Contents

1 Introduction **14**

1.1 Motivation 14

1.2 Objective 15

1.3 Statement of Originality 15

1.4 Overview 17

2 Background **19**

2.1 Introduction 19

2.2 Processing In Memory Systems 22

2.2.1 Implementation of Processing In Memory Systems 23

2.2.2 Existing Architectures 27

2.2.3 Processing near Storage 33

2.3 Memory Architectures 34

2.3.1 Memory Hierarchy and Model 34

2.3.2 Application Specific Memory Architectures 37

2.3.3 Streaming Architectures 38
2.3.4 Physical Memory Implementation 38
2.3.5 Memory Communication Bandwidth 39
2.4 Address Generation Units 39
2.5 Synthesis of Data Arrays to Memory 41
2.6 Reconfigurable Hardware Architectures 43
2.6.1 Reconfigurable Processing and Configuration 43
2.6.2 Reconfigurable Routing Resources 44
2.6.3 Embedded Memory Blocks 45
2.7 Summary 47

3 Abstractions for Structured Data Access 49

3.1 Introduction 49
3.2 Design 51
3.2.1 First-In-First-Out (FIFO) or Circular Buffer 52
3.2.2 First-In-Multiple-Out (FIMO) 53
3.2.3 Last-In-First-Out (LIFO) or Stack 55
3.2.4 Swinging Buffer 56
3.2.5 Striped Access 57
3.2.6 Bypass Mode 59
3.2.7 Hybrid Mode 60
3.2.8 Address Generation Unit Implementation 60

3.2.9 Configuration of these Modes 65
3.3 Efficiency of Memory Design Abstractions 66
3.3.1 Experimental Setup 66
3.3.2 Performance and Area for Dedicated and Reconfigurable Implementations 67
3.3.3 Influence of Address-width on Area and Performance 71
3.3.4 Relation between Memory and AGU size 73
3.3.5 Performance Reduction in Relation to AMB 75
3.3.6 Applicability of AMB to Applications 75
3.4 Summary 77

4 AMB Concatenation for Scalable Structured Access 79

4.1 Introduction 79
4.2 Data Splitting 80
4.3 AMB Concatenation 84
4.3.1 FIFO, FIMO and Stack 84
4.3.2 Swinging Buffer 86
4.3.3 Striped Access 88
4.3.4 Hybrid Mode 98
4.4 Alternative Address Generation Methods for Concatenating Blocks . . . 99
4.5 Evaluation 101
4.6 Summary 106

5 Case Studies **108**

5.1 Introduction 108

5.2 Image Registration 109

5.2.1 Introduction 110

5.2.2 Application Overview 111

5.2.3 Motion Vector Estimation 116

5.2.4 Image Transformation 132

5.3 JPEG Encoding 134

5.3.1 Application Overview 134

5.3.2 JPEG Implementation based on a 2D DCT 137

5.3.3 JPEG Implementation based on two 1D DCTs 139

5.4 Summary 142

6 Conclusions and Future Work **144**

6.1 Conclusions 144

6.2 Future Work 147

Glossary **150**

Bibliography **157**

List of Figures

2.1 Diagram indicating the location of PIM systems when integrated in the normal PC architecture 27

2.2 Parallel processing PIM-based systems 31

2.3 General block diagram for processing near storage systems 33

2.4 Memory hierarchy present in current general-purpose computers with respective access times and storage capacity 35

2.5 Different memory models according to the location of the interconnection network. (P=processor, M=memory, C=cache) (a) Interconnection is between processor and memory; (b) each processor has its local cache, but shares the main memory and (c) only permanent storage resources are shared 36

2.6 Location of the heterogeneous embedded memory blocks in the Altera Stratix devices 46

3.1 Block Diagram of Address Generation Unit (AGU) in FIFO mode . . . 53

3.2 Indication of FIMO mode operation: overwriting of oldest item (a), followed by reading the full buffer (b), and again overwriting the oldest item (c) and reading the full buffer (d) 54

3.3 (a) 1D "Striped" access mode; (b) 2D "Striped" access mode 57

3.4 Hierarchical pattern required within the striped access mode (Different peak heights indicate a different jump-size, whereas the same height represents identical jump-size.) 58

3.5 Autonomous Memory Block (AMB) with Bypass mode 59

3.6 Diagram indicating implementation details of the AGU for AMB 61

3.7 Block diagram of the address generation counters within the current AGU implementation 62

3.8 Diagram indicating implementation details of the flow-control counters within the AMBs AGU 63

3.9 Influence of different address-width on area and performance of dedicated AGU 71

3.10 Influence of different address-width on performance of reconfigurable implementation for each of the different modes 72

3.11 Influence of different address-width on area for a reconfigurable implementation of the different modes 72

3.12 Gate count for memory and AGU as a function of address-width 73

4.1 Overview of the possible interleaving patterns for splitting data (The different numbers refer to different memory blocks) 83

4.2 Concatenation of two AMBs to provide larger buffering in the case of FIFO, FIMO and stack 85

4.3 Representation of the difference between a block and a buffer in relation to splitting for parallelism in swinging buffer mode 87

4.4 Data splitting for striped access by interleaving 90

4.5 Example indicating the first restriction for interleaving data in memory for striped access mode (interleaved storage using 2 memory blocks) . . 91

4.6 Example for second constraint of the interleaving approach (interleaved storage using 4 memory blocks) 92

4.7 Appending and Overwriting Scheme (AOS): appending data for striped access can be performed by overwriting a number of items and adding a minor amount of new data at the end 93

4.8 Data splitting for striped access using duplication 95

4.9 Data splitting for striped access by shifting data through registers in combination with FIFOs 97

5.1 Approach to image registration for watermarked data 112

5.2 Motion Vector Estimation: (a) & (b) Division into blocks and determined motion vector; (c) Motion Vector map for perspective distortion of similar type as shown in (b) 114

5.3 Indication of the different buffering levels investigated for the Motion Vector Estimation implementation (A single square represents a block of 4 by 4 pixels.) 118

5.4 Block diagram for the baseline JPEG encoder 135

5.5 Block diagram of implementing a JPEG encoder with 2D DCT, including memory architecture and processing 137

5.6 Block diagram for the 1D DCT JPEG encoder 140

List of Tables

3.1 Area and speed for a 9-bit dedicated AGU in comparison to the use of reconfigurable resources for address generation 68

3.2 Percentage of CLBs saved, depending on memory access mode and size of Xilinx Virtex-II device 69

3.3 Area and performance for library modules with 9-bit address-width . . . 70

3.4 Size of AGU for the different types of memory in Stratix 74

3.5 Influence on area and performance of including the scoreboard for striped access flow-control (9-bit address-width) 75

4.1 Difference in area and performance for the dedicated normal AMB version and the one that allows concatenation 101

4.2 Difference in area and performance for the reconfigurable FIFO and striped access mode implementation with and without concatenation possibility 102

4.3 Size of controlling memory wrapper to combine dedicated AMBs 103

4.4 Area and maximum performance for the implementation of a 2D filter (window size of 3 by 3) using the different striped access approaches . . 105

5.1 Required resources and respective parallelism for storing an array with block-width length (16 pixels) 120

5.2 The different memory architectures and their respective resource consumption for buffering a processing block 123

5.3 Possible memory architectures for the Motion Vector Estimation implementation when a block of block-height and search-width is buffered on-chip 125

5.4 Required amount of storage resources and available parallelism when the full search-area is buffered on-chip 127

5.5 Required resources and achievable performance when an area search-size high and with full image width is buffered 129

5.6 Required resources and available parallelism for buffering the whole image on-chip 131