bully 有深度的反馈意见。最后也要感谢 Keith Bostic、Evi Nemeth、Pat Parseghian、Steven Rago、Margo Seltzer、Richard Stevens 以及 Lev Vaitzblit 审核了本书的部分章节。

我还要感谢我的经理 Percy Tzelnic 对我在写作本书过程中的支持和理解。最后，我要感谢出版人 Alan Apt，是他促成了本书，并在每个阶段都提供了帮助。还要感谢 Prentice-Hall 和 Spectrum Publisher Services 出版社的团队成员，特别感谢 Shirlery McGuire、Sondra Chavez 和 Kelly Ricci 的支持和帮助。

参考文献

[Bach 86] Bach, M.J., *The Design of the UNIX Operating System*, Prentice-Hall, 1986.

[Bapa 94] Bapat, S.G., *Object-Oriented Networks*, Prentice-Hall, 1994.

[Good 94] Goodheart, B., and Cox, J., *The Magic Garden Explained—The Internals of UNIX System V Release 4, An Open Systems Design*, Prentice-Hall, 1994.

[Leff 89] Leffler, S.J., McKusick, M.K., Karels, M.J., and Quarterman, J.S., *The Design and Implementation of the 4.3 BSD UNIX Operating System*, Addison-Wesley, 1989.

目 录

UNIX Internals: The New Frontiers