

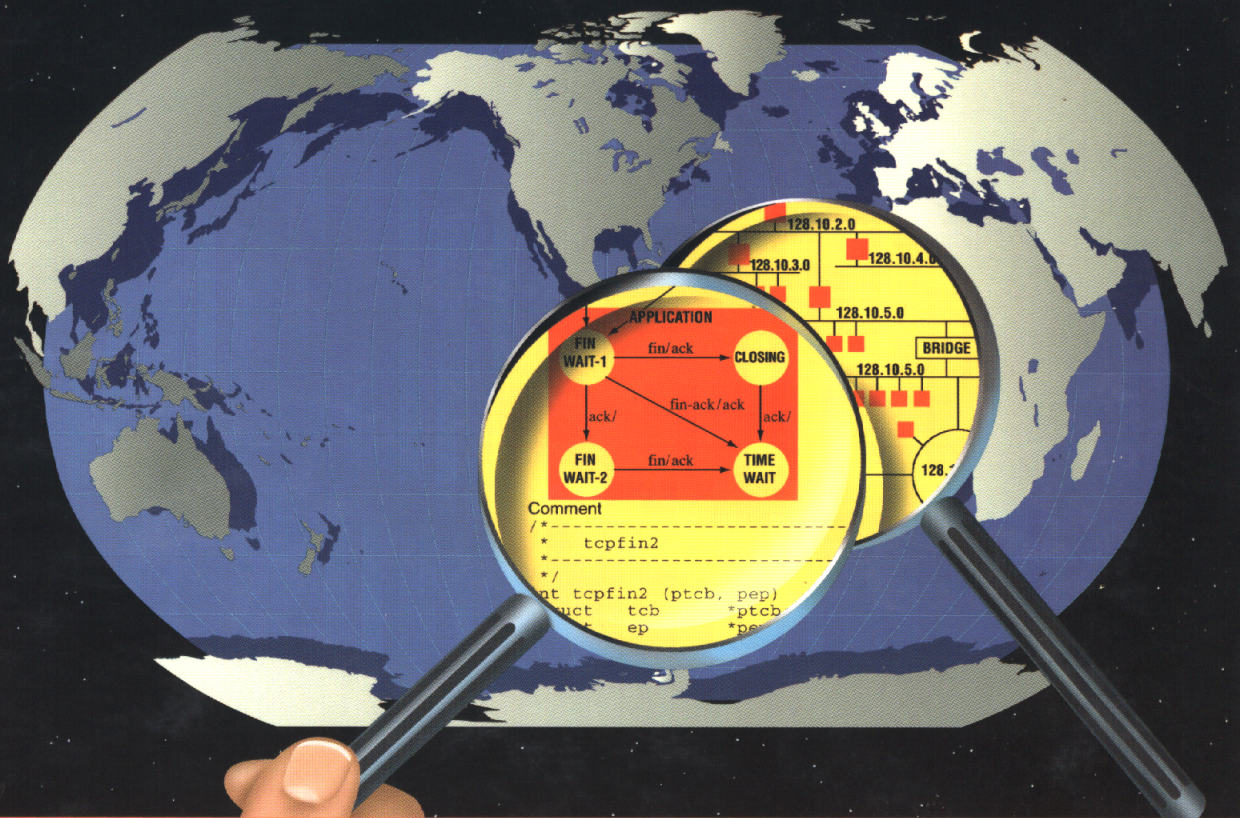
Third Edition

ANSI C
Version

INTERNETWORKING WITH TCP/IP

VOLUME II

DESIGN, IMPLEMENTATION,
AND INTERNALS



DOUGLAS E. COMER
DAVID L. STEVENS

Internetworking With TCP/IP

Vol II:

Design, Implementation, and Internals

THIRD EDITION

DOUGLAS E. COMER

*Department of Computer Sciences
Purdue University
West Lafayette, IN 47907*

and

DAVID L. STEVENS

*Sequent Computer Systems, Inc.
Beaverton, OR 97006*



PRENTICE HALL
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data
(Revised for vol. 2)

Comer, Douglas
Internetworking with TCP/IP.

Vol. 2 by Douglas E. Comer and David L. Stevens.
Includes bibliographical references and index.
Contents: Vol. 1. Principles, protocols, and
architecture—v. 2. Design, implementation, and
internals.

1. Computer networks. 2. Computer network protocols.
3. Data transmission systems. I. Stevens, David L.,
1962- . II. Title.
TK5105.5.C39 1991 004.6 90-7829
ISBN 0-13-468505-9 (v. 1)
ISBN 0-13-472242-6 (v. 2)

Publisher: **ALAN APT**
Development Editor: **SONDRA CHAVEZ**
Editor-in-chief: **MARCIA HORTON**
Production editor: **JOAN SKIDMORE - ICC OREGON**
Managing editor: **EILEEN CLARK**
Director of production and manufacturing: **DAVID W. RICCARDI**
Cover director: **HEATHER SCOTT**
Manufacturing buyer: **DONNA M. SULLIVAN**
Manufacturing manager: **TRUDY PISCIOTTI**
Editorial assistant: **TONI HOLM**

Prentice
Hall

© 1995 by Prentice-Hall, Inc.
Upper Saddle River, New Jersey 07458

All rights reserved. No part of this book may be
reproduced, in any form or by any means,
without permission in writing from the publisher.

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with or arising out of the furnishing, performance, or use of these programs.

Printed in the United States of America

10 9 8 7 6 5 4

ISBN 0-13-973843-6

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Pearson Education Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

*To the memory of net 10 and our
connections: 10.0.0.37 and 10.2.0.37*

The Comer Series Of Books On Networking And Internetworking

Internetworking With TCP/IP Volume 1: Principles Protocols, and Architecture, third edition, 1995, by Douglas Comer ISBN 0-13-468505-9

Internetworking With TCP/IP Volume II: Design, Implementation, and Internals, second edition, 1998, by Douglas Comer and David Stevens ISBN 0-13-973843-6

Internetworking With TCP/IP Volume III: Client-Server Programming and Applications, BSD Socket Version, second edition 1996, by Douglas Comer and David Stevens ISBN 0-13-260969-X

Internetworking With TCP/IP Volume III: Client-Server Programming and Applications, AT&T TLI Version, 1993, by Douglas Comer and David Stevens ISBN 0-13-474230-3

Internetworking With TCP/IP Volume III: Client-Server Programming and Applications, Window Sockets Version, 1997, by Douglas Comer and David Stevens ISBN 0-13-848714-6

Computer Networks And Internets, 1996, by Douglas Comer (with a CD-ROM by Ralph Droms) ISBN 0-13-239070-1

The Internet Book: Everything you need to know about computer networking and how the Internet works, second Edition 1997, by Douglas Comer ISBN 0-13-890161-9

Contents

Foreword	xvii
Preface	xix
Chapter 1 Introduction And Overview	1
1.1 TCP/IP Protocols	1
1.2 The Need To Understand Details	1
1.3 Complexity Of Interactions Among Protocols	2
1.4 The Approach In This Text	2
1.5 The Importance Of Studying Code	3
1.6 The Xinu Operating System	3
1.7 Organization Of The Remainder Of The Book	4
1.8 Summary	4
Chapter 2 The Structure Of TCP/IP Software In An Operating System	7
2.1 Introduction	7
2.2 The Process Concept	8
2.3 Process Priority	9
2.4 Process Synchronization	9
2.5 Interprocess Communication	12
2.6 Device Drivers, Input, And Output	14
2.7 Network Input and Interrupts	14
2.8 Passing Packets To Higher Level Protocols	16
2.9 Passing Datagrams From IP To Transport Protocols	16

- 2.10 *Delivery To Application Programs* 18
- 2.11 *Information Flow On Output* 19
- 2.12 *From TCP Through IP To Network Output* 20
- 2.13 *UDP Output* 21
- 2.14 *Summary* 21

Chapter 3 Network Interface Layer

27

- 3.1 *Introduction* 27
- 3.2 *The Network Interface Abstraction* 28
- 3.3 *Ethernet Definitions* 30
- 3.4 *Logical State Of An Interface* 34
- 3.5 *Local Host Interface* 35
- 3.6 *Buffer Management* 36
- 3.7 *Demultiplexing Incoming Packets* 38
- 3.8 *Summary* 40

Chapter 4 Address Discovery And Binding (ARP)

41

- 4.1 *Introduction* 41
- 4.2 *Conceptual Organization Of ARP Software* 42
- 4.3 *Example ARP Design* 42
- 4.4 *Data Structures For The ARP Cache* 43
- 4.5 *ARP Output Processing* 46
- 4.6 *ARP Input Processing* 51
- 4.7 *ARP Cache Management* 56
- 4.8 *ARP Initialization* 60
- 4.9 *ARP Configuration Parameters* 61
- 4.10 *Summary* 61

Chapter 5 IP: Global Software Organization

63

- 5.1 *Introduction* 63
- 5.2 *The Central Switch* 63
- 5.3 *IP Software Design* 64
- 5.4 *IP Software Organization And Datagram Flow* 65
- 5.5 *Byte-Ordering In The IP Header* 78
- 5.6 *Sending A Datagram To IP* 80
- 5.7 *Table Maintenance* 83
- 5.8 *Summary* 84

Chapter 6 IP: Routing Table And Routing Algorithm 87

- 6.1 *Introduction* 87
- 6.2 *Route Maintenance And Lookup* 87
- 6.3 *Routing Table Organization* 88
- 6.4 *Routing Table Data Structures* 89
- 6.5 *Origin Of Routes And Persistence* 91
- 6.6 *Routing A Datagram* 91
- 6.7 *Periodic Routing Table Maintenance* 98
- 6.8 *IP Options Processing* 106
- 6.9 *Summary* 107

Chapter 7 IP: Fragmentation And Reassembly 109

- 7.1 *Introduction* 109
- 7.2 *Fragmenting Datagrams* 109
- 7.3 *Implementation Of Fragmentation* 110
- 7.4 *Datagram Reassembly* 115
- 7.5 *Maintenance Of Fragment Lists* 124
- 7.6 *Initialization* 126
- 7.7 *Summary* 126

Chapter 8 IP: Error Processing (ICMP) 129

- 8.1 *Introduction* 129
- 8.2 *ICMP Message Formats* 129
- 8.3 *Implementation Of ICMP Messages* 129
- 8.4 *Handling Incoming ICMP Messages* 132
- 8.5 *Handling An ICMP Redirect Message* 134
- 8.6 *Setting A Subnet Mask* 135
- 8.7 *Choosing A Source Address For An ICMP Packet* 137
- 8.8 *Generating ICMP Error Messages* 138
- 8.9 *Avoiding Errors About Errors* 141
- 8.10 *Allocating A Buffer For ICMP* 142
- 8.11 *The Data Portion Of An ICMP Message* 144
- 8.12 *Generating An ICMP Redirect Message* 146
- 8.13 *Summary* 147

Chapter 9 IP: Multicast Processing (IGMP) 149

- 9.1 *Introduction* 149
- 9.2 *Maintaining Multicast Group Membership Information* 149
- 9.3 *A Host Group Table* 150
- 9.4 *Searching For A Host Group* 152
- 9.5 *Adding A Host Group Entry To The Table* 153
- 9.6 *Configuring The Network Interface For A Multicast Address* 155
- 9.7 *Translation Between IP and Hardware Multicast Addresses* 157
- 9.8 *Removing A Multicast Address From The Host Group Table* 159
- 9.9 *Joining A Host Group* 160
- 9.10 *Maintaining Contact With A Multicast Router* 161
- 9.11 *Implementing IGMP Membership Reports* 163
- 9.12 *Computing A Random Delay* 165
- 9.13 *A Process To Send IGMP Reports* 166
- 9.14 *Handling Incoming IGMP Messages* 167
- 9.15 *Leaving A Host Group* 169
- 9.16 *Initialization Of IGMP Data Structures* 170
- 9.17 *Summary* 171

Chapter 10 UDP: User Datagrams 173

- 10.1 *Introduction* 173
- 10.2 *UDP Ports And Demultiplexing* 173
- 10.3 *UDP Input Processing* 177
- 10.4 *UDP Output Processing* 187
- 10.5 *Summary* 190

Chapter 11 TCP: Data Structures And Input Processing 193

- 11.1 *Introduction* 193
- 11.2 *Overview Of TCP Software* 194
- 11.3 *Transmission Control Blocks* 194
- 11.4 *TCP Segment Format* 199
- 11.5 *Sequence Space Comparison* 200
- 11.6 *TCP Finite State Machine* 202
- 11.7 *Example State Transition* 204
- 11.8 *Declaration Of The Finite State Machine* 204
- 11.9 *TCB Allocation And Initialization* 206
- 11.10 *Implementation Of The Finite State Machine* 208
- 11.11 *Handling An Input Segment* 209

11.12 *Summary* 218

Chapter 12 TCP: Finite State Machine Implementation

221

- 12.1 *Introduction* 221
- 12.2 *CLOSED State Processing* 221
- 12.3 *Graceful Shutdown* 222
- 12.4 *Timed Delay After Closing* 222
- 12.5 *TIME-WAIT State Processing* 223
- 12.6 *CLOSING State Processing* 225
- 12.7 *FIN-WAIT-2 State Processing* 226
- 12.8 *FIN-WAIT-1 State Processing* 227
- 12.9 *CLOSE-WAIT State Processing* 229
- 12.10 *LAST-ACK State Processing* 231
- 12.11 *ESTABLISHED State Processing* 232
- 12.12 *Processing Urgent Data In A Segment* 233
- 12.13 *Processing Other Data In A Segment* 235
- 12.14 *Keeping Track Of Received Octets* 237
- 12.15 *Aborting A TCP Connection* 240
- 12.16 *Establishing A TCP Connection* 241
- 12.17 *Initializing A TCB* 241
- 12.18 *SYN-SENT State Processing* 243
- 12.19 *SYN-RECEIVED State Processing* 244
- 12.20 *LISTEN State Processing* 247
- 12.21 *Initializing Window Variables For A New TCB* 248
- 12.22 *Summary* 250

Chapter 13 TCP: Output Processing

251

- 13.1 *Introduction* 251
- 13.2 *Controlling TCP Output Complexity* 251
- 13.3 *The Four TCP Output States* 252
- 13.4 *TCP Output As A Process* 252
- 13.5 *TCP Output Messages* 253
- 13.6 *Encoding Output States And TCB Numbers* 254
- 13.7 *Implementation Of The TCP Output Process* 254
- 13.8 *Mutual Exclusion* 255
- 13.9 *Implementation Of The IDLE State* 256
- 13.10 *Implementation Of The PERSIST State* 256
- 13.11 *Implementation Of The TRANSMIT State* 257
- 13.12 *Implementation Of The RETRANSMIT State* 259
- 13.13 *Sending A Segment* 259

- 13.14 *Computing The TCP Data Length* 263
- 13.15 *Computing Sequence Counts* 264
- 13.16 *Other TCP Procedures* 265
- 13.17 *Summary* 271

Chapter 14 TCP: Timer Management

273

- 14.1 *Introduction* 273
- 14.2 *A General Data Structure For Timed Events* 273
- 14.3 *A Data Structure For TCP Events* 274
- 14.4 *Timers, Events, And Messages* 275
- 14.5 *The TCP Timer Process* 276
- 14.6 *Deleting A TCP Timer Event* 278
- 14.7 *Deleting All Events For A TCB* 280
- 14.8 *Determining The Time Remaining For An Event* 281
- 14.9 *Inserting A TCP Timer Event* 282
- 14.10 *Starting TCP Output Without Delay* 283
- 14.11 *Summary* 285

Chapter 15 TCP: Flow Control And Adaptive Retransmission

287

- 15.1 *Introduction* 287
- 15.2 *The Difficulties With Adaptive Retransmission* 288
- 15.3 *Tuning Adaptive Retransmission* 288
- 15.4 *Retransmission Timer And Backoff* 288
- 15.5 *Window-Based Flow Control* 291
- 15.6 *Maximum Segment Size Computation* 295
- 15.7 *Congestion Avoidance And Control* 299
- 15.8 *Slow-Start And Congestion Avoidance* 300
- 15.9 *Round Trip Estimation And Timeout* 303
- 15.10 *A Miscellaneous Note* 309
- 15.11 *Summary* 310

Chapter 16 TCP: Urgent Data Processing And The Push Function

313

- 16.1 *Introduction* 313
- 16.2 *Out-Of-Band Signaling* 313
- 16.3 *Urgent Data* 314
- 16.4 *Interpreting The Standard* 314
- 16.5 *Configuration For Berkeley Urgent Pointer Interpretation* 317
- 16.6 *Informing An Application* 317

- 16.7 *Reading Data From TCP* 318
- 16.8 *Sending Urgent Data* 320
- 16.9 *TCP Push Function* 321
- 16.10 *Interpreting Push With Out-Of-Order Delivery* 322
- 16.11 *Implementation Of Push On Input* 323
- 16.12 *Summary* 324

Chapter 17 Socket-Level Interface 327

- 17.1 *Introduction* 327
- 17.2 *Interfacing Through A Device* 327
- 17.3 *TCP Connections As Devices* 329
- 17.4 *An Example TCP Client Program* 330
- 17.5 *An Example TCP Server Program* 331
- 17.6 *Implementation Of The TCP Master Device* 333
- 17.7 *Implementation Of A TCP Slave Device* 341
- 17.8 *Initialization Of A Slave Device* 355
- 17.9 *Summary* 356

Chapter 18 RIP: Active Route Propagation And Passive Acquisition 359

- 18.1 *Introduction* 359
- 18.2 *Active And Passive Mode Participants* 360
- 18.3 *Basic RIP Algorithm And Cost Metric* 360
- 18.4 *Instabilities And Solutions* 361
- 18.5 *Message Types* 364
- 18.6 *Protocol Characterization* 365
- 18.7 *Implementation Of RIP* 366
- 18.8 *The Principle RIP Process* 369
- 18.9 *Responding To An Incoming Request* 375
- 18.10 *Generating Update Messages* 377
- 18.11 *Initializing Copies Of An Update Message* 379
- 18.12 *Generating Periodic RIP Output* 384
- 18.13 *Limitations Of RIP* 385
- 18.14 *Summary* 385

Chapter 19 OSPF: Route Propagation With An SPF Algorithm 387

- 19.1 *Introduction* 387
- 19.2 *OSPF Configuration And Options* 388
- 19.3 *OSPF's Graph-Theoretic Model* 388

19.4	<i>OSPF Declarations</i>	392
19.5	<i>Adjacency And Link State Propagation</i>	398
19.6	<i>Discovering Neighboring Gateways With Hello</i>	399
19.7	<i>Sending Hello Packets</i>	401
19.8	<i>Designated Router Concept</i>	407
19.9	<i>Electing A Designated Router</i>	407
19.10	<i>Reforming Adjacencies After A Change</i>	411
19.11	<i>Handling Arriving Hello Packets</i>	414
19.12	<i>Adding A Gateway To The Neighbor List</i>	416
19.13	<i>Neighbor State Transitions</i>	418
19.14	<i>OSPF Timer Events And Retransmissions</i>	420
19.15	<i>Determining Whether Adjacency Is Permitted</i>	422
19.16	<i>Handling OSPF input</i>	423
19.17	<i>Declarations And Procedures For Link State Processing</i>	426
19.18	<i>Generating Database Description Packets</i>	429
19.19	<i>Creating A Template</i>	430
19.20	<i>Transmitting A Database Description Packet</i>	431
19.21	<i>Handling An Arriving Database Description Packet</i>	433
19.22	<i>Handling Link State Request Packets</i>	440
19.23	<i>Building A Link State Summary</i>	442
19.24	<i>OSPF Utility Procedures</i>	443
19.25	<i>Summary</i>	446

Chapter 20 SNMP: MIB Variables, Representations, And Bindings **449**

20.1	<i>Introduction</i>	449
20.2	<i>Server Organization And Name Mapping</i>	450
20.3	<i>MIB Variables</i>	451
20.4	<i>MIB Variable Names</i>	452
20.5	<i>Lexicographic Ordering Among Names</i>	453
20.6	<i>Prefix Removal</i>	453
20.7	<i>Operations Applied To MIB Variables</i>	454
20.8	<i>Names For Tables</i>	454
20.9	<i>Conceptual Threading Of The Name Hierarchy</i>	455
20.10	<i>Data Structure For MIB Variables</i>	456
20.11	<i>A Data Structure For Fast Lookup</i>	459
20.12	<i>Implementation Of The Hash Table</i>	460
20.13	<i>Specification Of MIB Bindings</i>	460
20.14	<i>Internal Variables Used In Bindings</i>	467
20.15	<i>Hash Table Lookup</i>	469
20.16	<i>SNMP Structures And Constants</i>	471
20.17	<i>ASN.1 Representation Manipulation</i>	477
20.18	<i>Summary</i>	488

Chapter 21 SNMP: Client And Server 491

- 21.1 *Introduction* 491
- 21.2 *Data Representation In The Server* 491
- 21.3 *Server Implementation* 492
- 21.4 *Parsing An SNMP Message* 495
- 21.5 *Converting ASN.1 Names In The Binding List* 500
- 21.6 *Resolving A Query* 501
- 21.7 *Interpreting The Get-Next Operation* 504
- 21.8 *Indirect Application Of Operations* 504
- 21.9 *Indirection For Tables* 507
- 21.10 *Generating A Reply Message Backward* 509
- 21.11 *Converting From Internal Form to ASN.1* 512
- 21.12 *Utility Functions Used By The Server* 514
- 21.13 *Implementation Of An SNMP Client* 515
- 21.14 *Initialization Of Variables* 517
- 21.15 *Summary* 519

Chapter 22 SNMP: Table Access Functions 521

- 22.1 *Introduction* 521
- 22.2 *Table Access* 522
- 22.3 *Object Identifiers For Tables* 522
- 22.4 *Address Entry Table Functions* 522
- 22.5 *Net-To-Media Table Functions* 530
- 22.6 *Network Interface Table Functions* 541
- 22.7 *Routing Table Functions* 550
- 22.8 *TCP Connection Table Functions* 560
- 22.9 *UDP Listener Table* 569
- 22.10 *Utility Routines To Convert IP Addresses* 576
- 22.11 *Summary* 578

Chapter 23 Implementation In Retrospect 579

- 23.1 *Introduction* 579
- 23.2 *Statistical Analysis Of The Code* 579
- 23.3 *Lines Of Code For Each Protocol* 580
- 23.4 *Functions And Procedures For Each Protocol* 582
- 23.5 *Summary* 583

Appendix 1 Cross Reference Of Procedure Calls	585
Appendix 2 Cross Reference Of C Structures Used In The Code	607
Appendix 3 Xinu Functions And Constants Used In The Code	613
Bibliography	631
Index	639