

BSD 操作系统设计者亲自撰写

The Design and Implementation of the 4.4BSD Operating System

4.4BSD

操作系统 设计与实现 (中文版)

Marshall Kirk McKusick

Keith Bostic

Michael J. Karels

John S. Quarterman 著

李善平 刘大峰 马天驰 尹奇韡 等译



John Lasseter 94


ADDISON
WESLEY



中国电力出版社

www.infopower.com.cn

4.4BSD

操作系统 设计与实现 (中文版)

Marshall Kirk McKusick

Keith Bostic

Michael J. Karels

John S. Quarterman 著

李善平 刘大峰 马天驰 尹奇韡 等译

中国电力出版社

内 容 提 要

本书描述了 4.4BSD 的内部结构, 以及概念、数据结构和在实现 4.4BSD 系统功能时采用的算法, 侧重于 UNIX 系统伯克利变体的功能、数据结构和采用的算法。本书从 4.4BSD 的系统调用层往下——从接口到内核再到硬件。内核包含了系统功能, 例如进程管理、虚拟内存、系统 I/O、文件系统、套接字 IPC 机制和实现网络协议。除了把和终端接口及系统启动相关的内容保留了外, 系统调用层以上的内容——例如库、shell、命令、程序语言和其他用户界面都没有包括在内。

本书适用对象是: 操作系统实现者、系统程序员、UNIX 应用程序开发人员、系统管理员和有兴趣的读者。

图书在版编目(CIP)数据

4.4BSD 操作系统设计与实现/(美)麦库斯克等著;李善平等译. —北京: 中国电力出版社, 2003

ISBN 7-5083-1508-1

I .4... II .①麦... ②李... III. 计算机网络—操作系统, 4.4BSD
IV.TP316.89

中国版本图书馆 CIP 数据核字(2003)第 028781 号

著作权合同登记号 图字: 01-2000-3085 号

本书英文版原名: The Design and Implementation of the 4.4BSD Operating System
Published by arrangement with Addison Wesley Longman, Inc.

All rights reserved.

本书中文版由美国培生集团授权出版, 版权所有。

中国电力出版社出版、发行

(北京三里河路 6 号 100044 <http://www.infopower.com.cn>)

北京市铁成印刷厂印刷

各地新华书店经售

*

2003 年 7 月第 版 2003 年 7 月北京第一次印刷

787 毫米×1092 毫米 16 开本 27.5 印张 613 千字

定价 59.00 元

版 权 所 有 翻 印 必 究

(本书如有印装质量问题, 我社发行部负责退换)

译 序

我们是 UNIX 操作系统爱好者，在浙江大学从事操作系统及其相关领域的研究和开发，对 UNIX BSD、LINUX 等优秀的系统软件“情有独钟”。承担《4.4BSD 的设计与实现》一书的翻译任务，既是对我们的考验，更重要的，是给我们提供了一次仔细、系统地了解 UNIX 4.4 BSD 的机会。

《4.4BSD 的设计与实现》的翻译和校对工作由我和我的学生刘文峰、马天驰、尹奇韡、王伟波，以及王焕龙、彭刚、金海荣、张玉龙、胡玉杰、谢科先、虞浩泽、单宏伟、倪昕、彭瑞、秦延涛、胡斌、雍国恩、王鹏、张继南等共同完成。虽然经过我们的努力，稿件仍会存在这样那样的问题，但我们大家都付出了、尽力了。

李善平
于美国硅谷

前言

本书是扩充修订版，首次从权威角度全面地介绍了美国加利福尼亚州（简称加州）大学伯克利分校开发的 UNIX 系统研究版本的设计和实现。该书主要介绍 4.4BSD，4.4BSD 融合了前几个版本的改进之处。4.4BSD 包含将近 500 个除内核之外的实用程序，但本书仅集中介绍内核。

UNIX 系统

UNIX 系统在小到 PC 机，大到巨型计算机上都能运行。它可以作为多数多处理器计算机、图形处理机（graphics）和矢量处理器（vector-processing systems）的操作系统，并且由于其所具备的分时功能而受到广泛的应用。它是因特网上提供网络服务（从 FTP 到 WWW）的最通用的平台；是所有开发的操作系统中可移植性最好的系统。它的可移植性，部分归功于它的开发语言 C[Kernighan & Ritchie, 1978]（使用最广泛的可移植语言之一），部分归功于它的良好设计。它的许多功能被其他的系统所模仿[O'Dell, 1987]。

自从 1969 年开发出 UNIX 系统以来[Ritchie & Thompson, 1978]，UNIX 系统的开发经过了一系列的分分合合。最初的开发者不断地进行改进，在美国 AT&T 贝尔实验室开发出第九、第十版 UNIX，以及他们企图使之作为 UNIX 继承者的 Plan 9。同时，AT&T 批准让 UNIX 系统 V 成为商业产品，之后卖给了 Novell。Novell 把 UNIX 商标给了 X/OPEN，并把源代码和发行权卖给了 Santa Cruz Operation 公司（SCO）。系统 V 和第九版的 UNIX 都明显地受到由美国加州大学伯克利分校计算机系统研究小组（CSRG）开发的 BSD（Berkeley Software Distributions）的影响。

伯克利软件产品

这些伯克利系统使 UNIX 家族增加了一些有用的程序和功能：

- 2BSD（伯克利 PDP-11 系统）：文本编辑器 vi。
- 3BSD（第一个伯克利 VAX 系统）：支持按需分页虚拟内存。
- 4.0BSD：性能提高。
- 4.1BSD：任务控制、自动配置、长型 C 标识符。
- 4.2 和 4.3BSD：可靠的信号量；快速的文件系统；更高的网络性能，包括标准的 TCP/IP 实现；完善的进程间通信（IPC）原语，以及更多的性能提高。
- 4.4BSD：一个新的虚拟内存系统；一个可重叠和扩展的 vnode 接口；一个网络文件系统（NFS）；一个日志结构的文件系统，众多的文件类型，包括 loopback、联合、uid/gid 映射层；ISO9660 文件系统（例如 CD-ROM）；ISO 网络协议；支持 68K、SPARC、MIPS 和 PC 体系结构；支持 POSIX，包括 termios、会话和大多的功能；

每个接口可有多个 IP 地址；磁盘标志，以及启动效率的提高。

4.2BSD、4.3BSD 和 4.4BSD 是许多商业 UNIX 系统的基础，并被许多其他厂商的开发小组内部使用。在 BSD 系统上的许多改进也被合并到 System V 中，有的是由那些产品基于 SystemV 的商家添加进来的。

在 4.2BSD 和 4.3BSD 中，TCP/IP 网络协议组的实现和系统的可用性，解释了为什么 TCP/IP 网络协议组能够在全世界广泛地应用。许多厂商的产品都和伯克利网络兼容，不论它们是基于 4.2BSD、4.3BSD、4.4BSD、系统 V、DEC 的 VMS，还是微软在 Windows 95 和 Windows NT 上的 Winsock 接口。

4BSD 对 POSIX (IEEE Std 1003.1) 操作系统接口标准和相关的标准都产生了很深远的影响。一些 4.3BSD 的特征（如可靠信号量、任务控制、每个进程有多个存取组和路径操作例程）都被 POSIX 所采用。

本书涉及的内容

本书描述了 4.4BSD[Quarterman et al.1985]的内部结构，以及概念、数据结构和在实现 4.4BSD 系统功能时采用的算法。书的深度和 Bach 写的关于 UNIX SystemV[Bach, 1986]的那本书差不多；然而本书侧重于 UNIX 系统伯克利变体的功能、数据结构和采用的算法。本书从 4.4BSD 的系统调用层往下——从接口到内核再到硬件。内核包含了系统功能，例如进程管理、虚拟内存、系统 I/O、文件系统、套接字 IPC 机制和实现网络协议。除了把和终端接口及系统启动相关的内容保留了外，系统调用层以上的内容——例如库、Shell、命令、程序语言和其他用户接口都没有包括在内。像 Organick 写的关于 Multics[Organick, 1975]的书一样，本书深入介绍了一个流行的操作系统。

当涉及到特殊硬件时，本书参考了 Hewlett-Packard HP300（基于 Motorola 68000）的体系结构。因为 4.4BSD 是在 HP300 上开发的，所以完全支持 HP300。HP300 为 4.4BSD 提供了一个方便的参考点。

本书适用对象是：操作系统实现者、系统程序员、UNIX 应用程序开发人员、系统管理员和有兴趣的读者。本书可以和系统源代码结合起来读，细节处也可参考操作手册[CSRG, 1994]和代码。但本书既不是 UNIX 编程手册也不是用户教程（教程见[Libes & Ressler, 1988]）。熟悉 UNIX 系统的一些版本（见[Kernighan & Pike, 1984]）和 C 程序设计语言（见[Kernighan & Ritchie, 1988]）对阅读本书会有帮助。

操作系统课程的使用方法

本书适合作为高级操作系统课的参考书，以提供相关的背景知识。它不是用来作为入门类的操作系统教程的，读者应该已经接触了像内存管理、进程调度、I/O 系统[Silberschatz & Galvin, 1994]这类术语。熟悉网络协议的概念[Tanenbaum, 1988; Stallings, 1993; Schwarzl, 1987]对理解后面章节会有帮助。

每章后面都有习题。习题分为 3 个难度，用零个、一个或者两个星号表示。没有星号的习题的答案可以在书中找到；标有一个星号的习题除了参考书中的概念外，还需要进一步地推理；标有两个星号的习题表示是较大的设计方案或开放的研究问题。

书的结构

本书讨论原理和设计的问题，以及实际实现的细节。通常从系统调用层开始讨论，往下再到内核。使用表和图来清楚表示数据结构和控制流。使用类似 C 语言的伪码来表示算法。用**黑体字**表示程序名和文件系统的路径名。*斜体字*表示术语表中出现的术语并标志系统调用名、变量、例程和结构名。例程名（不是系统调用）后面都跟有一对圆括号（例如：`malloc()` 是例程名，而 `argv` 是变量名）。

本书分成以下五大部分：

- **第一部分，综述** 前面三章介绍了操作系统的内容和本书后面所要介绍的内容的背景知识。第一章综述了系统的历史发展，强调了系统的研究方向。第二章介绍了系统所提供的服务，简要介绍了内核的内部结构。也讨论了开发系统时所做的一些设计决策。其中，2.3 节和 2.4 节分别是第三、四章的概述。第三章解释了如何进行系统调用，并详细介绍了内核提供的一些基本服务。
- **第二部分，进程** 第四章是后面章节的基础，介绍了进程的结构、进程调度的算法和系统用来保持内核驻留数据结构访问一致性的同步机制。第五章详细讨论了虚拟内存管理系统。
- **第三部分，I/O 系统** 第六章讲解了 I/O 设备的系统接口并描述了支持该接口的功能结构。接下来的四章介绍了 I/O 系统主要部分的细节。第七章详细介绍了实现文件系统的数据结构和算法。第八章介绍了本地文件系统是如何与本地的存储介质联系起来的。第九章从服务端和客户端两方面介绍了网络文件系统。第十章讨论了对字符终端的支持，描述了面向字符的设备驱动程序。
- **第四部分，进程间通信** 第十一章介绍了相关或不相关进程之间的通信机制。第十二、十三章，这是相关联的，因为第十二章所介绍的功能是通过特殊的协议所实现的，例如 TCP/IP 协议组，协议在第十三章介绍。
- **第五部分，系统操作** 第十四章讨论了系统的启动、关闭和配置，介绍了系统在进程层的初始化，从内核初始化到用户登录。

建议按顺序阅读本书，但除第一部分外，其他部分是独立的，可以分开来看。第十四章应该最后看，专业知识丰富的读者会发现这一章很有用。

本书的最后是术语表，有主要术语的定义。每章包括了参考文献，有相关材料的引用。

如何得到 4.4BSD

当前关于 4.4BSD 的源代码的信息可以在以下站点找到。到出版时为止，Walnut Creek CDROM 可以提供 4.4BSD 的源码，包括 Lite Release 2 系统，4.4BSD 的 FreeBSD 版本（专门为 PC 及其兼容机编译和运行的）。Walnut Creek 的联系方式是：电话 1-800-786-9907，E-mail: orders@cdrom.com，网址 <http://www.cdrom.com/>。市面上发行的 NetBSD 是编译过的，可以在多数的工作站上运行。联系 NetBSD 项目组可以获得更多的信息，E-mail: majordomo@NetBSD.ORG（发送邮件正文信息：“lists”），网址 <http://www.NetBSD.ORG/>。发行的 OpenBSD 也是编译过的，可以在多种工作站上运行，并且其可靠性和安全性经过了多方面的检查。访问 OpenBSD 项目组的站点 <http://www.OpenBSD.org> 可以获得更多的信息。

功能齐全的商业版本 BSD/OS 可以从 Berkeley Software Design 得到, 公司电话 1-800-800-4273, E-mail: bsdi-info@bsdi.com, 网址 <http://www.bsdi.com/>。4.4BSD 的使用手册由 Usenix 和 O'Reilly 公司联合发行。O'Reilly 单本或成套销售这五卷书 (ISBN 1-56592-082-1), 电话 1-800-889-8969, E-mail: order@ora.com, 网址 <http://www.ora.com/>。

如果你看完了前言, 你或许能得到 T 恤作为奖励, T 恤上的图案是 John Lasseter 专门为本书封面所作的作品的复制品。对的! 他就是 Walt Disney/Pixar 的那个 John Lasseter, “Toy story”就是他的代表作。这些 T 恤是给那些为本书出版出过力的和那些首先指出书中错误的人而准备的。现在你有可能看到这些 T 恤与原来的不同之处(所以你不要失去可以得到这些别人很难得到的珍品的机会)。要获得进一步的相关信息, 请写信到:

M. K. McKusick
1614 Oxford St.
Berkeley, CA 94709-1608

请在信封上写明你的地址(美国居民请附上回复邮资)。你还可以访问“BSD T 恤的历史”网页, <http://www.mckusick.com/beastie/index.html>。

感谢

在此我们要特别感谢下列人员: Mike Hibler (University of Utah) 参与编写了第五章的内存管理方面的内容; Rick Macklem (University of Guelph), 他的 NFS 论文为第九章提供了许多材料。

感谢下列人员阅读本书并对本书提出意见: Paul Abrahams (顾问), Susan LoVerso (Orca Systems), George Neville—Neil (Wind River Systems) 和 Steve Stepanek (California State University, Northridge)。

感谢下列人员阅读本书初稿并提出意见: Eric Allman (Pangaea Reference Systems), Eric Anderson (University of California at Berkeley), Mark Andrews (Alias Research), Mike Beede (Secure Computing Corporation), Paul Borman (Berkeley Software Design), Peter Collinson (Hillside Systems), Ben Cottrell (NetBSD user), Patrick Cua (De La Salle University, Philippines), John Dyson (The FreeBSD Project), Sean Eric Fagan (BSD developer), Mike Fester (Medieus Systems Corporation), David Greenman (The FreeBSD Project), Wayne Hathaway (Auspex Systems), John Heidemann (University of California at Los Angeles), Jeff Honig (Berkeley Software Design), Gordon Irlam (Cygnus Support), Alan Langerman (Orca Systems), Sam Leffler (Silicon Graphics), Casimir Lesiak (NASA/Ames Research Center), Gavin Lim (De La Salle University, Philippines), Steve Lucco (Carnegie Mellon University), Jan-Simon Pendry (Sequent, UK), Arnold Robbins (Georgia Institute of Technology), Peter salus (UNIX historian), Wayne Sawdon (Carnegie Mellon University), Margo Seltzer (Harvard University), Keith Sklower (University of California at Berkeley), Keith Smith (Havard University) 以及 Humprey C.Sy (De La Salle University, Philippines)。

本书的编排上使用了 James Clark 的 pic、tbl、eqn 以及 groff。美工方面主要使用了 xfig。图片的排版以及段中短行的消除是由 groff 的宏 (macros) 来完成的, 但消除段中行被分离出

来的情况，以及偶数页底部的处理是手工完成的。

我们鼓励读者将您对于本书技术上或其他方面的建议和意见发送给我们，请发电子邮件至 bsdbook-bugs@McKusick.COM。

参考文献

Bach, 1986.

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

Bentley & Kernighan, 1986.

J. Bentley & B. Kernighan, "Tools for Printing Indexes," Computing Science Technical Report 128, AT&T Bell Laboratories, Murray Hill, NJ, 1986.

CSRG, 1994.

CSRG, in *4.4 Berkeley Software Distribution*, O'Reilly & Associates, Inc., Sebastopol, CA, 1994.

Kernighan & Pike, 1984.

B. W. Kernighan & R. Pike, *The UNIX Programming Environment*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

Kernighan & Ritchie, 1978.

B. W. Kernighan & D. M. Ritchie, *The C Programming Language*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

Kernighan & Ritchie, 1988.

B. W. Kernighan & D. M. Ritchie, *The C Programming Language*, 2nd ed, Prentice-Hall, Englewood Cliffs, NJ, 1988.

Libes & Ressler, 1988.

D. Libes & S. Ressler, *Life with UNIX*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

O'Dell, 1987.

M. O'Dell, "UNIX: The World View," *Proceedings of the 1987 Winter USENIX Conference*, pp. 35–45, January 1987.

Organick, 1975.

E. I. Organick, *The Multics System: An Examination of Its Structure*, MIT Press, Cambridge, MA, 1975.

Quarterman et al, 1985.

J. S. Quarterman, A. Silberschatz, & J. L. Peterson, "4.2BSD and 4.3BSD as Examples of the UNIX System," *ACM Computing Surveys*, vol. 17, no. 4, pp. 379–418, December 1985.

Ritchie & Thompson, 1978.

D. M. Ritchie & K. Thompson, "The UNIX Time-Sharing System," *Bell System Technical Journal*, vol. 57, no. 6, Part 2, pp. 1905–1929, July–August 1978. The original version [*Comm. ACM* vol. 7, no. 7, pp. 365–375 (July 1974)] described the 6th edition; this citation describes the 7th edition.

Schwartz, 1987.

M. Schwartz, *Telecommunication Networks*, Series in Electrical and

- Computer Engineering, Addison-Wesley, Reading, MA, 1987.
- Silberschatz & Galvin, 1994.
- A. Silberschatz & P. Galvin, *Operating System Concepts*, 4th Edition, Addison-Wesley, Reading, MA, 1994.
- Stallings, 1993.
- R. Stallings, *Data and Computer Communications*, 4th Edition, Macmillan, New York, NY, 1993.
- Tanenbaum, 1988.
- A. S. Tanenbaum, *Computer Networks*, 2nd ed, Prentice-Hall, Englewood Cliffs, NJ, 1988.

目 录

译 序
前 言

第一部分 综 述

第一章 BSD 系统的历史和目标	3
1.1 UNIX 系统的历史	3
1.2 BSD 和其他系统	9
1.3 4BSD 的设计目标	11
1.4 系统的发布	13
参考文献	14
第二章 4.4BSD 设计综述	18
2.1 4.4 BSD 模块与内核	18
2.2 内核结构	19
2.3 内核提供的服务	21
2.4 进程管理	22
2.5 内存管理	24
2.6 I/O 系统	26
2.7 文件系统	30
2.8 文件库 (filestore)	33
2.9 网络文件系统	33
2.10 终端	34
2.11 进程间通信	35
2.12 网络通信	35
2.13 网络实现	36
2.14 系统操作	36
习题	36
参考文献	37

第三章 内核服务	39
3.1 内核组织	39
3.2 系统调用	42
3.3 陷阱和中断	43
3.4 时钟中断	45
3.5 内存管理服务	47
3.6 时间服务	50
3.7 用户、组和其他身份标志	51
3.8 资源服务	54
3.9 系统操作服务	55
习题	56
参考文献	57

第二部分 进 程

第四章 进程管理	61
4.1 进程管理概述	61
4.2 进程状态	63
4.3 上下文切换	68
4.4 进程调度	72
4.5 进程创建	76
4.6 进程终止	77
4.7 信号	78
4.8 进程组和会话	84
4.9 进程调试	87
习题	88
参考文献	90
第五章 内存管理	91
5.1 术语	91
5.2 4.4BSD 虚拟内存系统综述	95
5.3 内核内存管理	97
5.4 进程独立拥有的资源	101
5.5 共享内存	105
5.6 新进程的创建	112

5.7 文件的执行	114
5.8 进程地址空间的操作.....	115
5.9 进程的终止	117
5.10 分页器接口	118
5.11 分页	123
5.12 页面交换	126
5.13 可移植性	131
习题	142
参考文献	143

第三部分 I/O 系统

第六章 I/O 系统综述.....	147
6.1 从用户到设备的 I/O 映射	147
6.2 块设备	150
6.3 字符设备	153
6.4 描述符管理和服务.....	156
6.5 虚拟文件系统的接口.....	166
6.6 独立于文件系统的服务.....	170
6.7 可堆叠 (stackable) 的文件系统.....	177
习题	182
参考文献	183
第七章 本地文件系统	184
7.1 文件系统的分层管理.....	184
7.2 inode 的结构.....	185
7.3 命名	188
7.4 配额	193
7.5 文件锁定	196
7.6 其他文件系统机制.....	200
习题	202
参考文献	202
第八章 本地文件库.....	203
8.1 文件库概述	203

8.2 Berkeley 快速文件系统	206
8.3 日志结构的文件系统	217
8.4 基于内存的文件系统	230
习题	233
参考文献	234
第九章 网络文件系统	237
9.1 历史和概况	237
9.2 NFS 结构和操作	239
9.3 提高性能的技术	249
习题	255
参考文献	256
第十章 终端处理	258
10.1 终端处理模式	258
10.2 行规程	259
10.3 用户接口	260
10.4 tty 数据结构	261
10.5 进程组、会话和终端控制	263
10.6 C-list	263
10.7 RS-232 和调制解调器控制	265
10.8 终端操作	265
10.9 其他的行规程	272
习题	273
参考文献	273

第四部分 进程间通信

第十一章 进程间通信	277
11.1 进程间通信模型	277
11.2 实现结构和概述	282
11.3 内存管理	283
11.4 数据结构	287
11.5 连接建立	291
11.6 数据传送	293

11.7 关闭 socket	299
习题	300
参考文献	301
第十二章 网络通信	302
12.1 内部结构	302
12.2 socket 到协议的接口	309
12.3 协议到协议的接口	313
12.4 协议到网络的接口	315
12.5 路由	318
12.6 缓存和拥塞控制	326
12.7 原始 socket	327
12.8 其他的网络子系统主题	329
习题	331
参考文献	331
第十三章 网络协议	333
13.1 Internet 网络协议	333
13.2 用户数据报协议 (UDP)	339
13.3 互联网协议 (IP)	342
13.4 传输控制协议 (TCP)	345
13.5 TCP 算法	350
13.6 TCP 输入处理	355
13.7 TCP 输出处理	358
13.8 Internet 控制报文协议 (ICMP)	365
13.9 OSI 实现中的问题	367
13.10 联网和进程间通信综述	368
习题	371
参考文献	372

第五部分 系统操作

第十四章 系统启动	377
14.1 概述	377
14.2 自举引导	378

14.3 内核的初始化.....	379
14.4 自动配置	381
14.5 独立于机器的初始化.....	385
14.6 用户级初始化.....	387
14.7 系统启动的相关话题.....	389
习题	392
参考文献	392
 术语表	 393



第一部分

综 述