HZ BOOKS
华章科技

PEARSON

华章专业开发者丛书

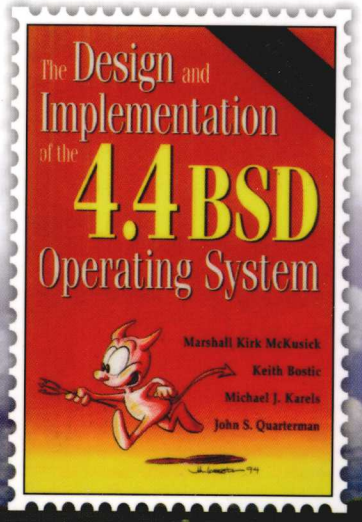
4.4BSD

操作系统设计与实现

The Design and Implementation of the 4.4BSD Operating System

（美）Marshall Kirk McKusick
Keith Bostic
Michael J. Karels
John S. Quarterman
著

李善平　刘文峰　马天驰　等译



机械工业出版社
China Machine Press

华章专业开发者丛书

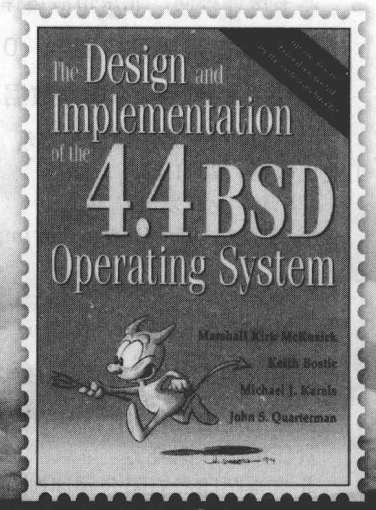
4.4BSD

操作系统设计与实现

The Design and Implementation of the 4.4BSD Operating System

(美) Marshall Kirk McKusick
Keith Bostic
Michael J. Karels
John S. Quarterman
著

李善平　刘文峰　马天驰　等译



机械工业出版社
China Machine Press

本书描述了4.4BSD的内部结构、概念、数据结构以及在实现4.4BSD系统功能时采用的算法，侧重于UNIX系统伯克利版本的功能、数据结构和采用的算法。本书从4.4BSD的系统调用层往下讲述，从接口到内核再到硬件。内核包含系统功能，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。除此之外，本书还详细地介绍了进程和内存管理的变化，描述了新的文件系统接口，更新了网络和进程间通信的相关信息。本书适合操作系统实现者、系统程序员、UNIX应用程序开发人员、系统管理员和对操作系统感兴趣的读者阅读。

本书版权登记号：图字：01-2010-6349

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（McKusick，M.K.）等著；李善平等译. —北京：机械工业出版社，2011.12

（华章专业开发者丛书）

书名原文：The Design and Implementation of the 4.4BSD Operating System

ISBN 978-7-111-36647-8

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（2011）第246815号

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：谢晓芳

北京京师印务有限公司印刷

2012年1月第1版第1次印刷

186mm×240mm • 26印张

标准书号：ISBN 978-7-111-36647-8

定价：79.00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

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@hzbook.com

前 言

本书是修订版，首次权威和完整地介绍了加利福尼亚大学伯克利分校开发的UNIX系统研究版本的设计和实现。本书主要介绍4.4BSD，4.4BSD融合了前几个版本改进的地方。虽然4.4BSD包含除内核之外的将近500个实用程序，但本书仅集中介绍内核。

UNIX系统

UNIX系统在小到家用计算机系统，大到巨型计算机上都能运行。它可以作为多数多处理器、图形和矢量处理系统的操作系统；是互联网上提供网络服务（从FTP到WWW）的最通用的平台；是所有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实现语言C[Kernighan & Ritchie，1978]（使用最广泛的可移植语言之一），部分归功于系统良好的设计。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。最初的开发者不断改进，在AT&T贝尔实验室开发出第9版和第10版UNIX，以及他们试图使之作为UNIX继承者的Plan 9。同时，AT&T批准了UNIX系统V作为一个商业产品，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9版本的UNIX都明显地受到由加利福尼亚大学伯克利分校计算机系统研究小组（Computer Systems Research Group，CSRG）开发的BSD（Berkeley Software Distribution，伯克利软件版本）的影响。

伯克利软件版本

这些伯克利系统在UNIX社区中引入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：作业控制、自动配置和长C标识符。
- 4.2BSD和4.3BSD：可靠的信号；快速的文件系统；改进网络性能，包括TCP/IP的一个引用实现；完善的进程间通信（Inter Process Communication，IPC）原语；以及更多的性能提升。
- 4.4BSD：一个新的虚拟内存系统；一个可堆叠和可扩展的vnode接口；一个网络文件系统（Network FileSystem，NFS）；一个日志结构的文件系统，众多的文件类型，包括回环、

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、会话和大多数的功能；每个接口可有多个 IP 地址；磁盘标签和启动效率的提高。

4.2BSD、4.3BSD 和 4.4BSD 是许多商业 UNIX 系统的基础，并被许多其他供应商的开发小组内部使用。在 BSD 系统上的许多改进也合并到系统 V 中，或者由其产品基于系统 V 的供应商添加进来。

在 4.2BSD 和 4.3BSD 中，TCP/IP 中网络协议套的实现和那些系统的可用性，解释了为什么 TCP/IP 的网络协议套在全世界广泛应用。许多供应商都采用了伯克利网络实现，不论它们的基本系统是 4.2BSD、4.3BSD、4.4BSD、系统 V，或是 DEC（Digital Equipment Corporation，数字设备公司）的 VMS，或是微软在 Windows 95 和 Windows/NT 上的 Winsock 接口。

4BSD 对 POSIX 操作系统接口标准（IEEE Std 1003.1）和相关的标准也产生了很大的影响。4.3BSD 的一些特征（如可靠信号、作业控制、每个进程的多个存取组和目录操作的例程）都被 POSIX 所采用。

本书涉及的内容

本书是关于 4.4BSD [Quarterman et al，1985] 的内部结构的，关于概念、数据结构和在实现 4.4BSD 系统功能时采用的算法。本书的深度和 Bach 写的关于系统 V [Bach，1986] 的那本书差不多；然而本书侧重于 UNIX 操作系统的伯克利版本的功能、数据结构和采用的算法。本书从系统调用层往下——从接口到内核再到硬件概述 4.4BSD。内核包括系统功能，例如进程管理、虚拟内存、系统 I/O、文件系统、套接字 IPC 机制和网络协议实现。排除了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，除了和终端接口及系统启动相关的内容。像 Organick 写的关于 Multics [Organick，1975] 的书，本书深入介绍了一个当代操作系统。

当涉及特殊硬件时，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，对 HP300 的体系结构完全支持，所以它提供了一个方便的参考点。

本书读者对象是：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，介于手册 [CSRG，1994] 和详细处理的代码之间。但本书既不是 UNIX 编程手册也不是用户教程（对于教程，见文献 [Libes & Ressler，1988]）。熟悉 UNIX 系统的一些版本（例如，见文献 [Kernighan & Pike，1984]）和 C 程序设计语言（例如，见文献 [Kernighan & Ritchie，1988]）对阅读本书会有帮助。

对学习操作系统课程的作用

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。它不是入门类的操作系统教材，本书假设读者应该已经接触过像内存管理、进程调度和 I/O 系统 [Silberschatz & Galvin，1994] 这类术语。熟悉网络协议的概念 [Tanenbaum，1988；Stallings，1993；Schwarlz，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，分别用零个、一个或两个星号表示。没有星号的习题的答案可以在本书中找到；标有一个星号的习题除了参考书中的概念外，还需要进一步推理；标有两个星号的习题表示是主要的设计方案或开放的研究问题。

本书结构

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语言的伪代码来表示算法。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）。

本书分成以下五部分：

第一部分　综述　这一部分的3章介绍了完整的操作系统的内容和本书下面所要介绍的内容。第1章综述了BSD系统的发展史，强调了系统的研究方向。第2章介绍了系统所提供的服务，概括介绍了内核的内部结构。它也讨论了开发系统时所做的一些设计决策。其中，2.3～2.14节分别是其对应章节的概述。第3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4章是后面章节的基础，介绍了进程的结构，调度进程执行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5章详细讨论了虚拟内存管理系统。

第三部分　I/O系统　第6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4章介绍了I/O系统主要部分的细节。第7章从应用程序的角度详细介绍了实现文件系统的数据结构和算法。第8章介绍了本地文件系统是如何与本地存储介质联系起来的。第9章从服务器和客户端两方面介绍了网络文件系统。第10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　通信　第11章介绍了相关或不相关进程之间的通信机制。第12～13章是相关联的，因为前者所介绍的功能是通过特定的协议所实现的，例如TCP/IP协议套，协议在第13章介绍。

第五部分　系统操作　第14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14章应该最后看，专业知识丰富的读者会发现单独的这一章很有用。

本书的最后是术语表，有主要术语的定义。每章最后包括了参考资源。

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的源代码，包括Lite Release 2系统，以及4.4BSD的FreeBSD版本（经过编译并可以运行在与PC兼容的硬件上）。Walnut Creek的联系方式是，电话1-800-786-9907，E-mail地址orders@cdrom.com，网址http://www.cdrom.com/。NetBSD版本是编译过的，可以在多数工作站体系结构上运行。联系NetBSD项目组可以获得更多的信息，E-mail地址

majordomo@NetBSD.ORG（发送“列表”邮件正文），网址 http://www.NetBSD.ORG/。OpenBSD 版本是编译过的，可以在多种工作站体系结构上运行，并且对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 公司得到，公司电话 1-800-800-4273，E-mail 地址 bsdi-info@bsdi.com，网址 http://www.bsdi.com/。4.4BSD 的使用手册由 Usenix 和 O'Reilly 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 地址 order@ora.com，网址 http://www.ora.com/。

致谢

我们在此感谢下列人员：Mike Hibler（犹他大学），他写了第 5 章的内容；Rick Mackiem（贵湖大学），他的 NFS 论文为第 9 章提供了许多材料。

我们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（加利福尼亚州立大学北岭分校）。

我们感谢下列人员阅读本书初稿并提出意见：Eric Allman（Pangaea Reference Systems）、Eric Anderson（加利福尼亚大学伯克利分校）、Mark Andrews（Alias Research）、Mike Beede（Secure Computing Corporation）、Paul Borman（Berkeley Software Design）、Peter Collinson（Hillside Systems）、Ben Cottrell（NetBSD 用户）、Patrick Cua（De La Salle University，Philippines）、John Dyson（FreeBSD 项目），Sean Eric fagan（BSD 开发者）、Mike Fester（Medieus Systems Corporation）、David Greenman（FreeBSD 项目）、Wayne Hathaway（Auspex Systems）、John Heidemann（加利福尼亚大学洛杉矶分校）、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（卡耐基梅隆大学）、Jan-Simon Pendry（Sequent、UK）、Arnold Robbins（乔治亚技术学院）、Peter salus（UNIX 历史学家）、Wayne Sawdon（卡耐基梅隆大学）、Margo Seltzer（哈佛大学）、Keith Sklower（加利福尼亚大学洛杉矶分校）、Keith Smith（哈佛大学），以及 Humprey C.Sy（De La Salle University，Philippines）。

本书在编排上使用了 James Clark 的 pic、tbl、eqn 以及 groff。美工方面主要使用了 xfig。图片的排版以及窗口消除是由 groff 的宏来完成的，但消除孤行以及每页底部的处理是手动完成的。

我们鼓励读者将本书的错误和修改意见发送给我们；请发电子邮件至 bsdbook-bugs@McKusick.COM。

参考资源

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

译者序
前　言

第一部分　综述

第 1 章　BSD 系统的历史和目标 …… 1
1.1　UNIX 系统的历史 …… 1
1.1.1　UNIX 系统的起源 …… 1
1.1.2　UNIX 系统的研究与发展 …… 3
1.1.3　AT&T 的 UNIX 系统Ⅲ和系统 V …… 5
1.1.4　其他组织 …… 5
1.1.5　关于 BSD 系统 …… 5
1.1.6　UNIX 世界 …… 6
1.2　BSD 和其他系统 …… 7
1.3　4BSD 的设计目标 …… 9
1.3.1　4.2BSD 设计目标 …… 9
1.3.2　4.3BSD 设计目标 …… 10
1.3.3　4.4BSD 设计目标 …… 10
1.4　系统的发布 …… 11
参考资源 …… 13
第 2 章　4.4BSD 设计综述 …… 16
2.1　4.4 BSD 模块与内核 …… 16
2.2　内核结构 …… 17
2.3　内核提供的服务 …… 19
2.4　进程管理 …… 19
2.4.1　信号 …… 20
2.4.2　进程组和会话 …… 21
2.5　内存管理 …… 22
2.5.1　BSD 内存管理设计要点 …… 22
2.5.2　内核中的内存管理 …… 23
2.6　I/O 系统 …… 24
2.6.1　描述符与 I/O …… 24
2.6.2　描述符管理 …… 25
2.6.3　设备 …… 25
2.6.4　套接字 IPC …… 26
2.6.5　分散 / 聚集 I/O …… 27
2.6.6　多文件系统支持 …… 27
2.7　文件系统 …… 27
2.8　文件库（filestore） …… 30
2.9　网络文件系统 …… 31
2.10　终端 …… 32
2.11　进程间通信 …… 32
2.12　网络通信 …… 33
2.13　网络实现 …… 33
2.14　系统操作 …… 34
习题 …… 34
参考资源 …… 34
第 3 章　内核服务 …… 36
3.1　内核组织 …… 36
3.1.1　系统进程 …… 36
3.1.2　系统入口 …… 36
3.1.3　内核的运行时结构 …… 37
3.1.4　内核的入口 …… 38
3.1.5　内核的返回 …… 39
3.2　系统调用 …… 39
3.2.1　结果处理 …… 39
3.2.2　系统调用的返回 …… 40
3.3　陷阱和中断 …… 40
3.3.1　陷阱 …… 40
3.3.2　I/O 设备中断 …… 41
3.3.3　软件中断 …… 41

3.4 时钟中断 …… 42
3.4.1 统计和进程调度 …… 42
3.4.2 超时 …… 43
3.5 内存管理服务 …… 44
3.6 时间服务 …… 46
3.6.1 标准时间 …… 47
3.6.2 调整时间 …… 47
3.6.3 外部表示 …… 47
3.6.4 间隔时间 …… 47
3.7 用户、组和其他标识符 …… 48
3.7.1 主机标识符 …… 50
3.7.2 进程组和会话 …… 50
3.8 资源服务 …… 51
3.8.1 进程优先级 …… 51
3.8.2 资源利用 …… 51
3.8.3 资源限制 …… 51
3.8.4 文件系统配额 …… 52
3.9 系统操作服务 …… 52
习题 …… 53
参考资源 …… 54

第二部分 进程

第 4 章 进程管理 …… 55
4.1 进程管理概述 …… 55
4.1.1 多程序机制 …… 56
4.1.2 调度 …… 56
4.2 进程状态 …… 57
4.2.1 进程结构 …… 58
4.2.2 用户结构 …… 61
4.3 上下文切换 …… 62
4.3.1 进程状态 …… 63
4.3.2 底层上下文切换 …… 63
4.3.3 主动上下文切换 …… 63
4.3.4 同步 …… 65
4.4 进程调度 …… 67
4.4.1 进程优先级的计算 …… 67
4.4.2 进程优先级例程 …… 68
4.4.3 进程运行队列和上下文切换 …… 69
4.5 进程创建 …… 70
4.6 进程终止 …… 72
4.7 信号 …… 72
4.7.1 与 POSIX 信号的比较 …… 74
4.7.2 发送信号 …… 75
4.7.3 传递信号 …… 77
4.8 进程组和会话 …… 78
4.8.1 会话 …… 79
4.8.2 作业控制 …… 80
4.9 进程调试 …… 81
习题 …… 83
参考资源 …… 84
第 5 章 内存管理 …… 85
5.1 术语 …… 85
5.1.1 进程与内存 …… 86
5.1.2 分页 …… 86
5.1.3 替换算法 …… 87
5.1.4 工作集模型 …… 87
5.1.5 交换 …… 88
5.1.6 虚拟内存的优点 …… 88
5.1.7 虚拟内存的硬件要求 …… 88
5.2 4.4BSD 虚拟内存系统综述 …… 89
5.3 内核内存管理 …… 91
5.3.1 内核映射和子映射 …… 91
5.3.2 内核地址空间的分配 …… 92
5.3.3 内核内存分配 …… 93
5.4 进程独立拥有的资源（Per-Process Resource） …… 95
5.4.1 4.4BSD 进程虚拟地址空间 …… 95
5.4.2 缺页调度 …… 96
5.4.3 映射对象 …… 97
5.4.4 对象 …… 98
5.4.5 页对象 …… 98
5.5 共享内存 …… 99
5.5.1 mmap 模型 …… 100
5.5.2 共享映射 …… 102

5.5.3 私有映射……102
5.5.4 压缩影子链……104
5.5.5 私有快照……105
5.6 新进程的创建……106
5.6.1 保留内核资源……106
5.6.2 复制用户地址空间……107
5.6.3 不采用复制技术创建新进程……108
5.7 文件的执行……108
5.8 进程地址空间的操作……109
5.8.1 进程大小的改变……109
5.8.2 文件映射……110
5.8.3 改变保护机制……111
5.9 进程的终止……112
5.10 分页器接口……112
5.10.1 vnode 分页器……114
5.10.2 设备分页器……115
5.10.3 交换分页器……115
5.11 分页……117
5.12 页面替换……121
5.12.1 换页参数……122
5.12.2 页面换出守护进程……123
5.12.3 交换……124
5.12.4 换入进程……125
5.13 可移植性……126
5.13.1 pmap 模块的角色……128
5.13.2 初始化和启动……130
5.13.3 映射的分配和释放……132
5.13.4 改变映射的访问和锁定属性……134
5.13.5 页表使用信息的管理……135
5.13.6 物理页面的初始化……135
5.13.7 内部数据结构的管理……136
习题……137
参考资源……137

第三部分 I/O 系统

第 6 章 I/O 系统综述……139
6.1 从用户到设备的 I/O 映射……139
6.1.1 设备驱动程序……140
6.1.2 I/O 队列……141
6.1.3 中断处理……141
6.2 块设备……142
6.2.1 块设备的入口点……142
6.2.2 磁盘 I/O 请求的排序……143
6.2.3 磁盘标签……144
6.3 字符设备……145
6.3.1 原始设备和物理 I/O……146
6.3.2 面向字符的设备……147
6.3.3 字符设备驱动程序的入口点……148
6.4 描述符管理和服务……148
6.4.1 打开文件入口……149
6.4.2 对描述符的管理……151
6.4.3 文件描述符的锁定……152
6.4.4 描述符上的多路复用 I/O 操作……154
6.4.5 select 的实现……155
6.4.6 在内核中数据的移动……157
6.5 虚拟文件系统的接口……159
6.5.1 vnode 的内容……159
6.5.2 对 vnode 的操作……160
6.5.3 路径名翻译……161
6.5.4 导出文件系统服务……162
6.6 独立于文件系统的服务……163
6.6.1 名字缓存……164
6.6.2 缓冲区管理……165
6.6.3 缓冲区管理的实现……167
6.7 可堆叠的文件系统……169
6.7.1 简单文件系统层……170
6.7.2 联合安装的文件系统……171
6.7.3 其他的文件系统……173
习题……174
参考资源……175
第 7 章 本地文件系统……176
7.1 文件系统的分层管理……176
7.2 索引节点的结构……177
7.3 命名……180

7.3.1 目录……180
7.3.2 在目录中查找名字……181
7.3.3 路径名转化……182
7.3.4 链接……182
7.4 配额……185
7.5 文件锁定……187
7.6 其他文件系统机制……191
7.6.1 大文件支持……191
7.6.2 文件标志……192
习题……193
参考资源……193
第 8 章 本地文件库……194
8.1 文件库概述……194
8.2 Berkeley 快速文件系统……196
8.2.1 Berkeley 快速文件系统的组织……197
8.2.2 存储策略的优化……198
8.2.3 文件的读 / 写操作……199
8.2.4 文件系统参数化……201
8.2.5 布局策略……202
8.2.6 分配机制……203
8.2.7 块的成簇……205
8.2.8 同步操作……207
8.3 日志结构的文件系统……208
8.3.1 日志结构的文件系统组织……209
8.3.2 索引文件……211
8.3.3 读日志……212
8.3.4 写日志……212
8.3.5 块统计……213
8.3.6 缓存……215
8.3.7 目录操作……215
8.3.8 文件的创建……216
8.3.9 读写文件……217
8.3.10 文件系统清理……217
8.3.11 文件系统参数化……219
8.3.12 文件系统的崩溃恢复……219
8.4 基于内存的文件系统……220
8.4.1 基于内存的文件系统的组织……221
8.4.2 文件系统性能……223
8.4.3 展望……223
习题……224
参考资源……225
第 9 章 网络文件系统……227
9.1 历史和概况……227
9.2 NFS 结构和操作……229
9.2.1 NFS 协议……231
9.2.2 4.4BSD 的 NFS 实现……233
9.2.3 客户端 / 服务器的交互……236
9.2.4 RPC 的传输问题……236
9.2.5 安全问题……237
9.3 提高性能的技术……239
9.3.1 租约……241
9.3.2 崩溃恢复……244
习题……245
参考资源……246
第 10 章 终端处理……248
10.1 终端处理模式……248
10.2 行规程……249
10.3 用户接口……250
10.4 tty 数据结构……251
10.5 进程组、会话和终端控制……253
10.6 字符列表……253
10.7 RS-232 和调制解调器控制……254
10.8 终端操作……255
10.8.1 打开……255
10.8.2 输出行规程……256
10.8.3 输出的上半部……257
10.8.4 输出的下半部……258
10.8.5 输入的下半部……258
10.8.6 输入的上半部……259
10.8.7 stop 例程……260
10.8.8 ioctl 例程……260
10.8.9 调制解调器转换……261
10.8.10 关闭终端设备……261
10.9 其他的行规程……262

10.9.1 串行线路 IP 规程 ……262
10.9.2 图表行规程 ……263
习题 ……263
参考资源 ……263

第四部分 通信

第 11 章 进程间通信 ……265
11.1 进程间通信模型 ……265
11.2 实现结构和概述 ……270
11.3 内存管理 ……271
11.3.1 mbufs ……271
11.3.2 存储管理算法 ……273
11.3.3 mbuf 操作例程 ……274
11.4 数据结构 ……275
11.4.1 通信域 ……275
11.4.2 套接字 ……276
11.4.3 套接字地址 ……278
11.5 建立连接 ……279
11.6 数据传送 ……281
11.6.1 传送数据 ……281
11.6.2 接收数据 ……283
11.6.3 传递访问权限 ……285
11.6.4 在本地域传递访问权限 ……286
11.7 关闭套接字 ……287
习题 ……288
参考资源 ……289
第 12 章 网络通信 ……290
12.1 内部结构 ……290
12.1.1 数据流 ……291
12.1.2 通信协议 ……291
12.1.3 网络接口 ……293
12.2 套接字到协议的接口 ……297
12.2.1 协议用户请求例程 ……298
12.2.2 内部请求 ……300
12.2.3 协议控制 - 输出例程 ……300
12.3 协议到协议的接口 ……301
12.3.1 pr_output ……301
12.3.2 pr_input ……301
12.3.3 pr_ctlinput ……301
12.4 协议和网络接口之间的接口 ……302
12.4.1 数据包的传送 ……302
12.4.2 数据包的接收 ……303
12.5 路由 ……305
12.5.1 内核路由表 ……306
12.5.2 路由查找 ……308
12.5.3 路由重定向 ……311
12.5.4 路由表接口 ……311
12.5.5 用户级的路由策略 ……312
12.5.6 用户级路由接口：路由套接字 ……312
12.6 缓存和拥塞控制 ……313
12.6.1 协议缓存策略 ……313
12.6.2 队列限制 ……314
12.7 原始套接字 ……314
12.7.1 控制块 ……314
12.7.2 输入处理 ……315
12.7.3 输出处理 ……315
12.8 其他的网络子系统主题 ……315
12.8.1 带外数据 ……316
12.8.2 地址解析协议 ……316
习题 ……317
参考资源 ……318
第 13 章 网络协议 ……320
13.1 Internet 网络协议 ……320
13.1.1 Internet 地址 ……322
13.1.2 子网 ……322
13.1.3 广播地址 ……324
13.1.4 Internet 多播 ……324
13.1.5 Internet 端口与关联 ……325
13.1.6 协议控制块 ……325
13.2 用户数据报协议（UDP） ……325
13.2.1 初始化 ……326
13.2.2 输出 ……327
13.2.3 输入 ……327
13.2.4 控制操作 ……328

13.3 互联网协议（IP）……328
13.3.1 输出……329
13.3.2 输入……330
13.3.3 转发……331
13.4 传输控制协议（TCP）……332
13.4.1 TCP 连接状态……333
13.4.2 序列变量……336
13.5 TCP 算法……337
13.5.1 定时器……338
13.5.2 往返程时间的估计……339
13.5.3 连接建立……340
13.5.4 连接关闭……341
13.6 TCP 输入处理……342
13.7 TCP 输出处理……344
13.7.1 数据的发送……345
13.7.2 避免傻瓜窗口综合征……346
13.7.3 避免小数据包……346
13.7.4 延迟的确认及窗口更新……347
13.7.5 重发状态……348
13.7.6 慢启动……348
13.7.7 源抑制的处理……349
13.7.8 缓冲区与窗口大小调整……349
13.7.9 使用慢启动避免拥塞……350
13.7.10 快速重发……351
13.8 Internet 控制报文协议（ICMP）……352
13.9 OSI 实现中的问题……353
13.10 联网和进程间通信综述……355
13.10.1 通信通道的创建……355
13.10.2 数据的发送与接收……356
13.10.3 数据传送或接收的终止……356
习题……357
参考资源……359

第五部分 系统操作

第 14 章 系统启动……361
14.1 概述……361
14.2 引导……362
14.3 内核的初始化……363
14.3.1 汇编语言启动……363
14.3.2 机器相关初始化……364
14.3.3 消息缓冲区……364
14.3.4 系统数据结构……364
14.4 自动配置……365
14.4.1 设备的探测……366
14.4.2 设备连接……367
14.4.3 新的自动配置数据结构……367
14.4.4 新的自动配置函数……368
14.4.5 设备命名……368
14.5 独立于机器的初始化……369
14.6 用户级初始化……371
14.6.1 /sbin/init……371
14.6.2 /etc/rc……372
14.6.3 /usr/libexe/getty……372
14.6.4 /usr/bin/login……372
14.7 系统启动的相关话题……373
14.7.1 内核的配置……373
14.7.2 系统关机与自动重启……373
14.7.3 系统调试……374
14.7.4 同内核来回传递信息……374
习题……375
参考资源……376
附录 术语表……377

第一部分
综　述

第①章
BSD 系统的历史和目标

1.1　UNIX 系统的历史

20 多年来，UNIX 操作系统被广泛应用在计算机界的各个领域。不计其数的公司和团体为 UNIX 系统的发展和完善而努力，从而使 UNIX 系统涌现出许多不同的版本。而本书主要集中精力对 UNIX 系统发展历史中的一个分支——BSD 系统进行讨论。UNIX 在这个分支上的发展主要由以下几个事件组成：

- 贝尔实验室（Bell Laboratories）创造出 UNIX。
- 加利福尼亚大学伯克利分校（University of California at Berkeley）的计算机系统研究小组（Computer Systems Research Group，CSRG）为 UNIX 系统实现虚拟内存机制和 TCP/IP 协议体系。
- 伯克利软件设计公司（Berkeley Software Design，Incorporated，BSDI）在 CSRG 研究的基础上，提出了 FreeBSD 和 NetBSD 的计划并进行开发。

1.1.1　UNIX 系统的起源

第一个 UNIX 系统的诞生是在 1969 年，贝尔实验室的 Ken Thompson 在一台已经过时的 PDP-7 上作为一个私人研发项目开始开发这个系统。很快，Dennis Ritchie，这位 C 语言的创始人加入了进来，用 C 语言重写整个系统，几乎所有的汇编程序都被 C 程序所代替。由于 UNIX 系统初始设计得简单、精练以及随后 15 年内的不断完善和发展［Ritvhie 1984a；Compton，1985］，该系统成为计算机界最重要和功能最完备的操作系统之一［Ritvhie，1987］。

Ritchie、Thompson 和早期的一些 UNIX 开发者曾经参与贝尔实验室的一个叫做 Multics 的研究项目［Peirce，1985；Organick，1975］，而该项目对新一代操作系统的发展起着不可磨灭的作用。UNIX 从某种角度也可以看做是一个 Multics：Multics 关注功能的齐全性，而 UNIX 致力于各项功能的完善。UNIX 中的很多设计方案和系统特征，都直接源于 Multics。例如文件系统的基本框架，分配一个用户进程作为命令行解释器的思想，文件系统界面的基本结构等。

许多其他操作系统（像麻省理工学院的 CTSS）的思想，也都融入了 UNIX 中。例如新建进程的 fork 操作，它源于伯克利分校的 GENIE（SDS-940，最新的 XDS-940）系统。用户可以轻易地创建进程，这样每条语句都使用一个进程，而不是被当做一个过程调用，这一想法是源于 Multics。

在 UNIX 系统的历史中至少有三条主线。图 1-1 描述了它们早期的形成过程，而图 1-2 则介

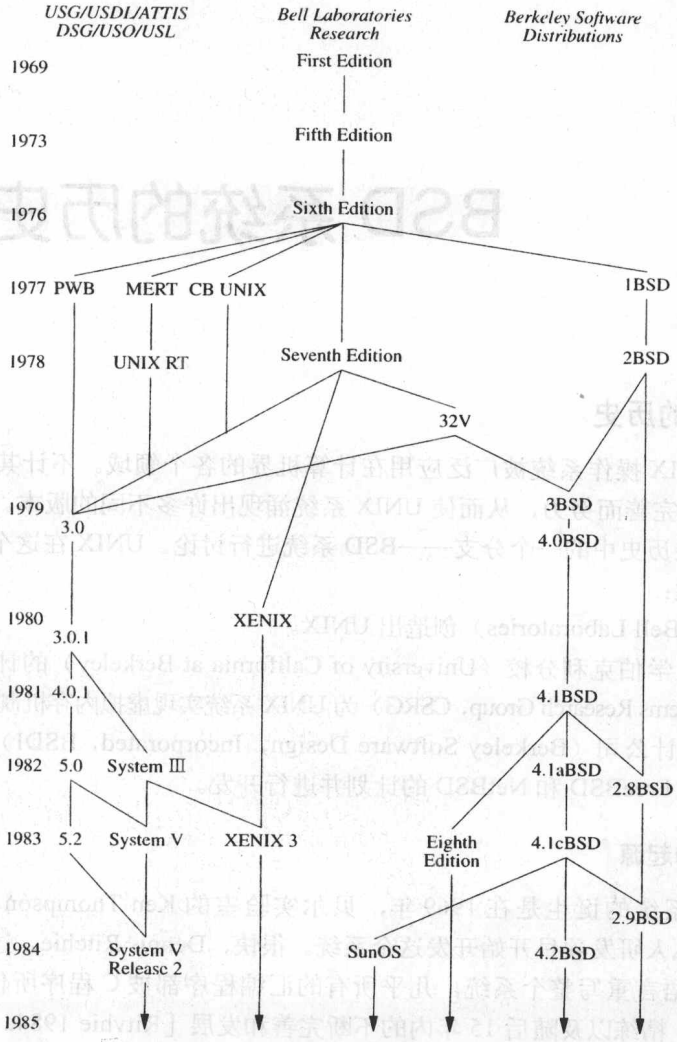


图 1-1　UNIX 系统树，1969 ～ 1985

绍了它们在近些年来的发展情况，特别是对那些后来发展为 4.4BSD 和 UNIX 系统 V［Chambers & Quarterman，1983；Uniejew Ski，1985］的分支给予了详细的介绍。在这两幅图中的时间只是一个大概时间，我们也不需要知道一个非常确切的发展过程。图 1-2 中一些系统的名字在下文中并没有提到。它们被收录这里只是为了让读者更清楚其与我们所要具体分析的 BSD 系统之间的关系。

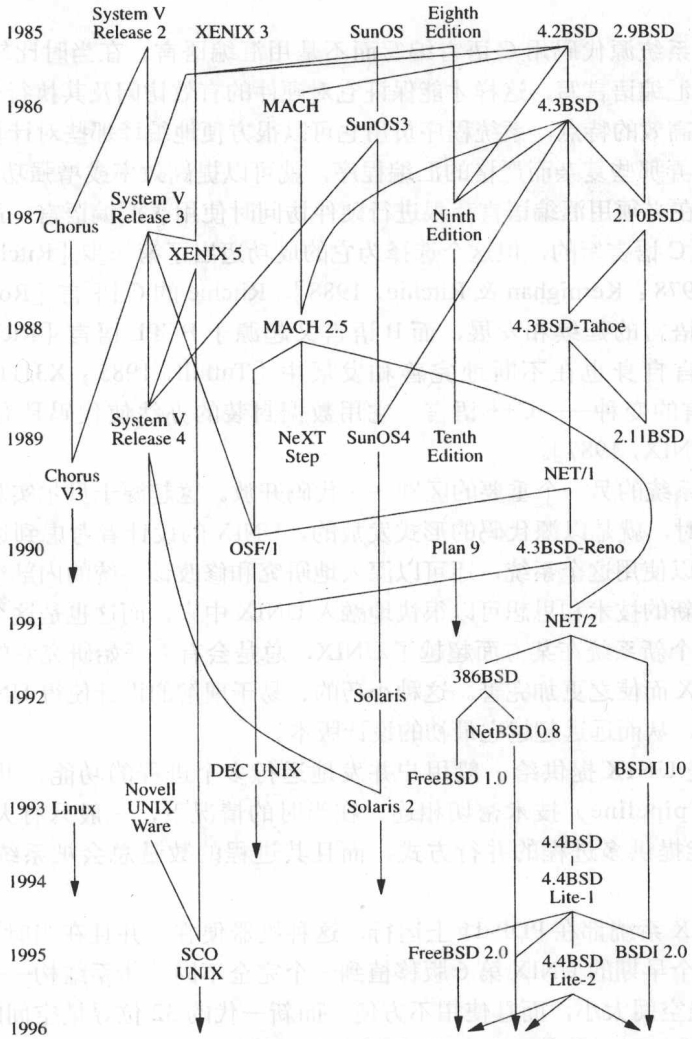


图 1-2　UNIX 系统树，1986 ～ 1996

1.1.2　UNIX 系统的研究与发展

UNIX 第一个重要的成型版本是在贝尔实验室成功开发的。它在该系统的早期版本上新增了分时系统（UNIX Time-Sharing System）第 6 版，即我们所熟知的 V6。1976 年，这种技术的第一个版本在贝尔实验室外被广泛使用。系统是以《UNIX 程序员手册》（*UNIX Programmer's Manual*）发布时的版本号来标识的。