

# TCP/IP Illustrated, Volume 3

---

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

W. Richard Stevens



ADDISON-WESLEY

Boston • San Francisco • New York • Toronto • Montreal  
London • Munich • Paris • Madrid  
Capetown • Sydney • Tokyo • Singapore • Mexico City

# Contents

|                                      |           |
|--------------------------------------|-----------|
| <b>Preface</b>                       | <b>xv</b> |
| <b>Part 1. TCP for Transactions</b>  | <b>1</b>  |
| <b>Chapter 1. T/TCP Introduction</b> | <b>3</b>  |
| 1.1 Introduction                     | 3         |
| 1.2 UDP Client-Server                | 3         |
| 1.3 TCP Client-Server                | 9         |
| 1.4 T/TCP Client-Server              | 17        |
| 1.5 Test Network                     | 20        |
| 1.6 Timing Example                   | 21        |
| 1.7 Applications                     | 22        |
| 1.8 History                          | 24        |
| 1.9 Implementations                  | 26        |
| 1.10 Summary                         | 28        |
| <b>Chapter 2. T/TCP Protocol</b>     | <b>29</b> |
| 2.1 Introduction                     | 29        |
| 2.2 New TCP Options for T/TCP        | 30        |
| 2.3 T/TCP Implementation Variables   | 33        |
| 2.4 State Transition Diagram         | 34        |
| 2.5 T/TCP Extended States            | 36        |
| 2.6 Summary                          | 38        |

|                   |                                                      |           |
|-------------------|------------------------------------------------------|-----------|
| <b>Chapter 3.</b> | <b>T/TCP Examples</b>                                | <b>39</b> |
| 3.1               | Introduction                                         | 39        |
| 3.2               | Client Reboot                                        | 40        |
| 3.3               | Normal T/TCP Transaction                             | 42        |
| 3.4               | Server Receives Old Duplicate SYN                    | 43        |
| 3.5               | Server Reboot                                        | 44        |
| 3.6               | Request or Reply Exceeds MSS                         | 45        |
| 3.7               | Backward Compatibility                               | 49        |
| 3.8               | Summary                                              | 51        |
| <b>Chapter 4.</b> | <b>T/TCP Protocol (Continued)</b>                    | <b>53</b> |
| 4.1               | Introduction                                         | 53        |
| 4.2               | Client Port Numbers and TIME_WAIT State              | 53        |
| 4.3               | Purpose of the TIME_WAIT State                       | 56        |
| 4.4               | TIME_WAIT State Truncation                           | 59        |
| 4.5               | Avoiding the Three-Way Handshake with TAO            | 62        |
| 4.6               | Summary                                              | 68        |
| <b>Chapter 5.</b> | <b>T/TCP Implementation: Socket Layer</b>            | <b>69</b> |
| 5.1               | Introduction                                         | 69        |
| 5.2               | Constants                                            | 70        |
| 5.3               | send Function                                        | 70        |
| 5.4               | Summary                                              | 72        |
| <b>Chapter 6.</b> | <b>T/TCP Implementation: Routing Table</b>           | <b>73</b> |
| 6.1               | Introduction                                         | 73        |
| 6.2               | Code Introduction                                    | 74        |
| 6.3               | radix_node_head Structure                            | 75        |
| 6.4               | rtentry Structure                                    | 75        |
| 6.5               | rt_metrics Structure                                 | 76        |
| 6.6               | in_inithead Function                                 | 76        |
| 6.7               | in_addroute Function                                 | 77        |
| 6.8               | in_rtroute Function                                  | 78        |
| 6.9               | in_clrroute Function                                 | 78        |
| 6.10              | in_rtqtime Function                                  | 79        |
| 6.11              | in_rtqkill Function                                  | 82        |
| 6.12              | Summary                                              | 85        |
| <b>Chapter 7.</b> | <b>T/TCP Implementation: Protocol Control Blocks</b> | <b>87</b> |
| 7.1               | Introduction                                         | 87        |
| 7.2               | in_pcbaddr Function                                  | 88        |
| 7.3               | in_pcbconnect Function                               | 89        |
| 7.4               | Summary                                              | 90        |
| <b>Chapter 8.</b> | <b>T/TCP Implementation: TCP Overview</b>            | <b>91</b> |
| 8.1               | Introduction                                         | 91        |
| 8.2               | Code Introduction                                    | 91        |

|                    |                                                    |            |
|--------------------|----------------------------------------------------|------------|
| 8.3                | TCP protocol Structure                             | 92         |
| 8.4                | TCP Control Block                                  | 93         |
| 8.5                | tcp_init Function                                  | 94         |
| 8.6                | tcp_slowtime Function                              | 94         |
| 8.7                | Summary                                            | 95         |
| <b>Chapter 9.</b>  | <b>T/TCP Implementation: TCP Output</b>            | <b>97</b>  |
| 9.1                | Introduction                                       | 97         |
| 9.2                | tcp_output Function                                | 97         |
| 9.3                | Summary                                            | 104        |
| <b>Chapter 10.</b> | <b>T/TCP Implementation: TCP Functions</b>         | <b>105</b> |
| 10.1               | Introduction                                       | 105        |
| 10.2               | tcp_newtcpcb Function                              | 105        |
| 10.3               | tcp_rtllookup Function                             | 106        |
| 10.4               | tcp_gettaocache Function                           | 108        |
| 10.5               | Retransmission Timeout Calculations                | 108        |
| 10.6               | tcp_close Function                                 | 112        |
| 10.7               | tcp_rssend Function                                | 113        |
| 10.8               | tcp_rssrcvd Function                               | 114        |
| 10.9               | tcp_dooptions Function                             | 121        |
| 10.10              | tcp_rss Function                                   | 122        |
| 10.11              | Summary                                            | 124        |
| <b>Chapter 11.</b> | <b>T/TCP Implementation: TCP Input</b>             | <b>125</b> |
| 11.1               | Introduction                                       | 125        |
| 11.2               | Preliminary Processing                             | 125        |
| 11.3               | Header Prediction                                  | 129        |
| 11.4               | Initiation of Passive Open                         | 130        |
| 11.5               | Initiation of Active Open                          | 134        |
| 11.6               | PAWS: Protection Against Wrapped Sequence Numbers  | 141        |
| 11.7               | ACK Processing                                     | 142        |
| 11.8               | Completion of Passive Opens and Simultaneous Opens | 142        |
| 11.9               | ACK Processing (Continued)                         | 143        |
| 11.10              | FIN Processing                                     | 145        |
| 11.11              | Summary                                            | 147        |
| <b>Chapter 12.</b> | <b>T/TCP Implementation: TCP User Requests</b>     | <b>149</b> |
| 12.1               | Introduction                                       | 149        |
| 12.2               | PRU_CONNECT Request                                | 149        |
| 12.3               | tcp_connect Function                               | 150        |
| 12.4               | PRU_SEND and PRU_SEND_EOF Requests                 | 154        |
| 12.5               | tcp_usrclosed Function                             | 155        |
| 12.6               | tcp_sysctl Function                                | 155        |
| 12.7               | T/TCP Futures                                      | 156        |
| 12.8               | Summary                                            | 158        |

---

**Part 2. Additional TCP Applications** **159**


---

|                    |                                             |            |
|--------------------|---------------------------------------------|------------|
| <b>Chapter 13.</b> | <b>HTTP: Hypertext Transfer Protocol</b>    | <b>161</b> |
| 13.1               | Introduction                                | 161        |
| 13.2               | Introduction to HTTP and HTML               | 162        |
| 13.3               | HTTP Protocol                               | 165        |
| 13.4               | An Example                                  | 170        |
| 13.5               | HTTP Statistics                             | 172        |
| 13.6               | Performance Problems                        | 173        |
| 13.7               | Summary                                     | 175        |
| <br>               |                                             |            |
| <b>Chapter 14.</b> | <b>Packets Found on an HTTP Server</b>      | <b>177</b> |
| 14.1               | Introduction                                | 177        |
| 14.2               | Multiple HTTP Servers                       | 180        |
| 14.3               | Client SYN Interarrival Time                | 181        |
| 14.4               | RTT Measurements                            | 185        |
| 14.5               | Listen Backlog Queue                        | 187        |
| 14.6               | Client SYN Options                          | 192        |
| 14.7               | Client SYN Retransmissions                  | 195        |
| 14.8               | Domain Names                                | 196        |
| 14.9               | Timing Out Persist Probes                   | 196        |
| 14.10              | Simulation of T/TCP Routing Table Size      | 200        |
| 14.11              | Mbuf Interaction                            | 202        |
| 14.12              | TCP PCB Cache and Header Prediction         | 203        |
| 14.13              | Summary                                     | 205        |
| <br>               |                                             |            |
| <b>Chapter 15.</b> | <b>NNTP: Network News Transfer Protocol</b> | <b>207</b> |
| 15.1               | Introduction                                | 207        |
| 15.2               | NNTP Protocol                               | 209        |
| 15.3               | A Simple News Client                        | 212        |
| 15.4               | A More Sophisticated News Client            | 214        |
| 15.5               | NNTP Statistics                             | 215        |
| 15.6               | Summary                                     | 216        |

---

**Part 3. The Unix Domain Protocols** **219**


---

|                    |                                            |            |
|--------------------|--------------------------------------------|------------|
| <b>Chapter 16.</b> | <b>Unix Domain Protocols: Introduction</b> | <b>221</b> |
| 16.1               | Introduction                               | 221        |
| 16.2               | Usage                                      | 222        |
| 16.3               | Performance                                | 223        |
| 16.4               | Coding Examples                            | 224        |
| 16.5               | Summary                                    | 225        |

|                     |                                                          |            |
|---------------------|----------------------------------------------------------|------------|
| <b>Chapter 17.</b>  | <b>Unix Domain Protocols: Implementation</b>             | <b>227</b> |
| 17.1                | Introduction                                             | 227        |
| 17.2                | Code Introduction                                        | 227        |
| 17.3                | Unix domain and protocol Structures                      | 228        |
| 17.4                | Unix Domain Socket Address Structures                    | 230        |
| 17.5                | Unix Domain Protocol Control Blocks                      | 231        |
| 17.6                | uipc_usrreq Function                                     | 233        |
| 17.7                | PRU_ATTACH Request and unpr_attach Function              | 233        |
| 17.8                | PRU_DETACH Request and unpr_detach Function              | 236        |
| 17.9                | PRU_BIND Request and unpr_bind Function                  | 237        |
| 17.10               | PRU_CONNECT Request and unpr_connect Function            | 240        |
| 17.11               | PRU_CONNECT2 Request and unpr_connect2 Function          | 245        |
| 17.12               | socketpair System Call                                   | 249        |
| 17.13               | pipe System Call                                         | 253        |
| 17.14               | PRU_ACCEPT Request                                       | 253        |
| 17.15               | PRU_DISCONNECT Request and unpr_disconnect Function      | 255        |
| 17.16               | PRU_SHUTDOWN Request and unpr_shutdown Function          | 257        |
| 17.17               | PRU_ABORT Request and unpr_drop Function                 | 258        |
| 17.18               | Miscellaneous Requests                                   | 259        |
| 17.19               | Summary                                                  | 261        |
| <b>Chapter 18.</b>  | <b>Unix Domain Protocols: I/O and Descriptor Passing</b> | <b>263</b> |
| 18.1                | Introduction                                             | 263        |
| 18.2                | PRU_SEND and PRU_RECV Requests                           | 263        |
| 18.3                | Descriptor Passing                                       | 269        |
| 18.4                | unpr_internalize Function                                | 274        |
| 18.5                | unpr_externalize Function                                | 276        |
| 18.6                | unpr_discard Function                                    | 277        |
| 18.7                | unpr_dispose Function                                    | 278        |
| 18.8                | unpr_scan Function                                       | 278        |
| 18.9                | unpr_gc Function                                         | 280        |
| 18.10               | unpr_mark Function                                       | 288        |
| 18.11               | Performance (Revisited)                                  | 288        |
| 18.12               | Summary                                                  | 289        |
| <b>Appendix A.</b>  | <b>Measuring Network Times</b>                           | <b>291</b> |
| A.1                 | RTT Measurements Using Ping                              | 292        |
| A.2                 | Protocol Stack Measurements                              | 294        |
| A.3                 | Latency and Bandwidth                                    | 300        |
| <b>Appendix B.</b>  | <b>Coding Applications for T/TCP</b>                     | <b>303</b> |
| <b>Bibliography</b> |                                                          | <b>309</b> |
| <b>Index</b>        |                                                          | <b>315</b> |