Contents

Foreword xiii
Ahmed JERRAYA

Acknowledgments xv
Liliana ANDRADE and Frédéric ROUSSEAU

Part 1. Processors 1

Chapter 1. Processors for the Internet of Things 3
Pieter VAN DER WOLF and Yankin TANURHAN

1.1. Introduction 3
1.2. Versatile processors for low-power IoT edge devices 4
1.2.1. Control processing, DSP and machine learning 4
1.2.2. Configurability and extensibility 6
1.3. Machine learning inference 8
1.3.1. Requirements for low/mid-end machine learning inference 10
1.3.2. Processor capabilities for low-power machine learning inference 14
1.3.3. A software library for machine learning inference 17
1.3.4. Example machine learning applications and benchmarks 20

1.4. Conclusion 23
1.5. References 24

Chapter 2. A Qualitative Approach to Many-core Architecture 27
Benoît DUPONT DE DINECHIN

2.1. Introduction 28
2.2. Motivations and context 29
2.2.1. Many-core processors 29
2.2.2. Machine learning inference 30
2.2.3. Application requirements 32
2.3. The MPPA3 many-core processor 34
2.3.1. Global architecture 34
2.3.2. Compute cluster 36
2.3.3. VLIW core 38
2.3.4. Coprocessor 39
2.4. The MPPA3 software environments 42
2.4.1. High-performance computing 42
2.4.2. KaNN code generator 43
2.4.3. High-integrity computing 46
2.5. Conclusion 47
2.6. References 48

Chapter 3. The Plural Many-core Architecture – High Performance at Low Power 53
Ran GINOSAR

3.1. Introduction 54
3.2. Related works 55
3.3. Plural many-core architecture 55
3.4. Plural programming model 56
3.5. Plural hardware scheduler/synchronizer 58
3.6. Plural networks-on-chip 61
3.6.1. Scheduler NoC 61
3.6.2. Shared memory NoC 61
3.7. Hardware and software accelerators for the Plural architecture 62
3.8. Plural system software 63
3.9. Plural software development tools 65
3.10. Matrix multiplication algorithm on the Plural architecture 65
3.11. Conclusion 67
3.12. References 67

Chapter 4. ASIP-Based Multi-Processor Systems for an Efficient Implementation of CNNs . . . 69
Andreas BYTYN, René AHLSDORF and Gerd ASCHEID

4.1. Introduction . . . 70
4.2. Related works . . . 71
4.3. ASIP architecture . . . 74
4.4. Single-core scaling . . . 75
4.5. MPSoC overview . . . 78
4.6. NoC parameter exploration . . . 79
4.7. Summary and conclusion . . . 82
4.8. References . . . 83

Part 2. Memory . . . 85

Chapter 5. Tackling the MPSoC Data Locality Challenge . . . 87
Sven RHEINDT, Akshay SRIVATSA, Oliver LENKE, Lars NOLTE, Thomas WILD and Andreas HERKERSDORF

5.1. Motivation . . . 88
5.2. MPSoC target platform . . . 90
5.3. Related work . . . 91
5.4. Coherence-on-demand: region-based cache coherence . . . 92
5.4.1. RBCC versus global coherence . . . 93
5.4.2. OS extensions for coherence-on-demand . . . 94
5.4.3. Coherency region manager . . . 94
5.4.4. Experimental evaluations . . . 97
5.4.5. RBCC and data placement . . . 99
5.5. Near-memory acceleration . . . 100
5.5.1. Near-memory synchronization accelerator . . . 102
5.5.2. Near-memory queue management accelerator . . . 104
5.5.3. Near-memory graph copy accelerator . . . 107
5.5.4. Near-cache accelerator . . . 110
5.6. The big picture . . . 111
5.7. Conclusion . . . 113
5.8. Acknowledgments . . . 114
5.9. References . . . 114

Chapter 6. mMPU: Building a Memristor-based General-purpose In-memory Computation Architecture . . . 119
Adi ELIAHU, Rotem BEN HUR, Ameer HAJ ALI and Shahar KVATINSKY

6.1. Introduction . . . 120
6.2. MAGIC NOR gate . . . 121

6.3. In-memory algorithms for latency reduction 122
6.4. Synthesis and in-memory mapping methods 123
6.4.1. SIMPLE 124
6.4.2. SIMPLER 126
6.5. Designing the memory controller 127
6.6. Conclusion 129
6.7. References 130

Chapter 7. Removing Load/Store Helpers in Dynamic Binary Translation 133
Antoine FARAVELON, Olivier GRUBER and Frédéric PÉTROT

7.1. Introduction 134
7.2. Emulating memory accesses 136
7.3. Design of our solution 140
7.4. Implementation 143
7.4.1. Kernel module 143
7.4.2. Dynamic binary translation 145
7.4.3. Optimizing our slow path 147
7.5. Evaluation 149
7.5.1. QEMU emulation performance analysis 150
7.5.2. Our performance overview 151
7.5.3. Optimized slow path 153
7.6. Related works 155
7.7. Conclusion 157
7.8. References 158

Chapter 8. Study and Comparison of Hardware Methods for Distributing Memory Bank Accesses in Many-core Architectures 161
Arthur VIANES and Frédéric ROUSSEAU

8.1. Introduction 162
8.1.1. Context 162
8.1.2. MPSoC architecture 163
8.1.3. Interconnect 164
8.2. Basics on banked memory 165
8.2.1. Banked memory 165
8.2.2. Memory bank conflict and granularity 166
8.2.3. Efficient use of memory banks: interleaving 168
8.3. Overview of software approaches 170
8.3.1. Padding 170
8.3.2. Static scheduling of memory accesses 172

8.3.3. The need for hardware approaches 172
8.4. Hardware approaches 172
8.4.1. Prime modulus indexing 172
8.4.2. Interleaving schemes using hash functions 174
8.5. Modeling and experimenting 181
8.5.1. Simulator implementation 182
8.5.2. Implementation of the Kalray MPPA cluster interconnect 182
8.5.3. Objectives and method 184
8.5.4. Results and discussion 185
8.6. Conclusion 191
8.7. References 192

Part 3. Interconnect and Interfaces 195

Chapter 9. Network-on-Chip (NoC): The Technology that Enabled Multi-processor Systems-on-Chip (MPSoCs) 197
K. Charles JANAC

9.1. History: transition from buses and crossbars to NoCs 198
9.1.1. NoC architecture 202
9.1.2. Extending the bus comparison to crossbars 207
9.1.3. Bus, crossbar and NoC comparison summary and conclusion . . . 207
9.2. NoC configurability 208
9.2.1. Human-guided design flow 208
9.2.2. Physical placement awareness and NoC architecture design 209
9.3. System-level services 211
9.3.1. Quality-of-service (QoS) and arbitration 211
9.3.2. Hardware debug and performance analysis 212
9.3.3. Functional safety and security 212
9.4. Hardware cache coherence 215
9.4.1. NoC protocols, semantics and messaging 216
9.5. Future NoC technology developments 217
9.5.1. Topology synthesis and floorplan awareness 217
9.5.2. Advanced resilience and functional safety for autonomous vehicles 218
9.5.3. Alternatives to von Neumann architectures for SoCs 219
9.5.4. Chiplets and multi-die NoC connectivity 221
9.5.5. Runtime software automation 222
9.5.6. Instrumentation, diagnostics and analytics for performance, safety and security 223
9.6. Summary and conclusion 224
9.7. References 224

Chapter 10. Minimum Energy Computing via Supply and Threshold Voltage Scaling 227
Jun SHIOMI and Tohru ISHIHARA

10.1. Introduction 228
10.2. Standard-cell-based memory for minimum energy computing 230
10.2.1. Overview of low-voltage on-chip memories 230
10.2.2. Design strategy for area- and energy-efficient SCMs 234
10.2.3. Hybrid memory design towards energy- and area-efficient memory systems 236
10.2.4. Body biasing as an alternative to power gating 237
10.3. Minimum energy point tracking 238
10.3.1. Basic theory 238
10.3.2. Algorithms and implementation 244
10.3.3. OS-based approach to minimum energy point tracking 246
10.4. Conclusion 249
10.5. Acknowledgments 249
10.6. References 250

Chapter 11. Maintaining Communication Consistency During Task Migrations in Heterogeneous Reconfigurable Devices 255
Arief WICAKSANA, Olivier MULLER, Frédéric ROUSSEAU and Arif SASONGKO

11.1. Introduction 256
11.1.1. Reconfigurable architectures 256
11.1.2. Contribution 257
11.2. Background 257
11.2.1. Definitions 258
11.2.2. Problem scenario and technical challenges 259
11.3. Related works 261
11.3.1. Hardware context switch 261
11.3.2. Communication management 262
11.4. Proposed communication methodology in hardware context switching 263
11.5. Implementation of the communication management on reconfigurable computing architectures 266
11.5.1. Reconfigurable channels in FIFO 267
11.5.2. Communication infrastructure 268
11.6. Experimental results 269
11.6.1. Setup 269
11.6.2. Experiment scenario 270
11.6.3. Resource overhead 271
11.6.4. Impact on the total execution time 273

11.6.5. Impact on the context extract and restore time 275
11.6.6. System responsiveness to context switch requests 276
11.6.7. Hardware task migration between heterogeneous FPGAs 280
11.7. Conclusion 282
11.8. References 283

List of Authors 287

Authors Biographies 291

Index 299