

Understanding Software Dynamics

Richard L. Sites

◆◆Addison-Wesley

Boston • Columbus • New York • San Francisco • Amsterdam • Cape Town
Dubai • London • Madrid • Milan • Munich • Paris • Montreal • Toronto • Delhi • Mexico City
São Paulo • Sydney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

Contents

Foreword	xix
Preface	xxi
Acknowledgments	xxv
About the Author	xxvii

I Measurement 1

1 My Program Is Too Slow 3

1.1 Datacenter Context	3
1.2 Datacenter Hardware	5
1.3 Datacenter Software	6
1.4 Long-Tail Latency	7
1.5 Thought Framework	9
1.6 Order-of-Magnitude Estimates	9
1.7 Why Are Transactions Slow?	11
1.8 The Five Fundamental Resources	12
1.9 Summary	12

2 Measuring CPUs 15

2.1 How We Got Here	15
2.2 Where Are We Now?	19
2.3 Measuring the Latency of an add Instruction	20
2.4 Straight-Line Code Fail	21
2.5 Simple Loop, Loop Overhead Fail, Optimizing Compiler Fail	21
2.6 Dead Variable Fail	24
2.7 Better Loop	25
2.8 Dependent Variables	26
2.9 Actual Execution Latency	26
2.10 More Nuance	27
2.11 Summary	28
Exercises	28

3 Measuring Memory 31

3.1 Memory Timing	31
3.2 About Memory	32
3.3 Cache Organization	34
3.4 Data Alignment	36
3.5 Translation Lookaside Buffer Organization	36

3.6	The Measurements	37
3.7	Measuring Cache Line Size	38
3.8	Problem: N+1 Prefetching	40
3.9	Dependent Loads	41
3.10	Non-random Dynamic Random-Access Memory	42
3.11	Measuring Total Size of Each Cache Level	43
3.12	Measuring Cache Associativity of Each Level	45
3.13	Translation Buffer Time	46
3.14	Cache Underutilization	46
3.15	Summary	46
	Exercises	47
4	CPU and Memory Interaction	49
4.1	Cache Interaction	49
4.2	Simple Matrix Multiply Dynamics	51
4.3	Estimates	51
4.4	Initialization, Cross-Checking, and Observing	52
4.5	Initial Results	53
4.6	Faster Matrix Multiply, Transpose Method	55
4.7	Faster Matrix Multiply, Subblock Method	57
4.8	Cache-Aware Computation	58
4.9	Summary	58
	Exercises	59
5	Measuring Disk/SSD	61
5.1	About Hard Disks	62
5.2	About SSDs	64
5.3	Software Disk Access and On-Disk Buffering	66
5.4	How Fast Is a Disk Read?	68
5.5	A Little Back-of-the-Envelope Calculation	71
5.6	How Fast Is a Disk Write?	72
5.7	Results	73
5.8	Reading from Disk	73
5.9	Writing to Disk	77
5.10	Reading from SSD	80
5.11	Writing to SSD	82
5.12	Multiple Transfers	82
5.13	Summary	83
	Exercises	84

6	Measuring Networks	85
6.1	About Ethernet	87
6.2	About Hubs, Switches, and Routers	89
6.3	About TCP/IP	89
6.4	About Packets	90
6.5	About Remote Procedure Calls (RPCs)	91
6.6	Slop	93
6.7	Observing Network Traffic	94
6.8	Sample RPC Message Definition	96
6.9	Sample Logging Design	99
6.10	Sample Client-Server System Using RPCs	100
6.11	Sample Server Program	101
6.12	Spinlocks	101
6.13	Sample Client Program	102
6.14	Measuring One Sample Client-Server RPC	105
6.15	Postprocessing RPC Logs	106
6.16	Observations	107
6.17	Summary	108
	Exercises	109
7	Disk and Network Database Interaction	111
7.1	Time Alignment	111
7.2	Multiple Clients	117
7.3	Spinlocks	118
7.4	Experiment 1	118
7.5	On-Disk Database	121
7.6	Experiment 2	121
7.7	Experiment 3	125
7.8	Logging	127
7.9	Understanding Transaction Latency Variation	128
7.10	Summary	128
	Exercises	129
II	Observation	131
8	Logging	133
8.1	Observation Tools	133
8.2	Logging	133
8.3	Basic Logging	134

8.4	Extended Logging	135
8.5	Timestamps	135
8.6	RPC IDs	136
8.7	Log File Formats	137
8.8	Managing Log Files	138
8.9	Summary	139
9	Aggregate Measures	141
9.1	Uniform vs. Bursty Event Rates	142
9.2	Measurement Intervals	143
9.3	Timelines	143
9.4	Further Summarizing of Timelines	145
9.5	Histogram Time Scales	147
9.6	Aggregating Per-Event Measurements	150
9.7	Patterns of Values Over Time	151
9.8	Update Intervals	152
9.9	Example Transactions	154
9.10	Conclusion	155
10	Dashboards	157
10.1	Sample Service	157
10.2	Sample Dashboards	159
10.3	Master Dashboard	159
10.4	Per-Instance Dashboards	163
10.5	Per-Server Dashboards	164
10.6	Sanity Checks	164
10.7	Summary	165
	Exercises	165
11	Other Existing Tools	167
11.1	Kinds of Observation Tools	167
11.2	Data to Observe	169
11.3	<code>top</code> Command	170
11.4	<code>/proc</code> and <code>/sys</code> Pseudofiles	171
11.5	<code>time</code> Command	171
11.6	<code>perf</code> Command	171
11.7	<code>oprofile</code> , CPU Profiler	173
11.8	<code>strace</code> , System Calls	176
11.9	<code>ltrace</code> , CPU C Library Calls	179

- 11.10 `ftrace`, CPU Trace 180
- 11.11 `mtrace`, Memory Malloc/Free 183
- 11.12 `blktrace`, Disk Trace 184
- 11.13 `tcpdump` and Wireshark, Network Trace 187
- 11.14 `locktrace`, Critical Section Locks 189
- 11.15 Offered Load, Outbound Calls, and Transaction Latency 189
- 11.16 Summary 191
- Exercises 191

12 Traces 193

- 12.1 Tracing Advantages 193
- 12.2 Tracing Disadvantages 194
- 12.3 The Three Starting Questions 194
- 12.4 Example: Early Program Counter Trace 197
- 12.5 Example: Per-Function Counts and Time 199
- 12.6 Case Study: Per-Function Trace of Gmail 203
- 12.7 Summary 207

13 Observation Tool Design Principles 209

- 13.1 What to Observe 209
- 13.2 How Frequently and For How Long? 210
- 13.3 How Much Overhead? 211
- 13.4 Design Consequences 212
- 13.5 Case Study: Histogram Buckets 212
- 13.6 Designing Data Display 214
- 13.7 Summary 215

III Kernel-User Trace 217

14 KUtrace: Goals, Design, Implementation 219

- 14.1 Overview 219
- 14.2 Goals 220
- 14.3 Design 221
- 14.4 Implementation 223
- 14.5 Kernel Patches and Module 224
- 14.6 Control Program 224
- 14.7 Postprocessing 225
- 14.8 A Note on Security 225
- 14.9 Summary 225

- 15 KUtrace: Linux Kernel Patches 227**
 - 15.1 Trace Buffer Data Structures 228
 - 15.2 Raw Traceblock Format 229
 - 15.3 Trace Entries 230
 - 15.4 IPC Trace Entries 232
 - 15.5 Timestamps 233
 - 15.6 Event Numbers 233
 - 15.7 Nested Trace Entries 233
 - 15.8 Code 234
 - 15.9 Packet Tracing 234
 - 15.10 AMD/Intel x86-64 Patches 236
 - 15.11 Summary 237
 - Exercises 237

- 16 KUtrace: Linux Loadable Module 239**
 - 16.1 Kernel Interface Data Structures 239
 - 16.2 Module Load/Unload 240
 - 16.3 Initializing and Controlling Tracing 241
 - 16.4 Implementing Trace Calls 241
 - 16.5 Insert1 241
 - 16.6 InsertN 243
 - 16.7 Switching to a New Traceblock 244
 - 16.8 Summary 244

- 17 KUtrace: User-Mode Runtime Control 245**
 - 17.1 Controlling Tracing 245
 - 17.2 Standalone kustrace_control Program 246
 - 17.3 The Underlying kustrace_lib Library 246
 - 17.4 The Control Interface to the Loadable Module 247
 - 17.5 Summary 247

- 18 KUtrace: Postprocessing 249**
 - 18.1 Postprocessing Details 249
 - 18.2 The rawtoevent Program 250
 - 18.3 The eventtospan Program 251
 - 18.4 The spantotrim Program 253
 - 18.5 The spantospan Program 253
 - 18.6 The samptoname_k and samptoname_u Programs 253
 - 18.7 The makeself Program 254

18.8 KUtrace JSON Format 254

18.9 Summary 256

19 KUtrace: Display of Software Dynamics 257

19.1 Overview 257

19.2 Region 1, Controls 258

19.3 Region 2, Y-axis 259

19.4 Region 3, Timelines 260

19.5 Region 4, IPC Legend 265

19.6 Region 5, X-axis 265

19.7 Region 6, Save/Restore 265

19.8 Secondary Controls 265

19.9 Summary 266

IV Reasoning 267

20 What to Look For 269

20.1 Overview 269

21 Executing Too Much 271

21.1 Overview 271

21.2 The Program 271

21.3 The Mystery 272

21.4 Exploring and Reasoning 273

21.5 Mystery Understood 277

21.6 Summary 277

22 Executing Slowly 279

22.1 Overview 279

22.2 The Program 279

22.3 The Mystery 280

22.4 Floating-Point Antagonist 282

22.5 Memory Antagonist 285

22.6 Mystery Understood 286

22.7 Summary 286

23 Waiting for CPU 289

23.1 The Program 289

23.2 The Mystery 289

23.3 Exploring and Reasoning 290

23.4 Mystery 2 292

23.5	Mystery 2 Understood	293
23.6	Bonus Mystery	295
23.7	Summary	297
	Exercises	297
24	Waiting for Memory	299
24.1	The Program	299
24.2	The Mystery	300
24.3	Exploring and Reasoning	300
24.4	Mystery 2: Access to a Page Table	304
24.5	Mystery 2 Understood	304
24.6	Summary	306
	Exercises	306
25	Waiting for Disk	307
25.1	The Program	307
25.2	The Mystery	307
25.3	Exploring and Reasoning	308
25.4	Reading 40MB	310
25.5	Reading Sequential 4KB Blocks	311
25.6	Reading Random 4KB Blocks	313
25.7	Writing and Sync of 40MB on SSD	314
25.8	Reading 40MB on SSD	315
25.9	Two Programs Accessing Two Files at Once	316
25.10	Mysteries Understood	317
25.11	Summary	317
	Exercises	317
26	Waiting for Network	319
26.1	Overview	319
26.2	The Programs	320
26.3	Experiment 1	321
26.4	Experiment 1 Mystery	322
26.5	Experiment 1 Exploring and Reasoning	323
26.6	Experiment 1 What About the Time Between RPCs?	327
26.7	Experiment 2	329
26.8	Experiment 3	329
26.9	Experiment 4	330
26.10	Mysteries Understood	333

26.11	Bonus Anomaly	334
26.12	Summary	336
27	Waiting for Locks	337
27.1	Overview	337
27.2	The Program	341
27.3	Experiment 1: Long Lock Hold Times	344
27.3.1	Simple Locking	344
27.3.2	Lock Saturation	345
27.4	Mysteries in Experiment 1	345
27.5	Exploring and Reasoning in Experiment 1	346
27.5.1	Lock Capture	347
27.5.2	Lock Starvation	348
27.6	Experiment 2: Fixing Lock Capture	348
27.7	Experiment 3: Fixing Lock Contention via Multiple Locks	349
27.8	Experiment 4: Fixing Lock Contention via Less Locked Work	351
27.9	Experiment 5: Fixing Lock Contention via RCU for Dashboard	353
27.10	Summary	355
28	Waiting for Time	357
28.1	Periodic Work	357
28.2	Timeouts	358
28.3	Timeslicing	358
28.4	Inline Execution Delays	359
28.5	Summary	359
29	Waiting for Queues	361
29.1	Overview	361
29.2	Request Distribution	363
29.3	Queue Structure	364
29.4	Worker Tasks	365
29.5	Primary Task	365
29.6	Dequeue	365
29.7	Enqueue	366
29.8	Spinlock	366
29.9	The “Work” Routine	367
29.10	Simple Examples	367

29.11	What Could Possibly Go Wrong?	368
29.12	CPU Frequency	369
29.13	Complex Examples	370
29.14	Waiting for CPUs: RPC Log	370
29.15	Waiting for CPUs: KUtrace	371
29.16	PlainSpinLock Flaw	374
29.17	Root Cause	375
29.18	PlainSpinLock Fixed: Observability	376
29.19	Load Balancing	377
29.20	Queue Depth: Observability	378
29.21	Spin at the End	378
29.22	One More Flaw	379
29.23	Cross-Checking	379
29.24	Summary	380
	Exercises	380
30	Recap	383
30.1	What You Learned	383
30.2	What We Haven't Covered	385
30.3	Next Steps	385
30.4	Summary (for the Entire Book)	386
A	Sample Servers	387
A.1	Sample Server Hardware	387
A.2	Connecting the Servers	388
B	Trace Entries	391
B.1	Fixed-Length Trace Entries	391
B.2	Variable-Length Trace Entries	392
B.3	Event Numbers	393
B.3.1	Events Inserted by Kernel-Mode KUtrace Patches	394
B.3.2	Events Inserted by User-Mode Code	395
B.3.3	Events Inserted by Postprocessing Code	395
	Glossary	397
	References	405
	Index	415