

Addison
Wesley

TURING

图灵原版计算机科学系列

TCP/IP 详解

卷2：实现

英文版

TCP/IP Illustrated

Volume 2: The Implementation

[美] Gary R. Wright 著
W. Richard Stevens



人民邮电出版社
POSTS & TELECOM PRESS

TCP/IP 详解

卷2：实现

英文版

TCP/IP Illustrated
Volume 2: The Implementation

[美] Gary R. Wright 著
W. Richard Stevens

人民邮电出版社
北京

图书在版编目 (CIP) 数据

TCP/IP详解. 卷2, 实现 = TCP/IP Illustrated,
Volume 2: The Implementation, First Edition: 英文
/ (美) 赖特 (Wright, G. R.), (美) 史蒂文斯
(Stevens, W. R.) 著. —北京: 人民邮电出版社,
2010.4

(图灵原版计算机科学系列)

ISBN 978-7-115-22248-0

I. ①T … II. ①赖… ②史… III. ①计算机网络—通
信协议—英文 IV. ①TN915.04

中国版本图书馆CIP数据核字 (2010) 第014975号

内 容 提 要

本书是TCP/IP领域的经典之作! 书中完整而详细地介绍了TCP/IP协议是如何实现的。本书介绍了一个实际的TCP/IP实现, 并给出了这一实现的完整源代码, 帮助读者全面掌握TCP/IP的实现。本书内容详尽且具权威性, 几乎每章都提供精选的习题, 并在附录中提供了部分习题的答案。

本书适合任何希望了解TCP/IP协议如何实现的读者阅读, 更是TCP/IP领域研究人员和开发人员的权威参考书。

图灵原版计算机科学系列

TCP/IP详解 卷2: 实现 (英文版)

-
- ◆ 著 [美] Gary R. Wright W. Richard Stevens
责任编辑 杨海玲
 - ◆ 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>
北京艺辉印刷有限公司印刷
 - ◆ 开本: 800×1000 1/16
印张: 74.75
字数: 1440 千字 2010年4月第1版
印数: 1-2 000 册 2010年4月北京第1次印刷

著作权合同登记号 图字: 01-2010-0310号

ISBN 978-7-115-22248-0

定价: 139.00元

读者服务热线: (010) 51095186 印装质量热线: (010) 67129223

反盗版热线: (010) 67171154

版 权 声 明

Original edition, entitled *TCP/IP Illustrated, Volume 2: The Implementation*, First Edition, 9780201633542 by Gary R.Wright and W.Richard Stevens, published by Pearson Education, Inc., publishing as Addison-Wesley, Copyright © 1995 by Pearson Education, Inc.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc.

China edition published by PEARSON EDUCATION ASIA LTD. and POSTS & TELECOM PRESS Copyright © 2010.

This edition is manufactured in the People's Republic of China, and is authorized for sale only in the People's Republic of China excluding Hong Kong, Macao and Taiwan.

本书英文版由Pearson Education Asia Ltd. 授权人民邮电出版社独家出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

仅限于中华人民共和国境内（香港、澳门特别行政区和台湾地区除外）销售发行。

本书封面贴有Pearson Education（培生教育出版集团）激光防伪标签，无标签者不得销售。

版权所有，侵权必究。

献给我的父母和姐姐，
感谢他们的爱与支持。
——Gary R. Wright

献给我的父母，
感谢他们让我接受了良好的教育，
并示我以优良的职业习惯。
——W. Richard Stevens

前言

概述

本书介绍并给出了TCP/IP的常见参考实现的源代码，即由加州大学伯克利分校计算机系统研究组（CSRG）研发的实现。历史上该实现是随4.x BSD（Berkeley Software Distribution）系统一起发布的。最早的发布版本出现于1982年，随后经过了多次大改、多次微调，并实现了映射到其他Unix及非Unix系统的众多端口。这个实现不可小觑，它是世界范围内无数系统中日常运行的TCP/IP实现的基础。该实现还提供了路由器功能，使我们可以展示出TCP/IP的主机实现与路由器的区别。

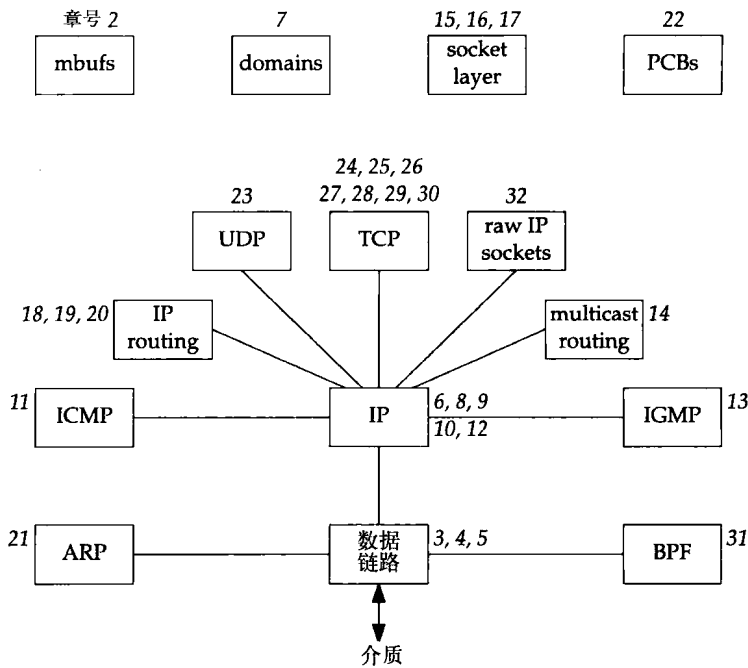
本书将描述该实现并给出TCP/IP核心实现的全部源代码，大约有15 000行C代码。书中介绍的Berkeley代码的版本是4.4BSD-Lite发布版本。这个代码于1994年4月公布，它在1988年的4.3BSD Tahoe发布版本、1990年的4.3BSD Reno发布版本以及1993年的4.4BSD发布版本基础上增加了大量的网络增强技术（附录B讲述了如何获取这些源代码）。4.4BSD发布版本提供了最新的TCP/IP特性，例如对多播（multicast）和长肥管道（long fat pipe）的支持（用于高带宽、长延迟的通道）。图1-1（第4页）提供了各版本的Berkeley网络代码的更多细节。

本书面向所有希望了解TCP/IP协议实现原理的读者：编写网络应用的程序员、利用TCP/IP维护计算机系统与网络的系统管理员以及希望了解实际操作系统中大型代码的集成

原理的程序员。

本书的结构

下面的图给出了本书涉及的各种协议和子系统，方框上的斜体数字指明了相应内容在哪一章讨论。



我们采用一种自底向上的方式来介绍TCP/IP协议族，首先是数据链路层，然后是网络层（IP、ICMP、IGMP、IP路由、多播路由）以及随后的套接字层，最后是传输层（UDP、TCP和原始IP）。

致读者

本书的读者应该对TCP/IP协议的工作原理有基本的了解。对TCP/IP协议不是很熟悉的读者应首先参考本套书的第1卷[Stevens, 1994]，该书对TCP/IP协议族有比较透彻的描述。本书将之前的第1卷称为卷1（Volume 1）。本书同时还需要读者对操作系统原理有初步的了解。

我们采用数据结构方法来描述协议的实现。每章不仅给出源代码，还用图片和文字描述源代码使用或者维护的数据结构。我们还展示了如何把这些数据结构集成到TCP/IP与内核所用的数据结构中。全书使用了大量的图表——超过250个。

这种数据结构方法使得读者可以用多种方式使用本书。那些对所有的实现细节都感兴趣的读者可以结合源代码，把本书从头一直读到最后。其他读者则可以通过理解数据

结构和阅读文字内容来理解协议的工作原理，而不一定非要参考源代码。

估计会有许多读者只是对本书中特定的章节感兴趣，并希望直接阅读这些章。全书贯穿了许多交叉引用，并提供了完整的索引，因此读者可以独立阅读各章的内容。索引后面还按照字母表顺序给出了书中所有函数和宏的交叉引用，以及相关详细信息的起始页码。每章的末尾都布置了一些习题，附录A给出了大部分习题的答案，这样做是为了本书更适合作为自学参考书。

源代码版权

本书中除了图1-2和图8-27之外的所有源代码都来自4.4BSD-Lite发布版。这个软件是公用程序，可以通过很多方式获得（详见附录B）。

书中的所有源代码版权说明如下：

```
/*
 * 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.
 */
```

致谢

感谢百忙之中拨冗阅读本书书稿并给出重要反馈的技术审稿人：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、Chakravarthi Ravi、Peter Salus、Doug Schmidt、Keith Sklower、Ian Lance Taylor和 G. N. Ananda Vardhana。特别感谢顾问编辑Brian Kernighan，在完成本书的过程中，他提出了很多及时、透彻、很有帮助的评审意见，并始终鼓励和支持着我。

我们（再次）感谢美国国家光学天文台，尤其是授权我们接入其网络和主机的Sidney Wolff、Richard Wolff和Steve Grandi。我们还要感谢加州大学伯克利分校计算机系统研究组的Keith Bostic和Kirk McKusick授权我们使用最新的4.4BSD系统，Keith Sklower修改了4.4BSD-Lite软件使其能运行于BSD/386 V1.1系统。

Gary R. Wright感谢John Wait多年以来的好心督促、Dave Schaller的鼓励以及Jim Hogue在本书写作及出版过程中的支持。

W. Richard Stevens再次感谢他的家人，他们又一次忍受了他的“小型”著书项目。感谢你们：Sally、Bill、Ellen和David。

Addison-Wesley团队的辛勤工作、专业水准和鼎力支持使作者的工作轻松了很多。我们特别感谢John Wait的指导和Kim Dawley的创新思维。

作者制作了本书的最终电子版。描述工业级强度的软件系统就需要使用工业级强度的文本处理系统。本书的一位作者选择使用James Clark的Groff包，另一位作者只能勉强同意。

欢迎读者以电子邮件的方式反馈意见、提出建议或订正错误。两位作者都很乐意指出对方残存的错误。

Gary R. Wright
于康涅狄格州米德尔敦市

W. Richard Stevens
于亚利桑那州图森市
1994年11月

Contents

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

Chapter 11.	ICMP: Internet Control Message Protocol	301
11.1	Introduction	301
11.2	Code Introduction	305
11.3	icmp Structure	308
11.4	ICMP protosw Structure	309
11.5	Input Processing: icmp_input Function	310
11.6	Error Processing	313
11.7	Request Processing	316
11.8	Redirect Processing	321
11.9	Reply Processing	323
11.10	Output Processing	324
11.11	icmp_error Function	324
11.12	icmp_reflect Function	328
11.13	icmp_send Function	333
11.14	icmp_sysctl Function	334
11.15	Summary	335
Chapter 12.	IP Multicasting	337
12.1	Introduction	337
12.2	Code Introduction	340
12.3	Ethernet Multicast Addresses	341
12.4	ether_multi Structure	342
12.5	Ethernet Multicast Reception	344
12.6	in_multi Structure	345
12.7	ip_moptions Structure	347
12.8	Multicast Socket Options	348
12.9	Multicast TTL Values	348
12.10	ip_setmoptions Function	351
12.11	Joining an IP Multicast Group	355
12.12	Leaving an IP Multicast Group	366
12.13	ip_getmoptions Function	371
12.14	Multicast Input Processing: ipintr Function	373
12.15	Multicast Output Processing: ip_output Function	375
12.16	Performance Considerations	379
12.17	Summary	379
Chapter 13.	IGMP: Internet Group Management Protocol	381
13.1	Introduction	381
13.2	Code Introduction	382
13.3	igmp Structure	384
13.4	IGMP protosw Structure	384
13.5	Joining a Group: igmp_joingroup Function	386
13.6	igmp_fasttimo Function	387
13.7	Input Processing: igmp_input Function	391
13.8	Leaving a Group: igmp_leavegroup Function	395
13.9	Summary	396

Chapter 14.	IP Multicast Routing	397
14.1	Introduction	397
14.2	Code Introduction	398
14.3	Multicast Output Processing Revisited	399
14.4	mrouted Daemon	401
14.5	Virtual Interfaces	404
14.6	IGMP Revisited	411
14.7	Multicast Routing	416
14.8	Multicast Forwarding: ip_mforward Function	424
14.9	Cleanup: ip_mrouter_done Function	433
14.10	Summary	434
Chapter 15.	Socket Layer	435
15.1	Introduction	435
15.2	Code Introduction	436
15.3	socket Structure	437
15.4	System Calls	441
15.5	Processes, Descriptors, and Sockets	445
15.6	socket System Call	447
15.7	getsock and sockargs Functions	451
15.8	bind System Call	453
15.9	listen System Call	455
15.10	tsleep and wakeup Functions	456
15.11	accept System Call	457
15.12	sonewconn and soisconnected Functions	461
15.13	connect System call	464
15.14	shutdown System Call	468
15.15	close System Call	471
15.16	Summary	474
Chapter 16.	Socket I/O	475
16.1	Introduction	475
16.2	Code Introduction	475
16.3	Socket Buffers	476
16.4	write, writev, sendto, and sendmsg System Calls	480
16.5	sendmsg System Call	483
16.6	sendit Function	485
16.7	sosend Function	489
16.8	read, readv, recvfrom, and recvmsg System Calls	500
16.9	recvmsg System Call	501
16.10	recvit Function	503
16.11	soreceive Function	505
16.12	soreceive Code	510
16.13	select System Call	524
16.14	Summary	534

Chapter 17.	Socket Options	537
17.1	Introduction	537
17.2	Code Introduction	538
17.3	setsockopt System Call	539
17.4	getsockopt System Call	545
17.5	fcntl and ioctl System Calls	548
17.6	getsockname System Call	554
17.7	getpeername System Call	554
17.8	Summary	557
Chapter 18.	Radix Tree Routing Tables	559
18.1	Introduction	559
18.2	Routing Table Structure	560
18.3	Routing Sockets	569
18.4	Code Introduction	570
18.5	Radix Node Data Structures	573
18.6	Routing Structures	578
18.7	Initialization: route_init and rtable_init Functions	581
18.8	Initialization: rn_init and rn_inithead Functions	584
18.9	Duplicate Keys and Mask Lists	587
18.10	rn_match Function	591
18.11	rn_search Function	599
18.12	Summary	599
Chapter 19.	Routing Requests and Routing Messages	601
19.1	Introduction	601
19.2	rtalloc and rtalloc1 Functions	601
19.3	RTFREE Macro and rtfree Function	604
19.4	rtrequest Function	607
19.5	rt_setgate Function	612
19.6	rtinit Function	615
19.7	rtredirect Function	617
19.8	Routing Message Structures	621
19.9	rt_missmsg Function	625
19.10	rt_ifmsg Function	627
19.11	rt_newaddrmsg Function	628
19.12	rt_msg1 Function	630
19.13	rt_msg2 Function	632
19.14	sysctl_rtable Function	635
19.15	sysctl_dumpentry Function	640
19.16	sysctl_iflist Function	642
19.17	Summary	644
Chapter 20.	Routing Sockets	645
20.1	Introduction	645
20.2	routedomain and protosw Structures	646
20.3	Routing Control Blocks	647

20.4	raw_init Function	647	
20.5	route_output Function	648	
20.6	rt_xaddrs Function	660	
20.7	rt_setmetrics Function	661	
20.8	raw_input Function	662	
20.9	route_usrreq Function	664	
20.10	raw_usrreq Function	666	
20.11	raw_attach, raw_detach, and raw_disconnect Functions		671
20.12	Summary	672	
Chapter 21.	ARP: Address Resolution Protocol		675
21.1	Introduction	675	
21.2	ARP and the Routing Table	675	
21.3	Code Introduction	678	
21.4	ARP Structures	681	
21.5	arpwhoas Function	683	
21.6	arprequest Function	684	
21.7	arpintr Function	687	
21.8	in_arpinput Function	688	
21.9	ARP Timer Functions	694	
21.10	arpresolve Function	696	
21.11	arplookup Function	701	
21.12	Proxy ARP	703	
21.13	arp_rtrequest Function	704	
21.14	ARP and Multicasting	710	
21.15	Summary	711	
Chapter 22.	Protocol Control Blocks		713
22.1	Introduction	713	
22.2	Code Introduction	715	
22.3	inpcb Structure	716	
22.4	in_pcballoc and in_pcbdetach Functions		717
22.5	Binding, Connecting, and Demultiplexing		719
22.6	in_pcblookup Function	724	
22.7	in_pcbbind Function	728	
22.8	in_pcbconnect Function	735	
22.9	in_pcbdisconnect Function	741	
22.10	in_setsockaddr and in_setpeeraddr Functions		741
22.11	in_pcbnotify, in_rtchange, and in_losing Functions		742
22.12	Implementation Refinements	750	
22.13	Summary	751	
Chapter 23.	UDP: User Datagram Protocol		755
23.1	Introduction	755	
23.2	Code Introduction	755	
23.3	UDP protosw Structure	758	

23.4	UDP Header	759	
23.5	udp_init Function	760	
23.6	udp_output Function	760	
23.7	udp_input Function	769	
23.8	udp_saveopt Function	781	
23.9	udp_ctlinput Function	782	
23.10	udp_usrreq Function	784	
23.11	udp_sysctl Function	790	
23.12	Implementation Refinements	791	
23.13	Summary	793	
Chapter 24.	TCP: Transmission Control Protocol		795
24.1	Introduction	795	
24.2	Code Introduction	795	
24.3	TCP protocol Structure	801	
24.4	TCP Header	801	
24.5	TCP Control Block	803	
24.6	TCP State Transition Diagram	805	
24.7	TCP Sequence Numbers	807	
24.8	tcp_init Function	812	
24.9	Summary	815	
Chapter 25.	TCP Timers		817
25.1	Introduction	817	
25.2	Code Introduction	819	
25.3	tcp_canceltimers Function	821	
25.4	tcp_fasttimo Function	821	
25.5	tcp_slowtimo Function	822	
25.6	tcp_timers Function	824	
25.7	Retransmission Timer Calculations	831	
25.8	tcp_newtcpcb Function	833	
25.9	tcp_setpersist Function	835	
25.10	tcp_xmit_timer Function	836	
25.11	Retransmission Timeout: tcp_timers Function	841	
25.12	An RTT Example	846	
25.13	Summary	848	
Chapter 26.	TCP Output		851
26.1	Introduction	851	
26.2	tcp_output Overview	852	
26.3	Determine if a Segment Should be Sent	852	
26.4	TCP Options	864	
26.5	Window Scale Option	866	
26.6	Timestamp Option	866	
26.7	Send a Segment	871	
26.8	tcp_template Function	884	
26.9	tcp_respond Function	885	
26.10	Summary	888	