

TCP/IP Illustrated, Volume 2

The Implementation

Gary R. Wright
W. Richard Stevens

ADDISON-WESLEY

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

Contents

Preface	xix
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

2.7	Summary of Mbuf Macros and Functions	51	
2.8	Summary of Net/3 Networking Data Structures	54	
2.9	n_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	protocol Structure	188
7.5	IP domain and protocol 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_insecoptions Function	265
9.9	ip_pchopts 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 protocol Structure	308
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	ip_multi Structure	345
12.7	ip_options Structure	347
12.8	Multicast Socket Options	348
12.9	Multicast TTL Values	348
12.10	ip_setoptions Function	351
12.11	Joining an IP Multicast Group	355
12.12	Leaving an IP Multicast Group	366
12.13	ip_getoptions 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 protocol Structure	384
13.5	Joining a Group: igmp_joingroup Function	386
13.6	igmp_fasttime 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	mouted Daemon	401
14.5	Virtual Interfaces	404
14.6	IGMP Revisited	411
14.7	Multicast Routing	416
14.8	Multicast Forwarding: <code>ip_mforward</code> Function	424
14.9	Cleanup: <code>ip_router_gone</code> 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	461
15.8	bind System Call	463
15.9	listen System Call	466
15.10	sleep and wakeup Functions	466
15.11	accept System Call	467
15.12	connect and disconnect Functions	481
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	osend 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	osrecv Function	505
16.12	osrecv 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	567
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	RTPREP 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_rtmmsg 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 protoaw Structures	646
20.3	Routing Control Blocks	647

20.4	<code>raw_init</code> Function	647	
20.5	<code>route_output</code> Function	648	
20.6	<code>rt_xaddrs</code> Function	660	
20.7	<code>rt_asmetrics</code> Function	661	
20.8	<code>raw_input</code> Function	662	
20.9	<code>route_uerror</code> Function	664	
20.10	<code>raw_uerror</code> Function	666	
20.11	<code>raw_attach</code> , <code>raw_detach</code> , and <code>raw_disconnect</code> 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	<code>arpwhohas</code> Function	683	
21.6	<code>arprequest</code> Function	684	
21.7	<code>arpinr</code> Function	687	
21.8	<code>in_arpinput</code> Function	688	
21.9	ARP Timer Functions	694	
21.10	<code>arpresolve</code> Function	696	
21.11	<code>arplookup</code> Function	701	
21.12	Proxy ARP	703	
21.13	<code>arp_rrequest</code> 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	<code>inpcb</code> Structure	716	
22.4	<code>in_pcballoc</code> and <code>in_pcbdetach</code> Functions	717	
22.5	Binding, Connecting, and Demultiplexing	719	
22.6	<code>in_pcblookup</code> Function	724	
22.7	<code>in_pcbbind</code> Function	728	
22.8	<code>in_pcbconnect</code> Function	735	
22.9	<code>in_pcbdisconnect</code> Function	741	
22.10	<code>in_getsockaddr</code> and <code>in_setpeeraddr</code> Functions	741	
22.11	<code>in_pcbnotify</code> , <code>in_rchange</code> , and <code>in_listen</code> 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 protocol 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_fasttime Function	821
25.5	tcp_slowtime 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

Chapter 27.	TCP Functions	891
27.1	Introduction	891
27.2	tcp_drain Function	892
27.3	tcp_drop Function	892
27.4	tcp_close Function	893
27.5	tcp_rss Function	897
27.6	tcp_ctlinput Function	904
27.7	tcp_notify Function	904
27.8	tcp_quench Function	906
27.9	TCP_REASS Macro and tcp_reass Function	906
27.10	tcp_trace Function	915
27.11	Summary	920
Chapter 28.	TCP Input	923
28.1	Introduction	923
28.2	Preliminary Processing	925
28.3	tcp_dooptions Function	933
28.4	Header Prediction	934
28.5	TCP Input: Slow Path Processing	941
28.6	Initiation of Passive Open, Completion of Active Open	942
28.7	PAWS: Protection Against Wrapped Sequence Numbers	951
28.8	Trim Segment so Data is Within Window	954
28.9	Self-Connects and Simultaneous Opens	960
28.10	Record Timestamp	963
28.11	RST Processing	963
28.12	Summary	965
Chapter 29.	TCP Input (Continued)	967
29.1	Introduction	967
29.2	ACK Processing Overview	967
29.3	Completion of Passive Opens and Simultaneous Opens	967
29.4	Fast Retransmit and Fast Recovery Algorithms	970
29.5	ACK Processing	974
29.6	Update Window Information	981
29.7	Urgent Mode Processing	983
29.8	tcp_pulloutofband Function	986
29.9	Processing of Received Data	988
29.10	FIN Processing	990
29.11	Final Processing	992
29.12	Implementation Refinements	994
29.13	Header Compression	995
29.14	Summary	1004
Chapter 30.	TCP User Requests	1007
30.1	Introduction	1007
30.2	tcp_usrreq Function	1007
30.3	tcp_attach Function	1018
30.4	tcp_disconnect Function	1019

30.5	<code>tcp_unreleased</code> Function	1021
30.6	<code>tcp_ctloutput</code> Function	1022
30.7	Summary	1025
Chapter 31.	BPF: BSD Packet Filter	1027
31.1	Introduction	1027
31.2	Code Introduction	1028
31.3	<code>bpf_if</code> Structure	1029
31.4	<code>bpf_d</code> Structure	1032
31.5	BPF Input	1040
31.6	BPF Output	1046
31.7	Summary	1047
Chapter 32.	Raw IP	1049
32.1	Introduction	1049
32.2	Code Introduction	1050
32.3	Raw IP protocol Structure	1051
32.4	<code>rip_init</code> Function	1053
32.5	<code>rip_input</code> Function	1053
32.6	<code>rip_output</code> Function	1056
32.7	<code>rip_unreleg</code> Function	1058
32.8	<code>rip_ctloutput</code> Function	1063
32.9	Summary	1065
Epilogue		1067
Appendix A.	Solutions to Selected Exercises	1069
Appendix B.	Source Code Availability	1093
Appendix C.	RFC 1122 Compliance	1097
C.1	Link-Layer Requirements	1097
C.2	IP Requirements	1098
C.3	IP Options Requirements	1102
C.4	IP Fragmentation and Reassembly Requirements	1104
C.5	ICMP Requirements	1105
C.6	Multicasting Requirements	1110
C.7	IGMP Requirements	1111
C.8	Routing Requirements	1111
C.9	ARP Requirements	1113
C.10	UDP Requirements	1113
C.11	TCP Requirements	1115
Bibliography		1125
Index		1133