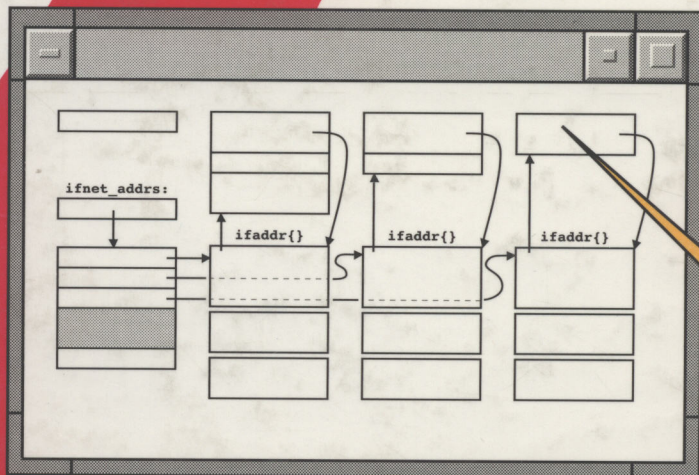


TCP/IP Illustrated, Volume 2

The Implementation

Gary R. Wright
W. Richard Stevens



ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES

TCP/IP Illustrated, Volume 2

The Implementation

Gary R. Wright
W. Richard Stevens

江苏工业学院图书馆
藏书章



ADDISON-WESLEY PUBLISHING COMPANY

Reading, Massachusetts Menlo Park, California New York

Don Mills, Ontario Wokingham, England Amsterdam

Bonn Sydney Singapore Tokyo Madrid San Juan

Seoul Milan Mexico City Taipei

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial caps or all caps.

The programs and applications presented in this book have been included for their instructional value. They have been tested with care, but are not guaranteed for any particular purpose. The publisher does not offer any warranties or representations, nor does it accept any liabilities with respect to the programs or applications.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate & Professional Publishing Group
Addison-Wesley Publishing Company
One Jacob Way
Reading, Massachusetts 01867

Library of Congress Cataloging-in-Publication Data

(Revised for vol. 2)

Stevens, W. Richard.
TCP/IP illustrated.

(Addison-Wesley professional computing series)

Vol. 2 by Gary R. Wright, W. Richard Stevens.

Includes bibliographical references and indexes.

Contents: v. 1. The protocols — v. 2. The implementation

1. TCP/IP (Computer network protocol) I. Wright,
Gary R. II. Title. III. Series.

TK5105.55.S74 1994 004.6'2 93-40000

ISBN 0-201-63346-9 (v. 1)

ISBN 0-201-63354-X (v. 2)

The BSD Daemon used on the cover of this book is reproduced with the permission of Marshall Kirk McKusick.

Copyright © 1995 by Addison-Wesley Publishing Company, Inc.

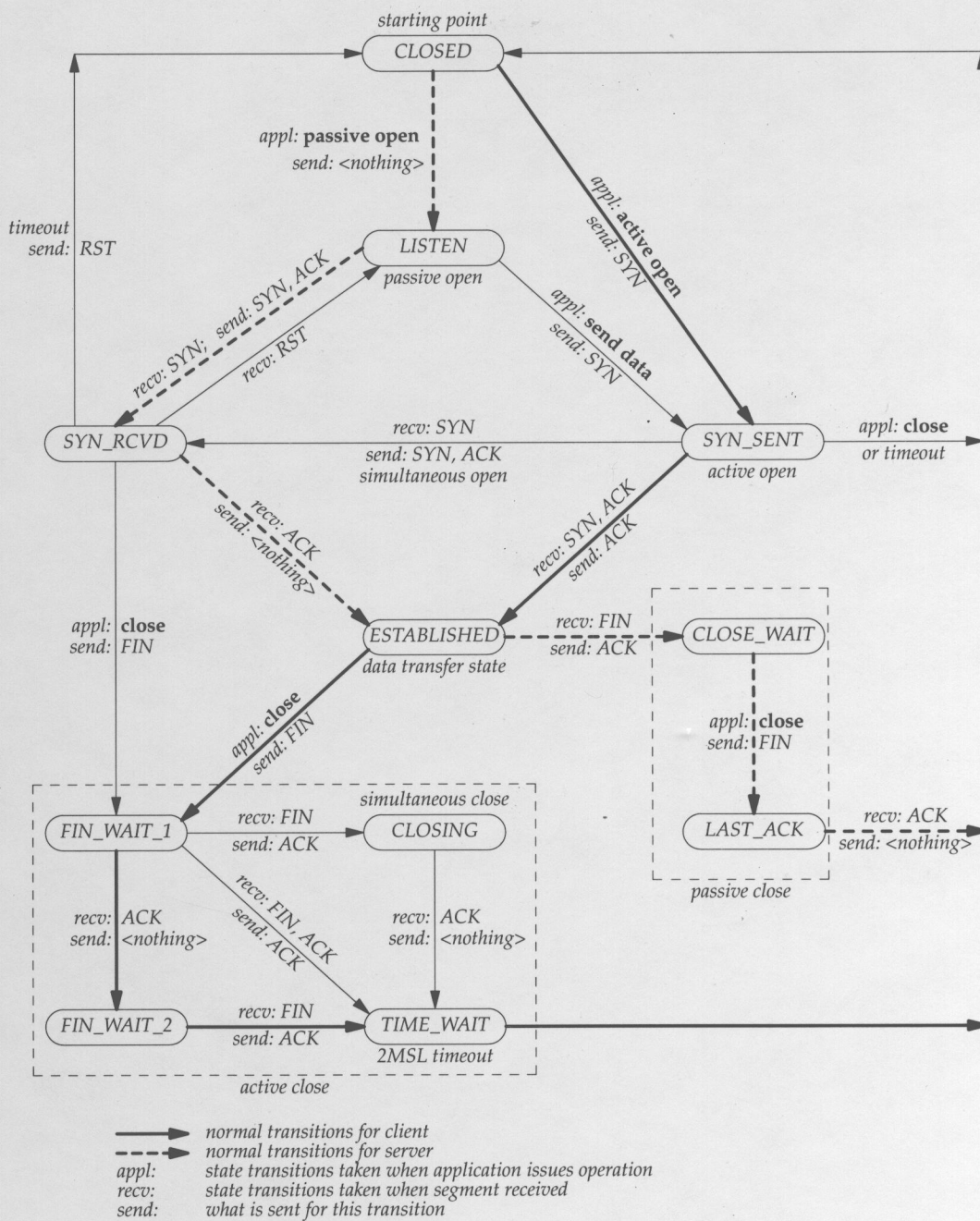
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form, or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior consent of the publisher. Printed in the United States of America. Published simultaneously in Canada.

Text printed on recycled and acid-free paper

ISBN 0-201-63354-X

3 4 5 6 7 8 9 10 11-CRW-99989796

Third printing, March 1996



TCP state transition diagram.

Structure Definitions

arpcom	80	mrt	419
arphdr	682	mrtctl	420
		msghdr	482
bpf_d	1033		
bpf_hdr	1029	osockaddr	75
bpf_if	1029		
		pdevinit	78
cmsghdr	482	protosw	188
domain	187	radix_mask	578
		radix_node	575
ether_arp	682	radix_node_head	574
ether_header	102	rawcb	647
ether_multi	342	route	220
		route_cb	625
icmp	308	rt_addrinfo	623
ifaddr	73	rtentry	579
ifa_msghdr	622	rt_metrics	580
ifconf	117	rt_msghdr	622
if_msghdr	622		
ifnet	67	selinfo	531
ifqueue	71	sl_softc	83
ifreq	117	sockaddr	75
igmp	384	sockaddr_dl	87
in_addr	160	sockaddr_in	160
in_aliasreq	174	sockaddr_inarp	701
in_ifaddr	161	sockbuf	476
in_multi	345	socket	438
inpcb	716	socket_args	444
iovec	481	sockproto	626
ip	211	sysent	443
ipasfrag	287		
ip_moptions	347	tcpcb	804
ip_mreq	356	tcp_debug	916
ipoption	265	tcphdr	801
ipovly	760	tcpihdr	803
ipq	286	timeval	106
ip_srcrt	258		
ip_timestamp	262	udphdr	759
		udpihdr	759
		uio	485
le_softc	80		
lgrplctl	411		
linger	542	vif	406
llinfo_arp	682	vifctl	407
mbuf	38	walkarg	632

Praise for

TCP/IP Illustrated, Volume 1: The Protocols

"*TCP/IP Illustrated* has already become my most-likely-to-have-the-answer reference book, the first resource I turn to with a networking question. The book is, all publisher hype aside, an instant classic, and I, for one, am thrilled that something like this is now available."

— Vern Paxson, *login:*, March/April 1994

"This is sure to be the bible for TCP/IP developers and users."

— Robert A. Ciampa, Network Engineer, Synernetics, division of 3COM

"... the difference is that Stevens wants to show as well as tell about the protocols. His principal teaching tools are straight-forward explanations, exercises at the ends of chapters, byte-by-byte diagrams of headers and the like, and listings of actual traffic as examples."

— Walter Zintz, *Unix World*, December 1993

"*TCP/IP Illustrated, Volume 1* is based on practical examples that reinforce the theory — distinguishing this book from others on the subject, and making it both readable and informative."

— Peter M. Haverlock, Consultant, IBM TCP/IP Development

"While all of Stevens' books are excellent, this new opus is awesome. Although many books describe the TCP/IP protocols, the author provides a level of depth and real-world detail lacking from the competition."

— Steven Baker, *Unix Review*, March 1994

"*TCP/IP Illustrated, Volume 1* is an excellent reference for developers, network administrators or anyone who needs to understand TCP/IP technology."

— Bob Williams, V.P. Marketing, NetManage, Inc.

"W. Richard Stevens has produced a fine text and reference work."

— Scott Bradner, Consultant, Harvard University OIT/NSD

"Even marketing weenies (of a technical bent) will appreciate this book, as it is clearly written, and uses lots of diagrams. I especially like the author's thoughtful use of asides—set in smaller type and indented—to explain this or that concept."

— Ron Jeffries, *ATM USER*, January 1994

"Stevens takes a subject that has been written about rather prolifically, TCP/IP, and does something fresh and useful with it."

— Jason Levitt, *Open Systems Today*, March 7, 1994

More Praise for

TCP/IP Illustrated, Volume 1: The Protocols

"This book is a stone jewel. ... Written by W. Richard Stevens, this book probably provides the most comprehensive view of TCP/IP available today in print."

— *Boardwatch*, April/May 1994

"...you can't get a better understanding of the workings of TCP/IP anywhere."

— Tom Nolle, *Netwatcher*, January 1994

"The book covers all the basic TCP/IP applications, including Telnet, NFS (Network File System), FTP (file transfer protocol) and TFTP (trivial FTP)."

— *Data Communications*, January 21, 1994

"The diagrams he uses are excellent and his writing style is clear and readable. Please read it and keep it on your bookshelf."

— Elizabeth Zinkann, *Sys Admin*, November 1993

"Stevens' Unix-oriented investigations will be invaluable to the network programmer or specialist who wishes to really understand how the TCP/IP stack is put together."

— Joel Snyder, *Internet World*, March/April 1994 issue

"All aspects of the transmission control protocol/Internet protocol (TCP/IP) are covered here, from link layer and static/dynamic routing implementations to applications such as SNMP and Telnet."

— *Telecommunications*, March 1994

"The author of *TCP/IP Illustrated* has succeeded in creating another indispensable tome of networking knowledge. This is the most comprehensible and complete book I have read on TCP/IP. It takes a different slant than other books, by presenting not only details of TCP, IP, ARP, ICMP, routing, etc., but actually shows these protocols (and common Internet tools) in action."

— Eli Charne, *ConneXions*, July 1994

"The word 'illustrated' distinguishes this book from its many rivals."

— Stan Kelly-Bootle, *Unix Review*, December 1993

TP393
w 6.2

TCP/IP Illustrated, Volume 2

Addison-Wesley Professional Computing Series

Brian W. Kernighan, Consulting Editor

Ken Arnold/John Peyton, *A C User's Guide to ANSI C*

Tom Cargill, *C++ Programming Style*

William R. Cheswick/Steven M. Bellovin, *Firewalls and Internet Security: Repelling the Wily Hacker*

David A. Curry, *UNIX® System Security: A Guide for Users and System Administrators*

Erich Gamma/Richard Helm/Ralph Johnson/John Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*

John Lakos, *Large-Scale C++ Software Design*

Scott Meyers, *Effective C++: 50 Specific Ways to Improve Your Programs and Designs*

Scott Meyers, *More Effective C++: 35 New Ways to Improve Your Programs and Designs*

Robert B. Murray, *C++ Strategies and Tactics*

David R. Musser/Atul Saini, *STL Tutorial and Reference Guide: C++ Programming with the Standard Template Library*

John K. Ousterhout, *Tcl and the Tk Toolkit*

Craig Partridge, *Gigabit Networking*

J. Stephen Pendergrast Jr., *Desktop KornShell Graphical Programming*

Radia Perlman, *Interconnections: Bridges and Routers*

David M. Piscitello/A. Lyman Chapin, *Open Systems Networking: TCP/IP and OSI*

Stephen A. Rago, *UNIX® System V Network Programming*

Curt Schimmel, *UNIX® Systems for Modern Architectures: Symmetric Multiprocessing and Caching for Kernel Programmers*

W. Richard Stevens, *Advanced Programming in the UNIX® Environment*

W. Richard Stevens, *TCP/IP Illustrated, Volume 1: The Protocols*

W. Richard Stevens, *TCP/IP Illustrated, Volume 3: TCP for Transactions, HTTP, NNTP, and the UNIX Domain Protocols*

Gary R. Wright/W. Richard Stevens, *TCP/IP Illustrated, Volume 2: The Implementation*

*To my parents and my sister,
for their love and support.
—G.R.W.*

Contents

*To my parents,
for the gift of an education,
and the example of a work ethic.
—W.R.S.*

Page

Chapter 1 Introduction 1

Chapter 2 The Problem 2

Chapter 3 The Literature 3

Chapter 4 The Method 4

Chapter 5 The Results 5

Chapter 6 The Discussion 6

Chapter 7 The Conclusion 7

Chapter 8 The Appendix 8

Chapter 9 The Bibliography 9

Chapter 10 The Index 10

Chapter 11 The Glossary 11

Chapter 12 The Acknowledgments 12

Chapter 13 The Dedication 13

Chapter 14 The Foreword 14

Chapter 15 The Preface 15

Chapter 16 The Epilogue 16

Chapter 17 The Afterword 17

Chapter 18 The Appendix 18

Chapter 19 The Bibliography 19

Chapter 20 The Index 20

Chapter 21 The Glossary 21

Chapter 22 The Acknowledgments 22

Preface

Introduction

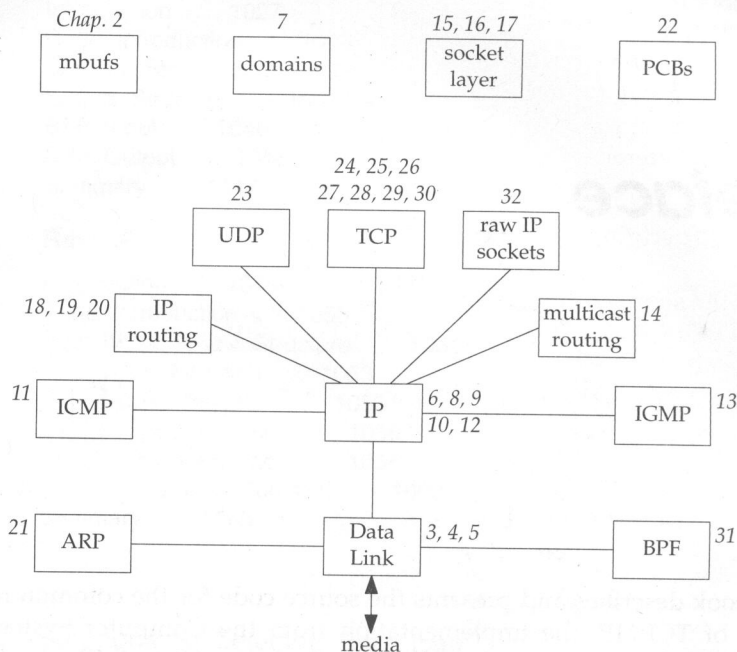
This book describes and presents the source code for the common reference implementation of TCP/IP: the implementation from the Computer Systems Research Group (CSRG) at the University of California at Berkeley. Historically this has been distributed with the 4.x BSD system (Berkeley Software Distribution). This implementation was first released in 1982 and has survived many significant changes, much fine tuning, and numerous ports to other Unix and non-Unix systems. This is not a toy implementation, but the foundation for TCP/IP implementations that are run daily on hundreds of thousands of systems worldwide. This implementation also provides router functionality, letting us show the differences between a host implementation of TCP/IP and a router.

We describe the implementation and present the entire source code for the kernel implementation of TCP/IP, approximately 15,000 lines of C code. The version of the Berkeley code described in this text is the 4.4BSD-Lite release. This code was made publicly available in April 1994, and it contains numerous networking enhancements that were added to the 4.3BSD Tahoe release in 1988, the 4.3BSD Reno release in 1990, and the 4.4BSD release in 1993. (Appendix B describes how to obtain this source code.) The 4.4BSD release provides the latest TCP/IP features, such as multicasting and long fat pipe support (for high-bandwidth, long-delay paths). Figure 1.1 (p. 4) provides additional details of the various releases of the Berkeley networking code.

This book is intended for anyone wishing to understand how the TCP/IP protocols are implemented: programmers writing network applications, system administrators responsible for maintaining computer systems and networks utilizing TCP/IP, and any programmer interested in understanding how a large body of nontrivial code fits into a real operating system.

Organization of the Book

The following figure shows the various protocols and subsystems that are covered. The *italic numbers* by each box indicate the chapters in which that topic is described.



We take a bottom-up approach to the TCP/IP protocol suite, starting at the data-link layer, then the network layer (IP, ICMP, IGMP, IP routing, and multicast routing), followed by the socket layer, and finishing with the transport layer (UDP, TCP, and raw IP).

Intended Audience

This book assumes a basic understanding of how the TCP/IP protocols work. Readers unfamiliar with TCP/IP should consult the first volume in this series, [Stevens 1994], for a thorough description of the TCP/IP protocol suite. This earlier volume is referred to throughout the current text as *Volume 1*. The current text also assumes a basic understanding of operating system principles.

We describe the implementation of the protocols using a data-structures approach. That is, in addition to the source code presentation, each chapter contains pictures and descriptions of the data structures used and maintained by the source code. We show how these data structures fit into the other data structures used by TCP/IP and the kernel. Heavy use is made of diagrams throughout the text—there are over 250 diagrams.

This data-structures approach allows readers to use the book in various ways. Those interested in all the implementation details can read the entire text from start to finish, following through all the source code. Others might want to understand how the

protocols are implemented by understanding all the data structures and reading all the text, but not following through all the source code.

We anticipate that many readers are interested in specific portions of the book and will want to go directly to those chapters. Therefore many forward and backward references are provided throughout the text, along with a thorough index, to allow individual chapters to be studied by themselves. The inside back covers contain an alphabetical cross-reference of all the functions and macros described in the book and the starting page number of the description. Exercises are provided at the end of the chapters; most solutions are in Appendix A to maximize the usefulness of the text as a self-study reference.

Source Code Copyright

All of the source code presented in this book, other than Figures 1.2 and 8.27, is from the 4.4BSD-Lite distribution. This software is publicly available through many sources (Appendix B).

All of this source code contains the following copyright notice.

```
/*
 * Copyright (c) 1982, 1986, 1988, 1990, 1993, 1994
 *   The Regents of the University of California. All rights reserved.
 *
 * Redistribution and use in source and binary forms, with or without
 * modification, are permitted provided that the following conditions
 * are met:
 * 1. Redistributions of source code must retain the above copyright
 *   notice, this list of conditions and the following disclaimer.
 * 2. Redistributions in binary form must reproduce the above copyright
 *   notice, this list of conditions and the following disclaimer in the
 *   documentation and/or other materials provided with the distribution.
 * 3. All advertising materials mentioning features or use of this software
 *   must display the following acknowledgement:
 *   This product includes software developed by the University of
 *   California, Berkeley and its contributors.
 * 4. Neither the name of the University nor the names of its contributors
 *   may be used to endorse or promote products derived from this software
 *   without specific prior written permission.
 *
 * THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS ``AS IS'' AND
 * ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
 * IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE
 * ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE
 * FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL
 * DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS
 * OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)
 * HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT
 * LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY
 * OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF
 * SUCH DAMAGE.
 */
```

Acknowledgments

We thank the technical reviewers who read the manuscript and provided important feedback on a tight timetable: Ragnvald Blindheim, Jon Crowcroft, Sally Floyd, Glen Glater, John Gulbenkian, Don Hering, Mukesh Kacker, Berry Kercheval, Brian W. Kernighan, Ulf Kieber, Mark Laubach, Steven McCanne, Craig Partridge, Vern Paxson, Steve Rago, Chakravardhi Ravi, Peter Salus, Doug Schmidt, Keith Sklower, Ian Lance Taylor, and G. N. Ananda Vardhana. A special thanks to the consulting editor, Brian Kernighan, for his rapid, thorough, and helpful reviews throughout the course of the project, and for his continued encouragement and support.

Our thanks (again) to the National Optical Astronomy Observatories (NOAO), especially Sidney Wolff, Richard Wolff, and Steve Grandi, for providing access to their networks and hosts. Our thanks also to the U.C. Berkeley CSRG: Keith Bostic and Kirk McKusick provided access to the latest 4.4BSD system, and Keith Sklower provided the modifications to the 4.4BSD-Lite software to run under BSD/386 V1.1.

G.R.W. wishes to thank John Wait, for several years of gentle prodding; Dave Schaller, for his encouragement; and Jim Hogue, for his support during the writing and production of this book.

W.R.S. thanks his family, once again, for enduring another “small” book project. Thank you Sally, Bill, Ellen, and David.

The hardwork, professionalism, and support of the team at Addison-Wesley has made the authors’ job that much easier. In particular, we wish to thank John Wait for his guidance and Kim Dawley for her creative ideas.

Camera-ready copy of the book was produced by the authors. It is only fitting that a book describing an industrial-strength software system be produced with an industrial-strength text processing system. Therefore one of the authors chose to use the Groff package written by James Clark, and the other author agreed begrudgingly.

We welcome electronic mail from any readers with comments, suggestions, or bug fixes: tcpi2-book@aw.com. Each author will gladly blame the other for any remaining errors.

Gary R. Wright

<http://www.connix.com/~gwright>
Middletown, Connecticut

November 1994

W. Richard Stevens

<http://www.noao.edu/~rstevens>
Tucson, Arizona

Contents

Preface

xix

Chapter 1. Introduction

1

- 1.1 Introduction 1
- 1.2 Source Code Presentation 1
- 1.3 History 3
- 1.4 Application Programming Interfaces 5
- 1.5 Example Program 5
- 1.6 System Calls and Library Functions 7
- 1.7 Network Implementation Overview 9
- 1.8 Descriptors 10
- 1.9 Mbufs (Memory Buffers) and Output Processing 15
- 1.10 Input Processing 19
- 1.11 Network Implementation Overview Revisited 22
- 1.12 Interrupt Levels and Concurrency 23
- 1.13 Source Code Organization 26
- 1.14 Test Network 28
- 1.15 Summary 29

Chapter 2. Mbufs: Memory Buffers

31

- 2.1 Introduction 31
- 2.2 Code Introduction 36
- 2.3 Mbuf Definitions 37
- 2.4 mbuf Structure 38
- 2.5 Simple Mbuf Macros and Functions 40
- 2.6 m_devget and m_pullup Functions 44

2.7	Summary of Mbuf Macros and Functions	51	
2.8	Summary of Net/3 Networking Data Structures		54
2.9	m_copy and Cluster Reference Counts	56	
2.10	Alternatives	60	
2.11	Summary	60	
Chapter 3.	Interface Layer		63
3.1	Introduction	63	
3.2	Code Introduction	64	
3.3	ifnet Structure	65	
3.4	ifaddr Structure	73	
3.5	sockaddr Structure	74	
3.6	ifnet and ifaddr Specialization	76	
3.7	Network Initialization Overview	77	
3.8	Ethernet Initialization	80	
3.9	SLIP Initialization	82	
3.10	Loopback Initialization	85	
3.11	if_attach Function	85	
3.12	ifinit Function	93	
3.13	Summary	94	
Chapter 4.	Interfaces: Ethernet		95
4.1	Introduction	95	
4.2	Code Introduction	96	
4.3	Ethernet Interface	98	
4.4	ioctl System Call	114	
4.5	Summary	125	
Chapter 5.	Interfaces: SLIP and Loopback		127
5.1	Introduction	127	
5.2	Code Introduction	127	
5.3	SLIP Interface	128	
5.4	Loopback Interface	150	
5.5	Summary	153	
Chapter 6.	IP Addressing		155
6.1	Introduction	155	
6.2	Code Introduction	158	
6.3	Interface and Address Summary	158	
6.4	sockaddr_in Structure	160	
6.5	in_ifaddr Structure	161	
6.6	Address Assignment	161	
6.7	Interface ioctl Processing	177	
6.8	Internet Utility Functions	181	
6.9	ifnet Utility Functions	182	
6.10	Summary	183	

Chapter 7.	Domains and Protocols	185
7.1	Introduction	185
7.2	Code Introduction	186
7.3	domain Structure	187
7.4	protosw Structure	188
7.5	IP domain and protosw Structures	191
7.6	pffindproto and pffindtype Functions	196
7.7	pfctlinput Function	198
7.8	IP Initialization	199
7.9	sysctl System Call	201
7.10	Summary	204
Chapter 8.	IP: Internet Protocol	205
8.1	Introduction	205
8.2	Code Introduction	206
8.3	IP Packets	210
8.4	Input Processing: ipintr Function	212
8.5	Forwarding: ip_forward Function	220
8.6	Output Processing: ip_output Function	228
8.7	Internet Checksum: in_cksum Function	234
8.8	setsockopt and getsockopt System Calls	239
8.9	ip_sysctl Function	244
8.10	Summary	245
Chapter 9.	IP Option Processing	247
9.1	Introduction	247
9.2	Code Introduction	247
9.3	Option Format	248
9.4	ip_dooptions Function	249
9.5	Record Route Option	252
9.6	Source and Record Route Options	254
9.7	Timestamp Option	261
9.8	ip_insertoptions Function	265
9.9	ip_pcbopts Function	269
9.10	Limitations	272
9.11	Summary	272
Chapter 10.	IP Fragmentation and Reassembly	275
10.1	Introduction	275
10.2	Code Introduction	277
10.3	Fragmentation	278
10.4	ip_optcopy Function	282
10.5	Reassembly	283
10.6	ip_reass Function	286
10.7	ip_slowtimo Function	298
10.8	Summary	300