Contents

List of Tables xv

List of Figures xvii

Listings xxiii

1 Introduction 1
1.1 Motivation 2
1.2 Main Contributions 4
1.3 Thesis Organization 6

2 Related Work 7
2.1 Video Decoding using Dedicated Hardware 7
2.2 Video Decoding using multi-core CPUs 8
2.3 Video Decoding using GPUs 9
2.3.1 GPU Parallelization for Video Coding Standards before HEVC . . . 9
2.3.2 GPU Parallelization for HEVC Video Coding Standard 11
2.4 Advances over State of the Art 12
2.5 Summary 13

3 Review of GPU Architecture and Video Decoding for Suitability Analysis 15
3.1 Review of General-purpose GPU Computation 15
3.1.1 Throughput-oriented Design of GPU Architecture 16
3.1.2 Programming Model for GPU Computation 18
3.1.3 Efficient Kernel Implementation on GPUs 26
3.2 Review of H.264 and HEVC Video Decoding 29
3.2.1 Anatomy of a Video Sequence 30
3.2.2 Video Decoding Kernels 31
3.2.3 Suitability Analysis of Video Decoding Kernels for GPU architecture 43
3.3 Summary 47

4 Offload H.264 Decoding onto GPUs 49
4.1 Introduction 49
4.2 GPU Parallelization of H.264 Inverse Transform 49
4.2.1 Inverse Transform in H.264 50
4.2.2 Frame Decoupling Code Adaptation 50
4.2.3 Proposed Implementation on GPU 52
4.2.4 Experimental Results 57

4.3 Parallelization of H.264 Motion Compensation 60
4.3.1 Motion Compensation in H.264 61
4.3.2 GPU Kernel Design 62
4.3.3 Experimental Setup 68
4.3.4 Experimental Results 70
4.4 Frame Pipelined H.264 Decoder 79
4.4.1 Sequential CPU+GPU Decoding 80
4.4.2 Frame Pipelined CPU+GPU Decoding 81
4.4.3 Experimental Evaluation 82
4.5 Summary 84

5 GPU Parallelization for HEVC In-loop Filters 87
5.1 Introduction 87
5.2 GPU Parallelization of HEVC Deblocking Filter at the Block Level 88
5.2.1 Block-level GPU Kernel Implementation for Deblocking Filter . . . 88
5.2.2 Experimental Setup 92
5.2.3 Experimental Results 92
5.3 GPU Parallelization of HEVC In-loop Filters at the Frame Level 93
5.3.1 CPU Frame-Decoupled (CPU-Frame) In-loop Filters 94
5.3.2 State-of-the-art GPU Parallelization for In-loop Filters 95
5.3.3 Optimizations over State-of-the-art GPU Parallelization 97
5.3.4 Experimental Setup 100
5.3.5 Experimental Results 102
5.4 Summary 108

6 Parallel HEVC Decoding on Heterogeneous CPU+GPU Systems 109
6.1 Introduction 109
6.2 Parallel HEVC Decoding Design for Heterogeneous Systems with CPU and GPU 109
6.2.1 Task Distribution and Parallel Decoding on Heterogeneous CPU+GPU Systems 110
6.2.2 Different Workload Balancing Schemes 114
6.2.3 Data Transfer Between the CPU and the GPU 116
6.3 Energy Measurement for Heterogeneous CPU+GPU Decoding 118
6.3.1 Energy Measurement of Intel CPUs 118
6.3.2 Energy Measurement of NVIDIA GPUs 119
6.4 Experimental Results 120
6.4.1 Performance Results 121
6.4.2 Energy Efficiency Results 129
6.5 Summary 135

7 Conclusions 137
7.1 Summary and Contributions 137
7.2 Conclusions 140

7.3 Suggestions for Future Work 141

Bibliography **143**