国外著名高等院校
信息科学与技术优秀教材

UNIX 系统内幕

UNIX Internals:
The New Frontiers

英文版

Uresh Vahalia

人民邮电出版社
POSTS & TELECOMMUNICATIONS PRESS

国外著名高等院校信息科学与技术优秀教材

UNIX 系统内幕

(英文版)

UNIX Internals：The New Frontiers

Uresh Vahalia

人 民 邮 电 出 版 社

图书在版编目（CIP）数据

UNIX 系统内幕 / （美）瓦哈兰（Vahalia, U. ）著. —北京：人民邮电出版社，2003.7
国外著名高等院校信息科学与技术优秀教材
ISBN 7-115-11300-9
I. U… II. 瓦… III. UNIX 操作系统—高等学校—教材—英文 IV. TP316.81
中国版本图书馆 CIP 数据核字（2003）第 040224 号

版权声明

国外著名高等院校信息科学与技术优秀教材

UNIX 系统内幕（英文版）

◆ Uresh Vahalia
责任编辑 李 际

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号
邮编 100061 电子函件 315@ptpress.com.cn
网址 http://www. ptpress.com.cn
读者热线 010-67132705
北京汉魂图文设计有限公司制作
北京朝阳展望印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：800×1000 1/16
印张：40
字数：836 千字 2003 年 7 月第 1 版
印数：1-3 000 册 2003 年 7 月北京第 1 次印刷

著作权合同登记 图字：01 - 2003 - 1470 号

ISBN 7-115-11300-9/TP · 3467

定价：58.00 元

本书如有印装质量问题，请与本社联系 电话：(010)67129223

内容提要

本书提供了最新和最全面的 UNIX 系统技术内幕资料。全书讨论了UNIX 的开发技术和原理，考察了现代 UNIX 系统的最新发展，比较分析了由最重要的 UNIX 系统和变体所提供的最新特性。本书涵盖的内容包括体现了20世纪90年代UNIX系统结构特色的多项技术：多线程内核、多处理机和实时系统，以及分布式文件系统。本书还讨论了内核的几个重要组成部分，比较了它们在几种不同 UNIX 变体中的设计，论述了其间的诸多权衡考虑，并且介绍了已被广泛采用的那些特性。

本书可作为大学计算机科学系高年级本科生和研究生的教材或参考书。本书也为专业 UNIX 程序员、系统管理员提供了极有价值的参考资料。

出版说明

2001 年，教育部印发了《关于“十五”期间普通高等教育教材建设与改革的意见》。该文件明确指出，“九五”期间原国家教委在“抓好重点教材，全面提高质量”方针指导下，调动了各方面的积极性，产生了一大批具有改革特色的新教材。然而随着科学技术的飞速发展，目前高校教材建设工作仍滞后于教学改革的实践，一些教材内容陈旧，不能满足按新的专业目录修订的教学计划和课程设置的需要。为此该文件明确强调，要加强国外教材的引进工作。当前，引进的重点是信息科学与技术和生物科学与技术两大学科的教材。要根据专业（课程）建设的需要，通过深入调查、专家论证，引进国外优秀教材。要注意引进教材的系统配套，加强对引进教材的宣传，促进引进教材的使用和推广。

邓小平同志早在 1977 年就明确指出：“要引进外国教材，吸收外国教材中有益的东西。”随着我国加入 WTO，信息产业的国际竞争将日趋激烈，我们必须尽快培养出大批具有国际竞争能力的高水平信息技术人才。教材是一个很关键的问题，国外的一些优秀教材不但内容新，而且还提供了很多新的研究方法和思考方式。引进国外原版教材，可以促进我国教学水平的提高，提高学生的英语水平和学习能力，保证我们培养出的学生具有国际水准。

为了贯彻中央“科教兴国”的方针，配合国内高等教育教材建设的需要，人民邮电出版社约请有关专家反复论证，与国外知名的教材出版公司合作，陆续引进一些信息科学与技术优秀教材。第一批教材针对计算机专业的主干核心课程，是国外著名高等院校所采用的教材，教材的作者都是在相关领域享有盛名的专家教授。这些教材内容新，反映了计算机科学技术的最新发展，对全面提高我国信息科学与技术的教学水平必将起到巨大的推动作用。

出版国外著名高等院校信息科学与技术优秀教材的工作将是一个长期的、坚持不懈的过程，我社网站（www.ptpress.com.cn）上介绍了我们陆续推出的图书的详细情况，敬请关注。希望广大教师和学生将使用中的意见和建议及时反馈给我们，我们将根据您的反馈不断改进我们的工作，推出更多更好的引进版信息科学与技术教材。

人民邮电出版社

序

首先要感谢 Vahalia 博士为读者带来了如此厚重的礼物！但我们或许不得不承认，这并不是一本献给 UNIX 初学者的经典。

虽然介绍 UNIX 的书种类繁多，但其中的大多数都侧重讲解 UNIX 的使用和管理，真正深入探讨 UNIX 内核原理和编程的书籍却屈指可数，至于经典名作就更是凤毛麟角了。在这方面，针对 UNIX 的两大流派 System V 和 Berkeley BSD，已各有其公认的经典：《The Design of UNIX Operating System》、《The Magic Garden Explained: The Internals of UNIX System V Release 4》和《The Design and Implementation of 4.4BSD Operating System》。

在 UNIX 的发展过程中，涌现出了诸多 UNIX 变体，它们或是由专业厂商所开发和销售，或是由富于自由精神的计算机 hacker 们开发和免费提供。但有一点却是共同的，那就是它们都会从上述两大流派中汲取优秀的设计思想和最新的系统技术，充实增强自身的功能。大多数 UNIX 用户和程序员在现实中间所面对的，一般都是这类或多或少融入了两大流派优势的“混血”UNIX——Solaris/SunOS、DEC UNIX、Linux，甚至现代的 BSD 家族等等，都是如此。

正因为这样，一本既介绍传统 UNIX 的内核结构，又全面阐述内核线程、用户空间线程、信号产生处理、作业控制、进程调度、IPC、文件系统、内存管理、多处理机同步等现代 UNIX 内核所具备的主要特性，而且还可以对各种系统的同种功能进行对比剖析的书，在 1996 年本书问世之前，对于大多数 UNIX 高级用户和系统程序员来说，都是一种奢望——Vahalia 博士的大作让这一奢望成为了现实。

另一方面，由于本书介绍的系统种类多，内容丰富，在每个系统高级专题上面都是开门见山、旁征博引，而且各个专题在整体结构上看都可以独立成篇，因此对于缺乏 UNIX 使用和操作系统原理方面基础的初学者来说，往往有相当大的难度，而对于有了一定的基础、希望深入研究现代 UNIX 内核一种或者多种新特性的读者来说，读起本书来，则会有一种驾轻就熟、爱不释手的畅快之感。

至于本书的缺憾，笔者以为有二：由于 AIX 和 HP-UX 具有其特殊性，本书在这两种大致属于 SVR 4 的操作系统上着墨不多。另外，本书也没有涉及 TCP/IP

协议栈的实现，读者可以参考其他详细介绍 TCP/IP 的书籍。但是瑕不掩玉，本书依然无愧于“为任何需要了解 UNIX 操作系统各种变体之区别的人士所必备”的评价。

张　辉
癸未年初于清华园

原书序

Peter H. Salus
——《Computing Systems》杂志执行编辑

UNIX 的风格类型比不少品牌的冰激凌的口味类型还要多。尽管业界在推动 X/Open 组织及其成员促成统一的 UNIX 规范，但是做到这一点似乎仍然超出了我们的掌控能力。事实上，这可能并不是一个重要的目标。自从 Interactive Systems 公司推出第一个商业版本的 UNIX 系统，以及 Whitesmiths 公司推出第一个 UNIX 克隆版本之后，广大用户所面对的就已是运行于各种平台之上的多种 UNIX 实现了。

UNIX 诞生于 1969 年，当它出现还不到 10 年的时候，其多种版本就开始到处流传。在它有 20 年历史之前，就已经出现了竞争性的联盟组织（Open Software Foundation 和 UNIX International）以及大量的不同版本。UNIX 的两种主要流派是 AT&T（现在是 Novell）的版本，以及加州大学伯克利分校（UC Berkeley）的版本。读者很容易就可以在 Maurice Bach 的[Bach 86]以及 Sam Leffler、Kirk McKusick、Mike Karels 和 John Quaterman 的[Leff 89]这两本书中获得对这两种 UNIX 的详尽介绍。

没有哪一本书能够给感兴趣的学生一个针对各种 UNIX 操作系统实现的完整描述。Uresh Vahalia 现在做到了。他大胆涉足前人从未触及的领域，详细阐述了 SVR 4、4.4BSD 和 Mach 的内幕。此外，他还仔细讨论了 Solaris/SunOS、Digital UNIX 和 HP-UX。

他的书内容清楚明了，没有像有些作者那样，对这种或者那种 UNIX 抱有偏见。鉴于像 Linux 这样新出现的 UNIX 克隆版本已经衍生出多种变体，而 Berkeley UNIX 的诸多分支也渐渐彼此独立发展，那么像本书这样，能够揭示出促进 UNIX 成长流行的内部机制和原理，便显得极富价值。

1972 年 6 月 12 日，Ken Thompson 和 Dennis Ritchie 发表了《UNIX Programmer's Manual》第二版。在这本手册的前言里，两位作者这样写到：“UNIX 的安装数量已经增加到了 10 个，预计还会增长。”他们绝对没有预料到实际上此刻已发生了什么样的重大事件。

我在另一本书[Salu 94]里详细回顾了 UNIX 系统的历史，而 Vahalia 却在本书中给予我们对各种 UNIX 系统独到全面的对比剖析。

参考文献

[Bash 86]　Bach, M.J., The Design of UNIX Operating System, Prentice-Hall, Englewood Cliffs, NJ, 1986（本书英文影印版已由人民邮电出版社出版，书号 11246）

[Leff 89] Leffler, S.J., McKusick, M.K., Karels, M.J., and Quarterman, J.S., The Design and Implementation of 4.3BSD UNIX Operating System, Addison-Wesley, Reading, MA, 1989.（本书第二版《The Design and Implementation of 4.4BSD UNIX Operating System》的英文影印版已由人民邮电出版社出版，书号 10348）

[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.

Foreword

Peter H. Salus
Managing Editor—Computing Systems

There are more flavors of UNIX than of most brands of ice cream. Despite the industrial impetus on the part of X/Open and its members, the single UNIX specification appears to be ever-further from our grasp. In fact, it may not be an important goal. Ever since Interactive Systems produced the first commercial UNIX system and Whitesmiths produced the first UNIX clone, the user community has been confronted by a variety of implementations running on multiple platforms.

Created in 1969, UNIX was not even a decade old when versions began to proliferate. Before it was 20 years old, there were rival consortial (the Open Software Foundation and UNIX International) and a large number of versions. The two main streams were those of AT&T (now Novell) and the University of California at Berkeley. Descriptions of those UNIXes were made easily available by Maurice Bach [Bach 86] and Sam Leffler, Kirk McKusick, Mike Karels, and John Quarterman [Leff 89].

No single book offered the interested student a view of the UNIX Operating System's various implementations. Uresh Vahalia has now done this. He has gone boldly where none have gone before and elucidated the internals of SVR4, 4.4BSD, and Mach. Even more, he presents elaborate discussions of both Solaris and SunOS, Digital UNIX, and HP-UX.

He has done so clearly and without the bias that some writers have displayed toward this UNIX or that. With relatively new UNIX clones such as Linux already developing variants and even Berkeley derivatives diverging from one another, a book like this, which exposes the internals and principles that motivated UNIX's growth and popularity is of exceptional value.

On June 12, 1972, Ken Thompson and Dennis Ritchie released the *UNIX Programmer's Manual,* Second Edition. In its Preface the authors remark: "The number of UNIX installations has grown to 10, with more expected." They could never have expected what has actually happened.

I have traced the paleontology and history of the system elsewhere [Salu 94], but Vahalia has given us a truly original and comprehensive view of the comparative anatomy of the species.

References

[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.

Preface

Since the early 1970s, the UNIX system has undergone considerable metamorphosis. It started as a small, experimental operating system distributed freely (almost) by Bell Telephone Laboratories to a growing band of loyal followers. Over the years, it absorbed contributions from numerous members of academia and industry, endured battles over ownership and standardization, and evolved into its current state as a stable, mature operating system. Today there are several commercial and research variants of the UNIX system, each different from the other in many respects, yet all similar enough to be recognizable as different members of the same family. A UNIX programmer who has gained experience on one specific UNIX system can be productive on a number of different hardware platforms and UNIX variants without skipping a beat.

Hundreds of books have described various features of the UNIX system. Although most of them describe user-visible aspects such as the command shell or the programming interface, only a small number of books discuss UNIX internals. UNIX internals refers to a study of the UNIX kernel, which comprises the heart of the operating system. To date, each book on UNIX internals has focused on one specific UNIX release. Bach's *The Design of the UNIX Operating System* [Bach 86] is a landmark book on the System V Release 2 (SVR2) kernel. Leffler et al.'s *The Design and Implementation of the 4.3BSD UNIX Operating System* [Leff 88] is a comprehensive description of the 4.3BSD release by some of its principal designers. Goodheart and Cox's *The Magic Garden Explained* [Good 94] describes the internals of System V Release 4.0 (SVR4).

Design Perspectives

This book views the UNIX kernel from a system design perspective. It describes a number of mainstream commercial and research UNIX variants. For each component of the kernel, the book explores its architecture and design, how the major UNIX systems have chosen to implement the component, and the advantages and drawbacks of alternative approaches. Such a comparative treatment gives the book a unique flavor and allows the reader to examine the system from a critical viewpoint. When studying an operating system, it is important to note both its strengths and its weaknesses. This is only possible by analyzing a number of alternatives.

UNIX Variants

Although this book gives most attention to SVR4.2, it also explores 4.4BSD, Solaris 2.*x*, Mach, and Digital UNIX in detail. Further, it describes interesting features of a number of other variants, including some research that has not yet made it into commercial releases. It analyzes the major developments in UNIX from the mid-1980s to the mid-1990s. For completeness it includes a brief description of traditional UNIX functionality and implementation. Where necessary, it provides an historical treatment, starting with the traditional approach, analyzing its drawbacks and limitations, and presenting the modern solutions.

Intended Audience

UNIX Internals is useful for university courses and as a professional reference. As a university text, it is suitable for an advanced undergraduate or graduate course on operating systems. It is not an introductory book and assumes knowledge of concepts such as the kernel, processes, and virtual memory. Each chapter contains a set of exercises designed to stimulate further thought and research, and to provide additional insight into the system design. Many of the exercises are open-ended, and some require additional reading on the part of the student. Each chapter also has an exhaustive list of references, which should be useful for the student seeking to explore further.

UNIX Internals is also suitable as a professional reference for operating system developers, application programmers, and system administrators. Operating system designers and architects can use it to study the kernel architecture in contemporary systems, evaluate the relative merits and drawbacks of different designs, and use the insight to develop the next generation of operating systems. Application programmers can use the knowledge of the system internals to write more efficient programs that take better advantage of the characteristics of the operating system. Finally, system administrators can do a better job of configuring and tuning their systems by understanding how various parameters and usage patterns affect the system behavior.

Organization of the Book

Chapter 1, "Introduction," traces the evolution of the UNIX system and analyzes the factors that have influenced major changes in the system. Chapters 2 through 7 describe the process subsystem. In particular, Chapter 2 describes the process and kernel architecture in traditional UNIX systems (SVR3, 4.3BSD, and earlier variants). Chapters 3 through 7 describe features of modern UNIX systems such as SVR4, 4.4BSD, Solaris 2.*x*, and Digital UNIX. Chapter 3 discusses threads and how they are implemented in the kernel and in user libraries. Chapter 4 describes signals, job control, and login session management. Chapter 5 describes the UNIX scheduler and the growing support for real-time applications. Chapter 6 deals with interprocess communications (IPC), including the set of features known as System V IPC. It also describes the Mach architecture, which uses IPC as the fundamental primitive for structuring the kernel. Chapter 7 discusses the synchronization frameworks used in modern uniprocessor and multiprocessor systems.

The next four chapters explore file systems. Chapter 8 describes the file system interface as seen by the user, and the vnode/vfs interface that defines the interactions between the kernel and the file system. Chapter 9 provides details of some specific file system implementations, including the original System V file system (s5fs), the Berkeley Fast File System (FFS), and many small, special-purpose file systems that take advantage of the vnode/vfs interface to provide useful services. Chapter 10 describes a number of distributed file systems, namely Sun Microsystems' Network File System (NFS), AT&T's Remote File Sharing (RFS), Carnegie-Mellon University's Andrew File System (AFS), and Transarc Corporation's Distributed File System (DFS). Chapter 11 describes some advanced file systems that use journaling to provide higher availability and performance, and a new file system framework based on stackable vnode layers.

Chapters 12 through 15 describe memory management. Chapter 12 discusses kernel memory allocation and explores several interesting allocation algorithms. Chapter 13 introduces the notion of virtual memory and uses the 4.3BSD implementation to illustrate several issues. Chapter 14 describes the virtual memory architecture of SVR4 and Solaris. Chapter 15 describes the Mach and 4.4BSD memory models. It also analyzes the effects of hardware features such as translation look-aside buffers and virtually addressed caches.

The last two chapters address the I/O subsystem. Chapter 16 describes the device driver framework, the interaction between the kernel and the I/O subsystem, and the SVR4 device driver interface/driver kernel interface specification. Chapter 17 talks about the STREAMS framework for writing network protocols and network and terminal drivers.

Typographical Conventions

I have followed a small set of typographical conventions throughout this book. All system calls, library routines, and shell commands are in italics (for instance, *fork*, *fopen*, and *ls -l*). The first occurrence of any term or concept is also italicized. Names of internal kernel functions and variables, as well as all code examples, are in fixed-width font, such as `ufs_lookup()`. When specifying the calling syntax, the system call name is italicized, but the arguments are in fixed-width font. Finally, all file and directory names are in bold face (for instance, **/etc/passwd**). In the figures, solid arrows represent direct pointers, whereas a dashed arrow implies that the relationship between the source and destination of the arrow is inferred indirectly.

Despite my best efforts, some errors are inevitable. Please send me all corrections, comments, and suggestions by electronic mail at **vahalia@acm.org**.

Acknowledgments

A number of people deserve credit for this book. First of all, I want to thank my son, Rohan, and my wife, Archana, whose patience, love, and sacrifice made this book possible. Indeed, the hardest thing about writing the book was justifying to myself the weekends and evenings that could have been spent with them. They have shared my travails with a smile and have encouraged me every step of the way. I also thank my parents for their love and support.

Next, I want to thank my friend Subodh Bapat, who gave me the confidence to undertake this project. Subodh has helped me maintain focus throughout the project and has spent countless hours advising, counseling, and encouraging me. I owe him special thanks for access to the tools, templates, and macros used for his book, *Object-Oriented Networks* [Bapa 94], for his meticulous reviews of my drafts, and for his lucid discourses on writing style.

A number of reviewers contributed an incredible amount of their time and expertise to improve the book, going through several drafts and providing invaluable comments and suggestions. I want to thank Peter Salus, for his constant encouragement and support, and Benson Marguiles, Terry Lambert, Mark Ellis, and William Bully for their in-depth feedback on the content and organization of my work. I also thank Keith Bostic, Evi Nemeth, Pat Parseghian, Steven Rago, Margo Seltzer, Richard Stevens, and Lev Vaitzblit, who reviewed parts of my book.

I want to thank my manager, Percy Tzelnic, for his support and understanding throughout my project. Finally, I want to thank my publisher Alan Apt, both for proposing the book and for helping me at every stage, and the rest of the team at Prentice-Hall and at Spectrum Publisher Services, in particular, Shirley McGuire, Sondra Chavez, and Kelly Ricci, for their help and support.

References

[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.

Contents