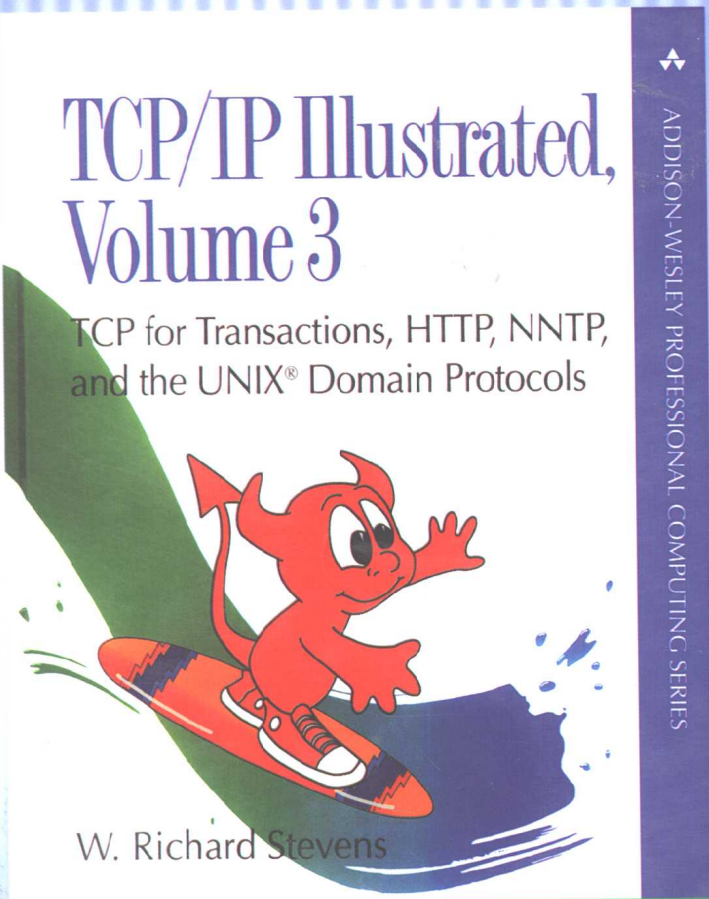
经 典 原 版 书 库

TCP/IP详解

卷3: TCP事务协议、HTTP、NNTP和UNIX域协议

（英文版）



（美）W. Richard Stevens 著

机械工业出版社
China Machine Press
Addison-Wesley

经 典 原 版 书 库

(英文版)

TCP/IP 详解

卷 3: TCP事务协议、HTTP、NNTP和UNIX域协议

TCP/IP Illustrated, Volume 3:

TCP for Transactions, HTTP, NNTP, and the UNIX® Domain Protocols

(美) W. Richard Stevens 著

机械工业出版社
China Machine Press

本书版权登记号：图字：01-2001-5012

图书在版编目（CIP）数据

TCP/IP详解　卷3：TCP事务协议、HTTP、NNTP和UNIX域协议（英文版）/（美）史蒂文斯（Stevens, W. R.）著. -北京：机械工业出版社，2002.1

（经典原版书库）

ISBN 7-111-09503-0

Ⅰ.T…　Ⅱ.史…　Ⅲ.计算机网络-通信协议-英文　Ⅳ.TN915.04

中国版本图书馆CIP数据核字（2001）第081796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：华　章

北京昌平奔腾印刷厂印刷·新华书店北京发行所发行

2002年1月第1版·2002年3月第2次印刷

880mm×1230mm 1/32·11.125印张

印数：2 001-4 000册

定价：28.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

ACRONYMS

ACK	acknowledgment flag; TCP header
ANSI	American National Standards Institute
API	application program interface
ARP	Address Resolution Protocol
ARPANET	Advanced Research Projects Agency network
ASCII	American Standard Code for Information Interchange
BPF	BSD Packet Filter
BSD	Berkeley Software Distribution
CC	connection count; T/TCP
CERT	Computer Emergency Response Team
CR	carriage return
DF	don't fragment flag; IP header
DNS	Domain Name System
EOL	end of option list
FAQ	frequently asked question
FIN	finish flag; TCP header
FTP	File Transfer Protocol
GIF	graphics interchange format
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ICMP	Internet Control Message Protocol
IEEE	Institute of Electrical and Electronics Engineers
INN	InterNet News
INND	InterNet News Daemon
IP	Internet Protocol
IPC	interprocess communication
IRTP	Internet Reliable Transaction Protocol
ISN	initial sequence number
ISO	International Organization for Standardization
ISS	initial send sequence number
LAN	local area network
LF	linefeed
MIME	multipurpose Internet mail extensions
MSL	maximum segment lifetime
MSS	maximum segment size
MTU	maximum transmission unit

NCSA	National Center for Supercomputing Applications
NFS	Network File System
NNRP	Network News Reading Protocol
NNTP	Network News Transfer Protocol
NOAO	National Optical Astronomy Observatories
NOP	no operation
OSF	Open Software Foundation
OSI	open systems interconnection
PAWS	protection against wrapped sequence numbers
PCB	protocol control block
POSIX	Portable Operating System Interface
PPP	Point-to-Point Protocol
PSH	push flag; TCP header
RDP	Reliable Datagram Protocol
RFC	Request for Comment
RPC	remote procedure call
RST	reset flag; TCP header
RTO	retransmission time out
RTT	round-trip time
SLIP	Serial Line Internet Protocol
SMTP	Simple Mail Transfer Protocol
SPT	server processing time
SVR4	System V Release 4
SYN	synchronize sequence numbers flag; TCP header
TAO	TCP accelerated open
TCP	Transmission Control Protocol
TTL	time-to-live
Telnet	remote terminal protocol
UDP	User Datagram Protocol
URG	urgent pointer flag; TCP header
URI	uniform resource identifier
URL	uniform resource locator
URN	uniform resource name
VMTP	Versatile Message Transaction Protocol
WAN	wide area network
WWW	World Wide Web

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭橥了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短、从业人员较少的现状下，美国等发达国家在其计算机科学发展的几十年间积淀的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章图文信息有限公司较早意识到“出版要为教育服务”。自1998年始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall，Addison-Wesley，McGraw-Hill，Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum，Stroustrup，Kernighan，Jim Gray等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及庋藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专诚为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：针对本科生的核心课程，剔抉外版菁华而成“国外经典教材”系列；对影印版的教材，则单独开辟出“经典原版书库”；定位在高级教程和专业参考的“计算机科学丛书”还将保持原来的风格，继续出版新的品种。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师们服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

“经典原版书库”是响应教育部提出的使用原版国外教材的号召，为国内高校的计算机教学度身订造的。在广泛地征求并听取丛书的“专家指导委员会”的意见后，我们最终选定了这30多种篇幅内容适度、讲解鞭辟入里的教材，其中的大部分已经被M.I.T.、Stanford、U.C. Berkley、C.M.U.等世界名牌大学采用。丛书不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程，而且各具特色——有的出自语言设计者之手、有的历三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下，读者必将在计算机科学的宫殿中由登堂而入室。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证，但我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。教材的出版只是我们的后续服务的起点。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

电子邮件：hzedu@hzbook.com
联系电话：（010）68995265
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

专家指导委员会

（按姓氏笔画顺序）

Praise for *TCP/IP Illustrated, Volume 3: TCP for Transactions, HTTP, NNTP, and the UNIX® Domain Protocols*

"An absolutely wonderful example of how to apply scientific thinking and analysis to a technological problem...it is the highest caliber of technical writing and thinking."

— Marcus J. Ranum, Firewall Architect

"A worthy successor that continues the series' standards of excellence for both clarity and accuracy. The coverage of T/TCP and HTTP is particularly timely, given the explosion of the World Wide Web."

— Vern Paxson, Network Research Group,
Lawrence Berkeley National Laboratory

"The coverage of the HTTP protocol will be invaluable to anyone who needs to understand the detailed behavior of web servers."

— Jeffrey Mogul, Digital Equipment Corporation

"Volume 3 *is a natural addition to the series. It covers the network aspects of Web services and transaction TCP in depth."*

— Pete Haverlock, Program Manager, IBM

"In this latest volume of TCP/IP Illustrated, *Rich Stevens maintains the high standards he set up in the previous volumes: clear presentation and technical accuracy to the finest detail."*

— Andras Olah, University of Twente

"This volume maintains the superb quality of the earlier ones in the series, extending the in-depth examination of networking implementation in new directions. The entire series is a must for anybody who is seriously interested in how the Internet works today."

— Ian Lance Taylor, Author of GNU/Taylor UUCP

Preface

Introduction and Organization of the Book

This book is a logical continuation of the *TCP/IP Illustrated* series: [Stevens 1994], which we refer to as *Volume 1*, and [Wright and Stevens 1995], which we refer to as *Volume 2*. This book is divided into three parts, each covering a different topic:

1. TCP for transactions, commonly called T/TCP. This is an extension to TCP designed to make client–server transactions faster, more efficient, and reliable. This is done by omitting TCP's three-way handshake at the beginning of a connection and shortening the TIME_WAIT state at the end of a connection. We'll see that T/TCP can match UDP's performance for a client–server transaction and that T/TCP provides reliability and adaptability, both major improvements over UDP.

   A transaction is defined to be a client request to a server, followed by the server's reply. (The term *transaction* does not mean a database transaction, with locking, two-phase commit, and backout.)

2. TCP/IP applications, specifically HTTP (the Hypertext Transfer Protocol, the foundation of the World Wide Web) and NNTP (the Network News Transfer Protocol, the basis for the Usenet news system).

3. The Unix domain protocols. These protocols are provided by all Unix TCP/IP implementations and on many non-Unix implementations. They provide a form of interprocess communication (IPC) and use the same sockets interface used with TCP/IP. When the client and server are on the same host, the Unix domain protocols are often twice as fast as TCP/IP.

Part 1, the presentation of T/TCP, is in two pieces. Chapters 1–4 describe the protocol and provide numerous examples of how it works. This material is a major expansion of the brief presentation of T/TCP in Section 24.7 of Volume 1. The second piece, Chapters 5–12, describes the actual implementation of T/TCP within the 4.4BSD-Lite networking code (i.e., the code presented in Volume 2). Since the first T/TCP implementation was not released until September 1994, about one year after Volume 1 was published and right as Volume 2 was being completed, the detailed presentation of T/TCP, with examples and all the implementation details, had to wait for another volume in the series.

Part 2, the HTTP and NNTP applications, are a continuation of the TCP/IP applications presented in Chapters 25–30 of Volume 1. In the two years since Volume 1 was published, the popularity of HTTP has grown enormously, as the Internet has exploded, and the use of NNTP has been growing about 75% per year for more than 10 years. HTTP is also a wonderful candidate for T/TCP, given its typical use of TCP: short connections with small amounts of data transferred, where the total time is often dominated by the connection setup and teardown. The heavy use of HTTP (and therefore TCP) on a busy Web server by thousands of different and varied clients also provides a unique opportunity to examine the actual packets at the server (Chapter 14) and look at many features of TCP/IP that were presented in Volumes 1 and 2.

The Unix domain protocols in Part 3 were originally considered for Volume 2 but omitted when its size reached 1200 pages. While it may seem odd to cover protocols other than TCP/IP in a series titled *TCP/IP Illustrated*, the Unix domain protocols were implemented almost 15 years ago in 4.2BSD alongside the first implementation of BSD TCP/IP. They are used heavily today in any Berkeley-derived kernel, but their use is typically "under the covers," and most users are unaware of their presence. Besides being the foundation for Unix pipes on a Berkeley-derived kernel, another heavy user is the X Window System, when the client and server are on the same host (i.e., on typical workstations). Unix domain sockets are also used to pass descriptors between processes, a powerful technique for interprocess communication. Since the sockets API (application program interface) used with the Unix domain protocols is nearly identical to the sockets API used with TCP/IP, the Unix domain protocols provide an easy way to enhance the performance of local applications with minimal code changes.

Each of the three parts can be read by itself.

Readers

As with the previous two volumes in the series, this volume is intended for anyone wishing to understand how the TCP/IP protocols operate: programmers writing network applications, system administrators responsible for maintaining computer systems and networks utilizing TCP/IP, and users who deal with TCP/IP applications on a daily basis.

Parts 1 and 2 assume 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. The first half of Part 1

(Chapters 1–4, the concepts behind T/TCP along with examples) can be read independent of Volume 2, but the remainder of Part 1 (Chapters 5–12, the implementation of T/TCP) assumes familiarity with the 4.4BSD-Lite networking code, as provided with Volume 2.

Many forward and backward references are provided throughout the text, to both topics within this text, and to relevant sections of Volumes 1 and 2 for readers interested in more details. A thorough index is provided, and a list of all the acronyms used throughout the text, along with the compound term for the acronym, appears on the inside front covers. The inside back covers contain an alphabetical cross-reference of all the structures, functions, and macros described in the book and the starting page number of the description. This cross-reference also refers to definitions in Volume 2, when that object is referenced from the code in this volume.

Source Code Copyright

All the source code in this book that is taken from the 4.4BSD-Lite release 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.
 */
```

The routing table code in Chapter 6 contains the following copyright notice:

```
/*
 * Copyright 1994, 1995 Massachusetts Institute of Technology
 *
 * Permission to use, copy, modify, and distribute this software and
 * its documentation for any purpose and without fee is hereby
 * granted, provided that both the above copyright notice and this
 * permission notice appear in all copies, that both the above
 * copyright notice and this permission notice appear in all
 * supporting documentation, and that the name of M.I.T. not be used
 * in advertising or publicity pertaining to distribution of the
 * software without specific, written prior permission.  M.I.T. makes
 * no representations about the suitability of this software for any
 * purpose.  It is provided "as is" without express or implied
 * warranty.
 *
 * THIS SOFTWARE IS PROVIDED BY M.I.T. ''AS IS''.  M.I.T. DISCLAIMS
 * ALL EXPRESS OR IMPLIED WARRANTIES WITH REGARD TO THIS SOFTWARE,
 * INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF
 * MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT
 * SHALL M.I.T. 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.
 */
```

Typographical Conventions

When we display interactive input and output we'll show our typed input in a **bold font**, and the computer output like this. *Comments are added in italics.*

```
sun % telnet www.aw.com 80          connect to the HTTP server
Trying 192.207.117.2...             this line and next output by Telnet client
Connected to aw.com.
```

We always include the name of the system as part of the shell prompt (sun in this example) to show on which host the command was run. The names of programs referred to in the text are normally capitalized (e.g., Telnet and Tcpdump) to avoid excessive font changes.

> Throughout the text we'll use indented, parenthetical notes such as this to describe historical points or implementation details.

Acknowledgments

First and foremost I thank my family, Sally, Bill, Ellen, and David, who have endured another book along with all my traveling during the past year. This time, however, it really is a "small" book.

I thank the technical reviewers who read the manuscript and provided important feedback on a tight timetable: Sami Boulos, Alan Cox, Tony DeSimone, Pete Haverlock, Chris Heigham, Mukesh Kacker, Brian Kernighan, Art Mellor, Jeff Mogul, Marianne Mueller, Andras Olah, Craig Partridge, Vern Paxson, Keith Sklower, Ian Lance Taylor, and Gary Wright. A special thanks to the consulting editor, Brian Kernighan, for his rapid, thorough, and helpful reviews throughout the course of the book, and for his continued encouragement and support.

Special thanks are also due Vern Paxson and Andras Olah for their incredibly detailed reviews of the entire manuscript, finding many errors and providing valuable technical suggestions. My thanks also to Vern Paxson for making available his software for analyzing Tcpdump traces, and to Andras Olah for his help with T/TCP over the past year. My thanks also to Bob Braden, the designer of T/TCP, who provided the reference source code implementation on which Part 1 of this book is based.

Others helped in significant ways. Gary Wright and Jim Hogue provided the system on which the data for Chapter 14 was collected. Doug Schmidt provided a copy of the public domain TTCP program that uses Unix domain sockets, for the timing measurements in Chapter 16. Craig Partridge provided a copy of the RDP source code to examine. Mike Karels answered lots of questions.

My thanks once again to the National Optical Astronomy Observatories (NOAO), Sidney Wolff, Richard Wolff, and Steve Grandi, for providing access to their networks and hosts.

Finally, my thanks to all the staff at Addison-Wesley, who have helped over the past years, especially my editor John Wait.

As usual, camera-ready copy of the book was produced by the author, a Troff diehard, using the Groff package written by James Clark. I welcome electronic mail from any readers with comments, suggestions, or bug fixes.

Tucson, Arizona
November 1995

W. Richard Stevens
rstevens@noao.edu
http://www.noao.edu/~rstevens

Contents