Addison Wesley

TURING

图灵原版计算机科学系列

TCP/IP详解

卷3：T/TCP、HTTP、NNTP和UNIX域协议 英文版

TCP/IP Illustrated

Volume 3: TCP for Transactions, HTTP, NNTP, and the UNIX Domain Protocols

[美] W. Richard Stevens 著

人民邮电出版社
POSTS & TELECOM PRESS

TCP/IP详解

卷3：T/TCP、HTTP、NNTP和UNIX域协议 英文版

TCP/IP Illustrated

Transactions, HTTP, NNTP, UNIX Domain Protocols

[美] W. Richard Stevens 著

人民邮电出版社
北京

图书在版编目（CIP）数据

TCP/IP详解. 卷3，T/TCP、HTTP、NNTP和UNIX域协议
=TCP/IP Illustrated, Volume 3: TCP for
Transactions, HTTP, NNTP, and the UNIX Domain
Protocols，First Edition：英文/（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 978-7-115-22254-1

I. ①T… II. ①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017228号

内 容 提 要

本书是TCP/IP领域的经典之作！书中重点讲述高级协议，覆盖了当今TCP/IP编程人员和网络管理员必须熟练掌握的T/TCP（TCP事务协议）、HTTP（超文本传送协议）、NNTP（网络新闻传送协议）和Unix域协议。与前面两卷一样，本书有丰富的例子和实现的细节。

本书适合希望了解TCP/IP协议如何实现的读者阅读，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3：T/TCP、HTTP、NNTP和UNIX域协议（英文版）

◆ 著 [美] 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：21.75
字数：418千字 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：01-2010-0312号
ISBN 978-7-115-22254-1

定价：59.00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

前　言

概述及本书的结构

本书是《TCP/IP详解》系列书的自然延续：[Stevens, 1994]，本书中称为卷1（Volume 1）；[Wright and Stevens, 1995]，本书中称为卷2（Volume 2）。本书可以分为三部分，每一部分包含一个不同的主题。

(1) TCP事务协议，一般简称T/TCP。这是TCP的扩展，用来使客户－服务器事务更快、更有效，同时也更加可靠。这是通过省略连接开始时的三次握手并缩短连接结束时的TIME_WAIT状态来实现的。我们将看到，对于客户－服务器事务，T/TCP可以达到UDP的性能，而T/TCP还提供了可靠性和适应性，这是与UDP相比的重要改进。

事务可以定义为客户向服务器提出的请求以及服务器相应的应答。（术语“事务”指的不是包含加锁、两段提交和回退过程的数据库事务。）

(2) TCP/IP应用具体是指HTTP（超文本传送协议，万维网的基础）和NNTP（网络新闻传送协议，Usenet新闻系统的基础）。

(3) Unix域协议。所有的Unix TCP/IP实现都提供这些协议，许多非Unix实现也提供这些协议。它们提供了一种进程间通信（IPC）的形式，并使用与TCP/IP一样的套接字接口。当客户和服务器在同一台主机上时，Unix域协议的速度一般是TCP/IP的两倍。

第一部分（T/TCP的描述）分为两块内容。第1章至第4章对这一协议进行了描述，并提供大量的示例说明其工作原理。卷1的24.7节曾对T/TCP进行了简单描述，本书的这部分内容对其进行了大幅扩展。第二块是第5章至第12章，描述的是4.4BSD-Lite网络代码（即卷2给出的代码）中T/TCP的实际实现。由于第一个T/TCP实现直到1994年9月才发布，而此时卷1已经出版一年，卷2也基本完成，因此T/TCP的示例和实现细节只能在本套书的这一卷中进行详细描述。

第二部分（HTTP和NNTP应用）是卷1的第25章至第30章介绍的TCP/IP应用的延续。在卷1出版后两年的时间里，HTTP技术随着因特网的兴起迅速流行开来，NNTP技术的使用在十几年时间中每年增长75%左右。由于常见的TCP使用方式是在数据交换极少的短连接里（连接的建立和销毁操作占用大部分时间），因此HTTP还是T/TCP的理想候补技术。在繁忙的Web服务器上由数以千计不同类型的客户大量使用HTTP（进而大量使用TCP）使我们可以检测服务器上的实际分组（第14章），并更好地理解卷1和卷2中描述的很多TCP/IP特性。

第三部分的Unix域协议本来是计划安排在卷2中的，但是由于卷2的篇幅已达到1200页，所以删掉了。在题为《TCP/IP详解》的一套书中讲述非TCP/IP协议看上去有点奇怪，但是Unix域协议早在将近15年前的4.2BSD版本中就首次实现了，与BSD TCP/IP的首次实现时间差不多。Berkeley衍生内核中大量使用了Unix域协议，但通常都是“在掩护下”使用的，大多数用户感觉不到它们的存在。除了作为Berkeley衍生内核中Unix管道的基础技术外，Unix域协议还大量用于客户和服务器在同一台主机（常见的工作站）上的X Window 系统。Unix域套接字技术用于在进程之间传递描述符，这是一种用于进程间通信的强大技术。由于Unix域协议中套接字API（应用程序接口）与TCP/IP中的套接字API几乎相同，因此只需要改动很少的代码，Unix域协议就可以轻松地提高应用程序的性能。

以上三部分内容可以独立阅读。

致读者

与前两卷相似，本卷面向所有希望了解TCP/IP协议运行原理的读者：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那些需要每天与TCP/IP应用打交道的用户。

前两部分内容要求读者对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首先应参考卷1[Stevens, 1994]，该书对TCP/IP协议族有比较透彻的讲述。第一部分的前一块内容（第1章至第4章，T/TCP基本概念及示例）可以独立于卷2阅读，但其余内容（第5~12章，T/TCP的实现）要求读者对卷2中提供的4.4BSD-Lite网络代码比较熟悉。

本书贯穿了一些交叉引用，不仅参考了本卷中的内容，还参考了卷1和卷2中相应的章节。本书提供了完整的索引，并把用到的所有缩略词及相应的复合术语都详细列在本

书的最前面。索引后还按照字母表顺序给出了书中所用到的结构体、函数和宏的交叉引用，以及相关详细信息的起始页码。当本卷的代码需要引用卷2中的内容时，交叉引用也会提及卷2中的相关定义。

源代码版权

本书中所有来自4.4BSD-Lite发布版本的源代码都包含如下的版权说明：

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

第6章的路由表代码包含如下的版权说明：

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

排版约定

在展示交互式的输入和输出时，我们用粗体显示键入内容，以等宽正体显示计算机的输出，以斜体显示注释，示例如下：

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

另外，我们将系统名（本例中是sun）作为shell提示符的一部分，以表明命令正在哪种主机上运行。正文中提到的程序的名字通常用首字母大写（如Telnet和Tcpdump）以避免过多的字体变化。

整本书中，我们随时会插入缩进的小字号段落来描述历史问题或实现细节。

致谢

首先我要感谢我的家人Sally、Bill、Ellen和David。在过去的一年中，他们又一次忍受了我外出旅行完成这本书的过程。不过，这一次做的确实是一本“小型”书。

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：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和Gary Wright。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。

特别感谢Vern Paxson和Andras Olah，他们对整部书稿进行了不可思议的细致审查，发现了许多错误，并提出了有价值的技术性建议。还要感谢Vern Paxson把他的软件提供给我来分析Tcpdump跟踪文件，感谢Andras Olah在过去一年中在T/TCP方面给予我的帮助。同样感谢T/TCP的设计者Bob Braden，他提供了参考源代码实现，这是本书第一部分

的基础。

还有一些人也提供了很重要的帮助。Gary Wright和Jim Hogue提供了第14章中采集数据所需要的系统。Doug Schmidt为第16章的时间度量提供了使用Unix域套接字的公共域TTCP程序的副本。Craig Partridge提供了一份RDP源代码的副本来帮助测试。Mike Karels解答了很多问题。

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。

最后，我要感谢Addison-Wesley公司的所有员工，特别是本书的编辑John Wait，感谢你们多年来的帮助。

跟以前一样，作者用James Clark编写的Groff包制作了本书的最终电子版——Troff硬拷贝。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。

W. Richard Stevens

1995年11月于亚利桑那州图森市

Contents

Part 1

TCP for Transactions

1

T/TCP Introduction

1.1 Introduction

This chapter introduces the concepts of a client–server transaction. We start with a UDP client–server application, the simplest possible. We then write the client and server using TCP and examine the resulting TCP/IP packets that are exchanged between the two hosts. Next we use T/TCP, showing the reduction in packets and the minimal source code changes required on both ends to take advantage of T/TCP.

We then introduce the test network used to run the examples in the text, and look at a simple timing comparison between the UDP, TCP, and T/TCP client–server applications. We look at some typical Internet applications that use TCP and see what would change if the two end systems supported T/TCP. This is followed by a brief history of transaction processing protocols within the Internet protocol suite, and a description of existing T/TCP implementations.

Throughout this text and throughout the T/TCP literature, the term *transaction* means a request sent by a client to a server along with the server's reply. A common Internet example is a client request to a Domain Name System (DNS) server, asking for the IP address corresponding to a domain name, followed by the server's response. We do *not* use the term to imply the semantics often associated with database transactions: locking, two-phase commit, backout, and so on.

1.2 UDP Client–Server

We begin with a simple UDP client–server example, showing the client source code in Figure 1.1. The client sends a request to the server, the server processes the request and sends back a reply.