
HIGH PERFORMANCE SWITCHES AND ROUTERS

H. JONATHAN CHAO and BIN LIU



WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

HIGH PERFORMANCE SWITCHES AND ROUTERS



THE WILEY BICENTENNIAL—KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

HIGH PERFORMANCE SWITCHES AND ROUTERS

H. JONATHAN CHAO and BIN LIU



WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2007 by John Wiley & Sons, Inc., All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data.

Chao, H. Jonathan, 1955-

High performance switches and routers / by H. Jonathan Chao, Bin Liu.
p. cm.

ISBN-13: 978-0-470-05367-6

ISBN-10: 0-470-05367-4

1. Asynchronous transfer mode. 2. Routers (Computer networks)
3. Computer network protocols. 4. Packet switching (Data transmission)

I. Liu, Bin. II. Title.

TK5105.35.C454 2007

621.382'16--dc22

2006026971

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE	xv
ACKNOWLEDGMENTS	xvii
1 INTRODUCTION	1
1.1 Architecture of the Internet: Present and Future / 2	
1.1.1 The Present / 2	
1.1.2 The Future / 4	
1.2 Router Architectures / 5	
1.3 Commercial Core Router Examples / 9	
1.3.1 T640 TX-Matrix / 9	
1.3.2 Carrier Routing System (CRS-1) / 11	
1.4 Design of Core Routers / 13	
1.5 IP Network Management / 16	
1.5.1 Network Management System Functionalities / 16	
1.5.2 NMS Architecture / 17	
1.5.3 Element Management System / 18	
1.6 Outline of the Book / 19	
2 IP ADDRESS LOOKUP	25
2.1 Overview / 25	
2.2 Trie-Based Algorithms / 29	
2.2.1 Binary Trie / 29	
2.2.2 Path-Compressed Trie / 31	

2.2.3	Multi-Bit Trie / 33
2.2.4	Level Compression Trie / 35
2.2.5	Lulea Algorithm / 37
2.2.6	Tree Bitmap Algorithm / 42
2.2.7	Tree-Based Pipelined Search / 45
2.2.8	Binary Search on Prefix Lengths / 47
2.2.9	Binary Search on Prefix Range / 48
2.3	Hardware-Based Schemes / 51
2.3.1	DIR-24-8-BASIC Scheme / 51
2.3.2	DIR-Based Scheme with Bitmap Compression (BC-16-16) / 53
2.3.3	Ternary CAM for Route Lookup / 57
2.3.4	Two Algorithms for Reducing TCAM Entries / 58
2.3.5	Reducing TCAM Power – CoolCAMs / 60
2.3.6	TCAM-Based Distributed Parallel Lookup / 64
2.4	IPv6 Lookup / 67
2.4.1	Characteristics of IPv6 Lookup / 67
2.4.2	A Folded Method for Saving TCAM Storage / 67
2.4.3	IPv6 Lookup via Variable-Stride Path and Bitmap Compression / 69
2.5	Comparison / 73

3 PACKET CLASSIFICATION

77

3.1	Introduction / 77
3.2	Trie-Based Classifications / 81
3.2.1	Hierarchical Tries / 81
3.2.2	Set-Pruning Trie / 82
3.2.3	Grid of Tries / 83
3.2.4	Extending Two-Dimensional Schemes / 84
3.2.5	Field-Level Trie Classification (FLTC) / 85
3.3	Geometric Algorithms / 90
3.3.1	Background / 90
3.3.2	Cross-Producing Scheme / 91
3.3.3	Bitmap-Intersection / 92
3.3.4	Parallel Packet Classification (P^2C) / 93
3.3.5	Area-Based Quadtree / 95
3.3.6	Hierarchical Intelligent Cuttings / 97
3.3.7	HyperCuts / 98
3.4	Heuristic Algorithms / 103
3.4.1	Recursive Flow Classification / 103
3.4.2	Tuple Space Search / 107

- 3.5 TCAM-Based Algorithms / 108
 - 3.5.1 Range Matching in TCAM-Based Packet Classification / 108
 - 3.5.2 Range Mapping in TCAMs / 110

4 TRAFFIC MANAGEMENT

114

- 4.1 Quality of Service / 114
 - 4.1.1 QoS Parameters / 115
 - 4.1.2 Traffic Parameters / 116
- 4.2 Integrated Services / 117
 - 4.2.1 Integrated Service Classes / 117
 - 4.2.2 IntServ Architecture / 117
 - 4.2.3 Resource ReSerVation Protocol (RSVP) / 119
- 4.3 Differentiated Services / 121
 - 4.3.1 Service Level Agreement / 122
 - 4.3.2 Traffic Conditioning Agreement / 123
 - 4.3.3 Differentiated Services Network Architecture / 123
 - 4.3.4 Network Boundary Traffic Classification and Conditioning / 124
 - 4.3.5 Per Hop Behavior (PHB) / 126
 - 4.3.6 Differentiated Services Field / 127
 - 4.3.7 PHB Implementation with Packet Schedulers / 128
- 4.4 Traffic Policing and Shaping / 129
 - 4.4.1 Location of Policing and Shaping Functions / 130
 - 4.4.2 ATM's Leaky Bucket / 131
 - 4.4.3 IP's Token Bucket / 133
 - 4.4.4 Traffic Policing / 134
 - 4.4.5 Traffic Shaping / 135
- 4.5 Packet Scheduling / 136
 - 4.5.1 Max-Min Scheduling / 136
 - 4.5.2 Round-Robin Service / 138
 - 4.5.3 Weighted Round-Robin Service / 139
 - 4.5.4 Deficit Round-Robin Service / 140
 - 4.5.5 Generalized Processor Sharing (GPS) / 141
 - 4.5.6 Weighted Fair Queuing (WFQ) / 146
 - 4.5.7 Virtual Clock / 150
 - 4.5.8 Self-Clocked Fair Queuing / 153
 - 4.5.9 Worst-Case Fair Weighted Fair Queuing (WF²Q) / 155
 - 4.5.10 WF²Q+ / 158
 - 4.5.11 Comparison / 159
 - 4.5.12 Priorities Sorting Using a Sequencer / 160

- 4.6 Buffer Management / 163
 - 4.6.1 Tail Drop / 163
 - 4.6.2 Drop on Full / 164
 - 4.6.3 Random Early Detection (RED) / 164
 - 4.6.4 Differential Dropping: RIO / 167
 - 4.6.5 Fair Random Early Detection (FRED) / 168
 - 4.6.6 Stabilized Random Early Detection (SRED) / 170
 - 4.6.7 Longest Queue Drop (LQD) / 172

5 BASICS OF PACKET SWITCHING 176

- 5.1 Fundamental Switching Concept / 177
- 5.2 Switch Fabric Classification / 181
 - 5.2.1 Time-Division Switching / 181
 - 5.2.2 Space-Division Switching / 183
- 5.3 Buffering Strategy in Switching Fabrics / 187
 - 5.3.1 Shared-Memory Queuing / 188
 - 5.3.2 Output Queuing (OQ) / 188
 - 5.3.3 Input Queuing / 189
 - 5.3.4 Virtual Output Queuing (VOQ) / 189
 - 5.3.5 Combined Input and Output Queuing / 190
 - 5.3.6 Crosspoint Queuing / 191
- 5.4 Multiplane Switching and Multistage Switching / 191
- 5.5 Performance of Basic Switches / 195
 - 5.5.1 Traffic Model / 196
 - 5.5.2 Input-Buffered Switches / 197
 - 5.5.3 Output-Buffered Switches / 199
 - 5.5.4 Completely Shared-Buffered Switches / 201

6 SHARED-MEMORY SWITCHES 207

- 6.1 Linked List Approach / 208
- 6.2 Content Addressable Memory Approach / 213
- 6.3 Space-Time-Space Approach / 215
- 6.4 Scaling the Shared-Memory Switches / 217
 - 6.4.1 Washington University Gigabit Switch / 217
 - 6.4.2 Concentrator-Based Growable Switch Architecture / 218
 - 6.4.3 Parallel Shared-Memory Switches / 218
- 6.5 Multicast Shared-Memory Switches / 220
 - 6.5.1 Shared-Memory Switch with a Multicast Logical Queue / 220
 - 6.5.2 Shared-Memory Switch with Cell Copy / 220
 - 6.5.3 Shared-Memory Switch with Address Copy / 222

7	INPUT-BUFFERED SWITCHES	225
7.1	Scheduling in VOQ-Based Switches /	226
7.2	Maximum Matching /	229
7.2.1	Maximum Weight Matching /	229
7.2.2	Approximate MWM /	229
7.2.3	Maximum Size Matching /	230
7.3	Maximal Matching /	231
7.3.1	Parallel Iterative Matching (PIM) /	232
7.3.2	Iterative Round-Robin Matching (<i>i</i> RRM) /	233
7.3.3	Iterative Round-Robin with SLIP (<i>i</i> SLIP) /	234
7.3.4	FIRM /	241
7.3.5	Dual Round-Robin Matching (DRRM) /	241
7.3.6	Pipelined Maximal Matching /	245
7.3.7	Exhaustive Dual Round-Robin Matching (EDRRM) /	248
7.4	Randomized Matching Algorithms /	249
7.4.1	Randomized Algorithm with Memory /	250
7.4.2	A Derandomized Algorithm with Memory /	250
7.4.3	Variant Randomize Matching Algorithms /	251
7.4.4	Polling Based Matching Algorithms /	254
7.4.5	Simulated Performance /	258
7.5	Frame-based Matching /	262
7.5.1	Reducing the Reconfiguration Frequency /	263
7.5.2	Fixed Size Synchronous Frame-Based Matching /	267
7.5.3	Asynchronous Variable-Size Frame-Based Matching /	270
7.6	Stable Matching with Speedup /	273
7.6.1	Output-Queuing Emulation with Speedup of 4 /	274
7.6.2	Output-Queuing Emulation with Speedup of 2 /	275
7.6.3	Lowest Output Occupancy Cell First (LOOFA) /	278
8	BANYAN-BASED SWITCHES	284
8.1	Banyan Networks /	284
8.2	Batcher-Sorting Network /	287
8.3	Output Contention Resolution Algorithms /	288
8.3.1	Three-Phase Implementation /	288
8.3.2	Ring Reservation /	288
8.4	The Sunshine Switch /	292
8.5	Deflection Routing /	294
8.5.1	Tandem Banyan Switch /	294
8.5.2	Shuffle-Exchange Network with Deflection Routing /	296
8.5.3	Dual Shuffle-Exchange Network with Error-Correcting Routing /	297

- 8.6 Multicast Copy Networks / 303
 - 8.6.1 Broadcast Banyan Network / 304
 - 8.6.2 Encoding Process / 308
 - 8.6.3 Concentration / 309
 - 8.6.4 Decoding Process / 310
 - 8.6.5 Overflow and Call Splitting / 310
 - 8.6.6 Overflow and Input Fairness / 311

9 KNOCKOUT-BASED SWITCHES 316

- 9.1 Single-Stage Knockout Switch / 317
 - 9.1.1 Basic Architecture / 317
 - 9.1.2 Knockout Concentration Principle / 318
 - 9.1.3 Construction of the Concentrator / 320
- 9.2 Channel Grouping Principle / 323
 - 9.2.1 Maximum Throughput / 324
 - 9.2.2 Generalized Knockout Principle / 325
- 9.3 Two-Stage Multicast Output-Buffered ATM Switch (MOBAS) / 327
 - 9.3.1 Two-Stage Configuration / 327
 - 9.3.2 Multicast Grouping Network (MGN) / 330
- 9.4 Appendix / 333

10 THE ABACUS SWITCH 336

- 10.1 Basic Architecture / 337
- 10.2 Multicast Contention Resolution Algorithm / 340
- 10.3 Implementation of Input Port Controller / 342
- 10.4 Performance / 344
 - 10.4.1 Maximum Throughput / 344
 - 10.4.2 Average Delay / 347
 - 10.4.3 Cell Loss Probability / 349
- 10.5 ATM Routing and Concentration (ARC) Chip / 351
- 10.6 Enhanced Abacus Switch / 354
 - 10.6.1 Memoryless Multi-Stage Concentration Network / 354
 - 10.6.2 Buffered Multi-Stage Concentration Network / 357
 - 10.6.3 Resequencing Cells / 359
 - 10.6.4 Complexity Comparison / 361
- 10.7 Abacus Switch for Packet Switching / 362
 - 10.7.1 Packet Interleaving / 362
 - 10.7.2 Cell Interleaving / 364

11 CROSSPOINT BUFFERED SWITCHES 367

- 11.1 Combined Input and Crosspoint Buffered Switches / 368

- 11.2 Combined Input and Crosspoint Buffered Switches with VOQ / 370
 - 11.2.1 CIXB with One-Cell Crosspoint Buffers (CIXB-1) / 371
 - 11.2.2 Throughput and Delay Performance / 371
 - 11.2.3 Non-Negligible Round-Trip Times in CIXB- k / 376
- 11.3 OCF_OCF: Oldest Cell First Scheduling / 376
- 11.4 LQF_RR: Longest Queue First and Round-Robin Scheduling in CIXB-1 / 378
- 11.5 MCBF: Most Critical Buffer First Scheduling / 379

12 CLOS-NETWORK SWITCHES 382

- 12.1 Routing Property of Clos Network Switches / 383
- 12.2 Looping Algorithm / 387
- 12.3 m -Matching Algorithm / 388
- 12.4 Euler Partition Algorithm / 388
- 12.5 Karol's Algorithm / 389
- 12.6 Frame-Based Matching Algorithm for Clos Network (f-MAC) / 391
- 12.7 Concurrent Matching Algorithm for Clos Network (c-MAC) / 392
- 12.8 Dual-Level Matching Algorithm for Clos Network (d-MAC) / 395
- 12.9 The ATLANTA Switch / 398
- 12.10 Concurrent Round-Robin Dispatching (CRRD) Scheme / 400
- 12.11 The Path Switch / 404
 - 12.11.1 Homogeneous Capacity and Route Assignment / 406
 - 12.11.2 Heterogeneous Capacity Assignment / 408

13 MULTI-PLANE MULTI-STAGE BUFFERED SWITCH 413

- 13.1 TrueWay Switch Architecture / 414
 - 13.1.1 Stages of the Switch / 415
- 13.2 Packet Scheduling / 417
 - 13.2.1 Partial Packet Interleaving (PPI) / 419
 - 13.2.2 Dynamic Packet Interleaving (DPI) / 419
 - 13.2.3 Head-of-Line (HOL) Blocking / 420
- 13.3 Stage-To-Stage Flow Control / 420
 - 13.3.1 Back-Pressure / 421
 - 13.3.2 Credit-Based Flow Control / 421
 - 13.3.3 The DQ Scheme / 422
- 13.4 Port-To-Port Flow Control / 424
 - 13.4.1 Static Hashing / 424
 - 13.4.2 Dynamic Hashing / 425
 - 13.4.3 Time-Stamp-Based Resequence / 428
 - 13.4.4 Window-Based Resequence / 428

- 13.5 Performance Analysis / 431
 - 13.5.1 Random Uniform Traffic / 431
 - 13.5.2 Hot-Spot Traffic / 432
 - 13.5.3 Bursty Traffic / 432
 - 13.5.4 Hashing Schemes / 432
 - 13.5.5 Window-Based Resequencing Scheme / 434
- 13.6 Prototype / 434

14 LOAD-BALANCED SWITCHES 438

- 14.1 Birkhoff–Von Neumann Switch / 438
- 14.2 Load-Balanced Birkhoff–von Neumann Switches / 441
 - 14.2.1 Load-Balanced Birkhoff–von Neumann Switch Architecture / 441
 - 14.2.2 Performance of Load-Balanced Birkhoff–von Neumann Switches / 442
- 14.3 Load-Balanced Birkhoff–von Neumann Switches With FIFO Service / 444
 - 14.3.1 First Come First Served (FCFS) / 446
 - 14.3.2 Earliest Deadline First (EDF) and EDF-3DQ / 450
 - 14.3.3 Full Frames First (FFF) / 451
 - 14.3.4 Full Ordered Frames First (FOFF) / 455
 - 14.3.5 Mailbox Switch / 456
 - 14.3.6 Byte-Focal Switch / 459

15 OPTICAL PACKET SWITCHES 468

- 15.1 Opto-Electronic Packet Switches / 469
 - 15.1.1 Hypass / 469
 - 15.1.2 Star-Track / 471
 - 15.1.3 Cisneros and Brackett / 472
 - 15.1.4 BNR (Bell-North Research) Switch / 473
 - 15.1.5 Wave-Mux Switch / 474
- 15.2 Optoelectronic Packet Switch Case Study I / 475
 - 15.2.1 Speedup / 476
 - 15.2.2 Data Packet Flow / 477
 - 15.2.3 Optical Interconnection Network (OIN) / 477
 - 15.2.4 Ping-Pong Arbitration Unit / 482
- 15.3 Optoelectronic Packet Switch Case Study II / 490
 - 15.3.1 Petabit Photonic Packet Switch Architecture / 490
 - 15.3.2 Photonic Switch Fabric (PSF) / 495
- 15.4 All Optical Packet Switches / 503
 - 15.4.1 The Staggering Switch / 503
 - 15.4.2 ATMOS / 504

- 15.4.3 Duan's Switch / 505
- 15.4.4 3M Switch / 506
- 15.5 Optical Packet Switch with Shared Fiber Delay Lines
Single-stage Case / 509
 - 15.5.1 Optical Cell Switch Architecture / 509
 - 15.5.2 Sequential FDL Assignment (SEFA) Algorithm / 512
 - 15.5.3 Multi-Cell FDL Assignment (MUFA) Algorithm / 518
- 15.6 All Optical Packet Switch with Shared Fiber Delay
Lines – Three Stage Case / 524
 - 15.6.1 Sequential FDL Assignment for
Three-Stage OCNS (SEFAC) / 526
 - 15.6.2 Multi-Cell FDL Assignment for
Three-Stage OCNS (MUFAC) / 526
 - 15.6.3 FDL Distribution in Three-Stage OCNS / 528
 - 15.6.4 Performance Analysis of SEFAC and MUFAC / 530
 - 15.6.5 Complexity Analysis of SEFAC and MUFAC / 532

16 HIGH-SPEED ROUTER CHIP SET

538

- 16.1 Network Processors (NPs) / 538
 - 16.1.1 Overview / 538
 - 16.1.2 Design Issues for Network Processors / 539
 - 16.1.3 Architecture of Network Processors / 542
 - 16.1.4 Examples of Network Processors – Dedicated Approach / 543
- 16.2 Co-Processors for Packet Classification / 554
 - 16.2.1 LA-1 Bus / 554
 - 16.2.2 TCAM-Based Classification Co-Processor / 556
 - 16.2.3 Algorithm-Based Classification Co-Processor / 562
- 16.3 Traffic Management Chips / 567
 - 16.3.1 Overview / 567
 - 16.3.2 Agere's TM Chip Set / 567
 - 16.3.3 IDT TM Chip Set / 573
 - 16.3.4 Summary / 579
- 16.4 Switching Fabric Chips / 579
 - 16.4.1 Overview / 579
 - 16.4.2 Switch Fabric Chip Set from Vitesse / 580
 - 16.4.3 Switch Fabric Chip Set from AMCC / 589
 - 16.4.4 Switch Fabric Chip Set from IBM (now of AMCC) / 593
 - 16.4.5 Switch Fabric Chip Set from Agere / 597

INDEX

606