

Contents

Abstract	iii
Contents	ix
1 Introduction	1
1.1 Motivation	1
1.2 Background	3
1.3 Thesis Objectives	10
1.4 Contributions	11
1.5 Outline	11
2 Impact of Memory-Level Parallelism on the Performance of GPU Coherence Protocols	13
2.1 Abstract	14
2.2 Introduction	14
2.3 GPU Architecture	15
2.4 Axes of Characterization	15
2.5 Experimental Evaluation	17
2.6 Related Work	22

2.7	Conclusions	23
3	Accurately Modeling the GPU Memory Subsystem	25
3.1	Abstract	26
3.2	Introduction	26
3.3	Related Work	28
3.4	Southern Islands GPU Architecture and Programming Model	29
3.5	Proposed Multi2Sim GPU Extensions	32
3.6	Experimental Results	37
3.7	Conclusions	41
4	Accurately Modeling the On-chip and Off-chip GPU Memory Subsystem	43
4.1	Abstract	44
4.2	Introduction	44
4.3	Related Work	46
4.4	Southern Islands GPU Programming Model and Architecture	47
4.5	Modeled Memory Subsystem Components	50
4.6	Experimental Results	56
4.7	Putting it All Together and Validation	63
4.8	Conclusions	64
5	Improving GPU Cache Hierarchy Performance with a Fetch and Replacement Cache	67
5.1	Abstract	68
5.2	Introduction	68
5.3	Background	70
5.4	Motivation	71
5.5	FRC Approach	73
5.6	Experimental Evaluation	75
5.7	Related Work	80

5.8 Conclusions.	81
6 Efficient Management of Cache Accesses to 2 Boost GPGPU Memory Subsystem	
Performance	83
6.1 Abstract	84
6.2 Introduction	84
6.3 Background.	87
6.4 Motivation	88
6.5 FRC Implementation	92
6.6 Experimental Evaluation.	94
6.7 Related Work	105
6.8 Conclusions.	108
7 Results Discussion	
7.1 Characterization of GPGPU Applications	112
7.2 Simulator Framework Improvements and Validation.	113
7.3 Proposed LLC Miss Management Approach	114
8 Conclusions	
8.1 Characterization of GPGPU Applications	118
8.2 Simulator Framework Improvements and Validation.	118
8.3 Proposed LLC Miss Management Approach	120
8.4 Other Indirectly Related Work	121
8.5 Future Work	122
Bibliography	123