Network Systems Design

using

Network Processors

DOUGLAS E. COMER

Contents

Foreword xxi

Preface xxiii

Chapter 1 Introduction And Overview 1

1.1 *Network Systems And The Internet* 1
1.2 *Applications Vs. Infrastructure* 1
1.3 *Network Systems Engineering* 2
1.4 *Packet Processing* 2
1.5 *Achieving High Speed* 3
1.6 *Network Speed* 3
1.7 *Hardware, Software, And Hybrids* 4
1.8 *Scope And Organization Of The Text* 5
1.9 *Summary* 5

Chapter 2 Basic Terminology And Example Systems 7

2.1 *Introduction* 7
2.2 *Networks And Packets* 7
2.3 *Connection-Oriented And Connectionless Paradigms* 8
2.4 *Digital Circuits* 8
2.5 *LAN And WAN Classifications* 9
2.6 *The Internet And Heterogeneity* 9
2.7 *Example Network Systems* 9
2.8 *Broadcast Domains* 10
2.9 *The Two Key Systems Used In The Internet* 11
2.10 *Other Systems Used In The Internet* 12
2.11 *Monitoring And Control Systems* 13
2.12 *Summary* 13

Chapter 3 Review Of Protocols And Packet Formats 15

3.1 *Introduction* 15
3.2 *Protocols And Layering* 15
3.3 *Layers 1 And 2 (Physical And Network Interface)* 17
3.4 *Layer 3 (Internet)* 19
3.5 *Layer 4 (Transport)* 20
3.6 *Protocol Port Numbers And Demultiplexing* 23
3.7 *Encapsulation And Transmission* 23
3.8 *Address Resolution Protocol* 24
3.9 *Summary* 24

PART I Traditional Protocol Processing Systems

Chapter 4 Conventional Computer Hardware Architecture 29

4.1 *Introduction* 29
4.2 *A Conventional Computer System* 29
4.3 *Network Interface Cards* 30
4.4 *Definition Of A Bus* 31
4.5 *The Bus Address Space* 32
4.6 *The Fetch-Store Paradigm* 33
4.7 *Network Interface Card Functionality* 34
4.8 *NIC Optimizations For High Speed* 34
4.9 *Onboard Address Recognition* 35
4.10 *Onboard Packet Buffering* 36
4.11 *Direct Memory Access* 37
4.12 *Operation And Data Chaining* 38
4.13 *Data Flow Diagram* 39
4.14 *Promiscuous Mode* 39
4.15 *Summary* 40

Chapter 5 Basic Packet Processing: Algorithms And Data Structures 43

5.1 *Introduction* 43
5.2 *State Information and Resource Exhaustion* 43
5.3 *Packet Buffer Allocation* 44
5.4 *Packet Buffer Size And Copying* 45
5.5 *Protocol Layering And Copying* 45
5.6 *Heterogeneity And Network Byte Order* 46
5.7 *Bridge Algorithm* 47

5.8 *Table Lookup And Hashing* 49
5.9 *IP Datagram Fragmentation And Reassembly* 50
5.10 *IP Datagram Forwarding* 56
5.11 *IP Forwarding Algorithm* 57
5.12 *High-Speed IP Forwarding* 57
5.13 *TCP Connection Recognition Algorithm* 59
5.14 *TCP Splicing Algorithm* 60
5.15 *Summary* 63

Chapter 6 Packet Processing Functions 67

6.1 *Introduction* 67
6.2 *Packet Processing* 68
6.3 *Address Lookup And Packet Forwarding* 68
6.4 *Error Detection And Correction* 69
6.5 *Fragmentation, Segmentation, And Reassembly* 70
6.6 *Frame And Protocol Demultiplexing* 70
6.7 *Packet Classification* 71
6.8 *Queueing And Packet Discard* 73
6.9 *Scheduling And Timing* 75
6.10 *Security: Authentication And Privacy* 76
6.11 *Traffic Measurement And Policing* 76
6.12 *Traffic Shaping* 77
6.13 *Timer Management* 79
6.14 *Summary* 80

Chapter 7 Protocol Software On A Conventional Processor 83

7.1 *Introduction* 83
7.2 *Implementation Of Packet Processing In An Application* 83
7.3 *Fast Packet Processing In Software* 84
7.4 *Embedded Systems* 84
7.5 *Operating System Implementations* 85
7.6 *Software Interrupts And Priorities* 85
7.7 *Multiple Priorities And Kernel Threads* 87
7.8 *Thread Synchronization* 88
7.9 *Software For Layered Protocols* 88
7.10 *Asynchronous Vs. Synchronous Programming* 92
7.11 *Summary* 93

Chapter 8 Hardware Architectures For Protocol Processing 97

8.1 *Introduction* 97
8.2 *Network Systems Architecture* 97
8.3 *The Traditional Software Router* 98
8.4 *Aggregate Data Rate* 99
8.5 *Aggregate Packet Rate* 99
8.6 *Packet Rate And Software Router Feasibility* 101
8.7 *Overcoming The Single CPU Bottleneck* 103
8.8 *Fine-Grain Parallelism* 104
8.9 *Symmetric Coarse-Grain Parallelism* 104
8.10 *Asymmetric Coarse-Grain Parallelism* 105
8.11 *Special-Purpose Coprocessors* 105
8.12 *ASIC Coprocessor Implementation* 106
8.13 *NICs With Onboard Processing* 107
8.14 *Smart NICs With Onboard Stacks* 108
8.15 *Cells And Connection-Oriented Addressing* 108
8.16 *Data Pipelines* 109
8.17 *Summary* 111

Chapter 9 Classification And Forwarding 115

9.1 *Introduction* 115
9.2 *Inherent Limits Of Demultiplexing* 115
9.3 *Packet Classification* 116
9.4 *Software Implementation Of Classification* 117
9.5 *Optimizing Software-Based Classification* 118
9.6 *Software Classification On Special-Purpose Hardware* 119
9.7 *Hardware Implementation Of Classification* 119
9.8 *Optimized Classification Of Multiple Rule Sets* 120
9.9 *Classification Of Variable-Size Headers* 122
9.10 *Hybrid Hardware/Software Classification* 123
9.11 *Dynamic Vs. Static Classification* 124
9.12 *Fine-Grain Flow Creation* 125
9.13 *Flow Forwarding In A Connection-Oriented Network* 126
9.14 *Connectionless Network Classification And Forwarding* 126
9.15 *Second Generation Network Systems* 127
9.16 *Embedded Processors In Second Generation Systems* 128
9.17 *Classification And Forwarding Chips* 129
9.18 *Summary* 130

Chapter 10 Switching Fabrics 133

10.1 Introduction 133
10.2 Bandwidth Of An Internal Fast Path 133
10.3 The Switching Fabric Concept 134
10.4 Synchronous And Asynchronous Fabrics 135
10.5 A Taxonomy Of Switching Fabric Architectures 136
10.6 Dedicated Internal Paths And Port Contention 136
10.7 Crossbar Architecture 137
10.8 Basic Queueing 139
10.9 Time Division Solutions: Sharing Data Paths 141
10.10 Shared Bus Architecture 141
10.11 Other Shared Medium Architectures 142
10.12 Shared Memory Architecture 143
10.13 Multistage Fabrics 144
10.14 Banyan Architecture 145
10.15 Scaling A Banyan Switch 146
10.16 Commercial Technologies 148
10.17 Summary 148

PART II Network Processor Technology

Chapter 11 Network Processors: Motivation And Purpose 153

11.1 Introduction 153
11.2 The CPU In A Second Generation Architecture 153
11.3 Third Generation Network Systems 154
11.4 The Motivation For Embedded Processors 155
11.5 RISC vs. CISC 155
11.6 The Need For Custom Silicon 156
11.7 Definition Of A Network Processor 157
11.8 A Fundamental Idea: Flexibility Through Programmability 158
11.9 Instruction Set 159
11.10 Scalability With Parallelism And Pipelining 159
11.11 The Costs And Benefits Of Network Processors 160
11.12 Network Processors And The Economics Of Success 161
11.13 The Status And Future Of Network Processors 162
11.14 Summary 162

Chapter 12 The Complexity Of Network Processor Design 165

12.1 Introduction 165
12.2 Network Processor Functionality 165
12.3 Packet Processing Functions 166
12.4 Ingress And Egress Processing 167
12.5 Parallel And Distributed Architecture 170
12.6 The Architectural Roles Of Network Processors 171
12.7 Consequences For Each Architectural Role 171
12.8 Macroscopic Data Pipelining And Heterogeneity 173
12.9 Network Processor Design And Software Emulation 173
12.10 Summary 174

Chapter 13 Network Processor Architectures 177

13.1 Introduction 177
13.2 Architectural Variety 177
13.3 Primary Architectural Characteristics 178
13.4 Architecture, Packet Flow, And Clock Rates 186
13.5 Software Architecture 189
13.6 Assigning Functionality To The Processor Hierarchy 189
13.7 Summary 191

Chapter 14 Issues In Scaling A Network Processor 195

14.1 Introduction 195
14.2 The Processing Hierarchy And Scaling 195
14.3 Scaling By Making Processors Faster 196
14.4 Scaling By Increasing The Number of Processors 196
14.5 Scaling By Increasing Processor Types 197
14.6 Scaling A Memory Hierarchy 198
14.7 Scaling By Increasing Memory Size 200
14.8 Scaling By Increasing Memory Bandwidth 200
14.9 Scaling By Increasing Types Of Memory 201
14.10 Scaling By Adding Memory Caches 202
14.11 Scaling With Content Addressable Memory 203
14.12 Using CAM for Packet Classification 205
14.13 Other Limitations On Scale 207
14.14 Software Scalability 208
14.15 Bottlenecks And Scale 209
14.16 Summary 209

Chapter 15 Examples Of Commercial Network Processors 213

15.1 Introduction 213
15.2 An Explosion Of Commercial Products 213
15.3 A Selection of Products 214
15.4 Multi-Chip Pipeline (Agere) 214
15.5 Augmented RISC Processor (Alchemy) 218
15.6 Embedded Processor Plus Coprocessors (AMCC) 219
15.7 Pipeline Of Homogeneous Processors (Cisco) 221
15.8 Configurable Instruction Set Processors (Cognigine) 222
15.9 Pipeline Of Heterogeneous Processors (EZchip) 223
15.10 Extensive And Diverse Processors (IBM) 225
15.11 Flexible RISC Plus Coprocessors (Motorola) 227
15.12 Summary 231

Chapter 16 Languages Used For Classification 233

16.1 Introduction 233
16.2 Optimized Classification 233
16.3 Imperative And Declarative Paradigms 234
16.4 A Programming Language For Classification 235
16.5 Automated Translation 235
16.6 Language Features That Aid Programming 236
16.7 The Relationship Between Language And Hardware 236
16.8 Efficiency And Execution Speed 237
16.9 Commercial Classification Languages 238
16.10 Intel's Network Classification Language (NCL) 238
16.11 An Example Of NCL Code 239
16.12 NCL Intrinsic Functions 242
16.13 Predicates 243
16.14 Conditional Rule Execution 243
16.15 Incremental Protocol Definition 244
16.16 NCL Set Facility 245
16.17 Other NCL Features 246
16.18 Agere's Functional Programming Language (FPL) 247
16.19 Two Pass Processing 247
16.20 Designating The First And Second Pass 249
16.21 Using Patterns For Conditionals 249
16.22 Symbolic Constants 251
16.23 Example FPL Code For Second Pass Processing 251
16.24 Sequential Pattern Matching Paradigm 252
16.25 Tree Functions And The BITS Default 254
16.26 Return Values 254

16.27 Passing Information To The Routing Engine 254
16.28 Access To Built-in And External Functions 255
16.29 Other FPL Features 255
16.30 Summary 257

Chapter 17 Design Tradeoffs And Consequences **261**

17.1 Introduction 261
17.2 Low Development Cost Vs. Performance 261
17.3 Programmability Vs. Processing Speed 262
17.4 Performance: Packet Rate, Data Rate, And Bursts 262
17.5 Speed Vs. Functionality 263
17.6 Per-Interface Rate Vs. Aggregate Data Rate 263
17.7 Network Processor Speed Vs. Bandwidth 263
17.8 Coprocessor Design: Lookaside Vs. Flow-Through 264
17.9 Pipelining: Uniform Vs. Synchronized 264
17.10 Explicit Parallelism Vs. Cost And Programmability 264
17.11 Parallelism: Scale Vs. Packet Ordering 265
17.12 Parallelism: Speed Vs. Stateful Classification 265
17.13 Memory: Speed Vs. Programmability 265
17.14 I/O Performance Vs. Pin Count 266
17.15 Programming Languages: A Three-Way Tradeoff 266
17.16 Multithreading: Throughput Vs. Programmability 266
17.17 Traffic Management Vs. Blind Forwarding At Low Cost 267
17.18 Generality Vs. Specific Architectural Role 267
17.19 Memory Type: Special-Purpose Vs. General-Purpose 267
17.20 Backward Compatibility Vs. Architectural Advances 268
17.21 Parallelism Vs. Pipelining 268
17.22 Summary 269

PART III Example Network Processor

Chapter 18 Overview Of The Intel Network Processor **273**

18.1 Introduction 273
18.2 Intel Terminology 273
18.3 IXA: Internet Exchange Architecture 274
18.4 IXP: Internet Exchange Processor 274
18.5 Basic IXP1200 Features 275
18.6 External Connections 275
18.7 Internal Components 278
18.8 IXP1200 Processor Hierarchy 279

18.9 IXP1200 Memory Hierarchy 281
18.10 Word And Longword Addressing 283
18.11 An Example Of Underlying Complexity 283
18.12 Other Hardware Facilities 285
18.13 Summary 285

Chapter 19 Embedded RISC Processor (StrongARM Core) 289

19.1 Introduction 289
19.2 Purpose Of An Embedded Processor 289
19.3 StrongARM Architecture 291
19.4 RISC Instruction Set And Registers 291
19.5 StrongARM Memory Architecture 292
19.6 StrongARM Memory Map 293
19.7 Virtual Address Space And Memory Management 294
19.8 Shared Memory And Address Translation 294
19.9 Internal Peripheral Units 295
19.10 Other I/O 296
19.11 User And Kernel Mode Operation 296
19.12 Coprocessor 15 297
19.13 Summary 297

Chapter 20 Packet Processor Hardware (Microengines And FBI) 301

20.1 Introduction 301
20.2 The Purpose Of Microengines 301
20.3 Microengine Architecture 302
20.4 The Concept Of Microsequencing 302
20.5 Microengine Instruction Set 303
20.6 Separate Memory Address Spaces 305
20.7 Execution Pipeline 305
20.8 The Concept Of Instruction Stalls 307
20.9 Conditional Branching And Pipeline Abort 308
20.10 Memory Access Delay 308
20.11 Hardware Threads And Context Switching 309
20.12 Microengine Instruction Store 311
20.13 Microengine Hardware Registers 312
20.14 General-Purpose Registers 312
20.15 Transfer Registers 314
20.16 Local Control And Status Registers (CSRs) 315
20.17 Inter-Processor Communication 315
20.18 FBI Unit 316

20.19 Transmit And Receive FIFOs 317
20.20 FBI Architecture And Push/Pull Engines 317
20.21 Scratchpad Memory 318
20.22 Hash Unit 319
20.23 Configuration, Control, and Status Registers 321
20.24 Summary 321

Chapter 21 Reference System And Software Development Kit (Bridal Veil, SDK) 325

21.1 Introduction 325
21.2 Reference Systems 325
21.3 The Intel Reference System 326
21.4 Host Operating System Choices 328
21.5 Operating System Used On The StrongARM 328
21.6 External File Access And Storage 329
21.7 PCI Ethernet Emulation 330
21.8 Bootstrapping The Reference Hardware 330
21.9 Running Software 331
21.10 System Reboot 332
21.11 Alternative Cross-Development Software 332
21.12 Summary 332

Chapter 22 Programming Model (ACE) 335

22.1 Introduction 335
22.2 The ACE Abstraction 335
22.3 ACE Definitions And Terminology 336
22.4 Four Conceptual Parts Of An ACE 336
22.5 Output Targets And Late Binding 337
22.6 An Example Of ACE Interconnection 337
22.7 Division Of An ACE Into Core And Microblock 338
22.8 Microblock Groups 339
22.9 Replicated Microblock Groups 340
22.10 Microblock Structure 340
22.11 The Dispatch Loop 341
22.12 Dispatch Loop Calling Conventions 342
22.13 Packet Queues 343
22.14 Exceptions 344
22.15 Crosscalls 345
22.16 Application Programs Outside The ACE Model 346
22.17 Summary 346

Chapter 23 ACE Run-Time Structure And StrongARM Facilities 349

23.1 Introduction 349
23.2 StrongARM Responsibilities 349
23.3 Principle Run-Time Components 350
23.4 Core Components Of ACEs 350
23.5 Object Management System (OMS) 351
23.6 Resource Manager 352
23.7 Operating System Specific Library (OSSL) 352
23.8 Action Services Library 353
23.9 Automated Microengine Assignment 353
23.10 ACE Program Structure 354
23.11 ACE Main Program And Event Loop 354
23.12 ACE Event Loop And Blocking 355
23.13 Asynchronous Programming Paradigm And Callbacks 356
23.14 Asynchronous Execution And Mutual Exclusion 358
23.15 Memory Allocation 359
23.16 Loading And Starting An ACE (ixstart) 360
23.17 ACE Data Allocation And Initialization 361
23.18 Crosscalls 362
23.19 Crosscall Declaration Using IDL 363
23.20 Communication Access Process (CAP) 364
23.21 Timer Management 364
23.22 NCL Classification, Actions, And Default 366
23.23 Summary 367

Chapter 24 Microengine Programming I 371

24.1 Introduction 371
24.2 Intel's Microengine Assembler 371
24.3 Microengine Assembly Language Syntax 372
24.4 Example Operand Syntax 373
24.5 Symbolic Register Names And Allocation 376
24.6 Register Types And Syntax 377
24.7 Local Register Scope, Nesting, And Shadowing 378
24.8 Register Assignments And Conflicts 379
24.9 The Macro Preprocessor 380
24.10 Macro Definition 380
24.11 Repeated Generation Of A Code Segment 382
24.12 Structured Programming Directives 383
24.13 Instructions That Can Cause A Context Switch 385
24.14 Indirect Reference 386
24.15 External Transfers 387

24.16 Library Macros And Transfer Register Allocation 388
24.17 Summary 389

Chapter 25 Microengine Programming II **393**

25.1 Introduction 393
25.2 Specialized Memory Operations 393
25.3 Buffer Pool Manipulation 394
25.4 Processor Coordination Via Bit Testing 394
25.5 Atomic Memory Increment 395
25.6 Processor Coordination Via Memory Locking 396
25.7 Control And Status Registers 397
25.8 Intel Dispatch Loop Macros 399
25.9 Packet Queues And Selection 400
25.10 Accessing Fields In A Packet Header 401
25.11 Initialization Required For Dispatch Loop Macros 402
25.12 Packet I/O And The Concept Of Mpackets 404
25.13 Packet Input Without Interrupts 405
25.14 Ingress Packet Transfer 406
25.15 Packet Egress 406
25.16 Other I/O Details 408
25.17 Summary 408

Chapter 26 An Example ACE **411**

26.1 Introduction 411
26.2 An Example Bump-In-The-Wire 411
26.3 Wwbump Design 412
26.4 Header Files 413
26.5 Microcode For Packet Classification And Processing 415
26.6 Microcode For The Dispatch Loop 419
26.7 Code For Core Component (Exception Handler) 422
26.8 ACE Structure 423
26.9 Code To Initialize And Finalize The Wwbump ACE 423
26.10 An Example Crosscall 426
26.11 Code For A Crosscall Function 430
26.12 System Configuration 432
26.13 A Potential Bottleneck In The Wwbump Design 437
26.14 Summary 437

Chapter 27 Intel's Second Generation Processors 441

27.1 Introduction 441
27.2 Use Of Dual Chips For Higher Data Rates 441
27.3 General Characteristics 442
27.4 Memory Hierarchy 443
27.5 External Connections And Buses 443
27.6 Flow Control Bus 443
27.7 Media Or Switch Fabric Interface 444
27.8 Internal Architecture 445
27.9 Physical Network Interfaces And Multiplexing 446
27.10 Microengine Enhancements 446
27.11 Support For Software Pipelining 447
27.12 The IXP2800 447
27.13 Summary 448

Appendix 1 Glossary Of Terms And Abbreviations 449

Bibliography 497

Index 501