

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Challenges	4
1.2.1	Deep Neural Networks	5
1.2.2	PIM Architectures	6
1.3	Scope of This Thesis	8
1.3.1	Hardware Design	8
1.3.2	Software Design	9
1.3.3	HW/SW Co-design	9
1.4	Contributions	10
1.5	Thesis Outline	12
1.5.1	Background (Chapter 2)	12
1.5.2	In-Memory Crossbar Design (Chapter 3)	13
1.5.3	In-memory Architecture for Multi-Precision Neural Networks (Chapter 4)	13
1.5.4	Quantization and Hardware Approximations (Chapter 5)	14
1.5.5	Variability Mitigation and Analog Approximate Computing (Chapter 6)	14
1.5.6	Extended In-Memory Architectures (Chapter 7)	15
1.5.7	Conclusion and Future Work (Chapter 8)	15
2	Background	17
2.1	Deep Neural Networks (DNNs)	17
2.1.1	Convolutional Neural Networks (CNNs)	17
2.1.2	CNN Quantization	20
2.2	CNN Hardware Acceleration Trends	25
2.2.1	CNN Parallelism	25
2.2.2	Data Reuse	27
2.2.3	Approximate Computing	30
2.3	Processing-in-Memory	32
2.3.1	Memory Technologies	32
2.3.2	PIM Computational Domain	35
2.4	DNN Models and Datasets	37
2.4.1	MNIST	37
2.4.2	CIFAR-10 & CIFAR-100	38
2.4.3	ImageNet	38

3 In-Memory Crossbar Macro Design	41
3.1 Related Work	42
3.2 Bit-Decomposing the MAC Operation	43
3.3 FeFET Memory Cell for PIM	44
3.4 Analog to Digital Converter	46
3.4.1 Sequential Staircase ADC	46
3.4.2 Sequential Ladder ADC	47
3.4.3 FeFET-based ADC	48
3.4.4 Temperature Immune ADC	49
3.5 Crossbar Organization & Operation	50
3.6 Evaluation	53
3.6.1 Experimental Setup	53
3.6.2 ADC Evaluation	54
3.6.3 Crossbar Evaluation	56
3.7 Conclusion	58
4 FELIX: A Ferroelectric FET Based Architecture for DNN Acceleration	59
4.1 Related Work	60
4.1.1 Crossbar Array Computing	60
4.1.2 Bit-wise Computing	60
4.2 FELIX Architecture	61
4.2.1 FeFET Processing Element Design	61
4.2.2 PE Cluster	63
4.2.3 System Tile	64
4.2.4 Top Level Architecture	65
4.3 Architecture Design Space Exploration	66
4.4 Mapping & Accelerator Operation	67
4.4.1 Kernel weights mapping to the crossbar array	67
4.4.2 Accelerator Operation Timeline	68
4.5 Evaluation	69
4.5.1 Experimental Setup	69
4.5.2 Benchmarks Accuracy	70
4.5.3 Performance Analysis	71
4.5.4 Energy Efficiency	73
4.5.5 Area Efficiency	74
4.5.6 Comparison to DNN Accelerators	74
4.6 Conclusion	76
5 Digital Approximations for Adaptive Hardware Design	77
5.1 Related Work	78
5.2 Hardware Aware DNN Optimization	79
5.2.1 Hardware Realization	80
5.2.2 Experimental Results	81
5.3 Bit-Decomposition/Bit-Serialization Hardware Optimization	83
5.3.1 Extended Truncation Approximation	83

5.4	Evaluation	86
5.4.1	Evaluation setup	86
5.4.2	Hardware Evaluation	87
5.4.3	Accuracy Evaluation	87
5.4.4	Comparison with state-of-the-art DNN Accelerators	90
5.5	Conclusion	90
6	Mixed-signal Approximations and Variability Immunity Techniques	91
6.1	Related Work	92
6.1.1	Variability of PIM Devices	92
6.1.2	Approximate Computing	92
6.2	Variability Mitigation Methods	93
6.2.1	Device Characteristics	93
6.2.2	Variability Mitigation Methods	95
6.2.3	Evaluation	96
6.2.4	Comparison with DNN Accelerators	97
6.3	Adaptable Analog Approximation	98
6.3.1	Approximated Operation	98
6.4	Cross-Layer Analog Approximation	100
6.5	Overall Evaluation	102
6.5.1	Evaluation Setup	102
6.5.2	Design Space Exploration	103
6.5.3	Hardware Performance Gain	105
6.5.4	Accuracy Evaluation	106
6.6	Conclusion	108
7	Extended In-Memory Architectures	109
7.1	Related Work	110
7.1.1	DRAM-based Architectures	110
7.1.2	Binary Architectures	111
7.2	Bit-Decomposition DRAM-Based Architecture	111
7.2.1	Architecture Organization	111
7.2.2	Evaluation	112
7.3	Logarithmic Quantized DRAM-Based Architecture	114
7.3.1	Retraining One-hot Quantized Weights	114
7.3.2	Architecture Organization	116
7.3.3	Evaluation	117
7.4	Binary FeFET-based Architecture	121
7.4.1	Accelerator Design	122
7.4.2	Convolution Operation	122
7.4.3	Evaluation	123
7.5	Conclusion	124
8	Summary, Conclusion & Future Work	127
8.1	Truncated multiplier	139

8.2	Full Approximation and Fine-Tuning (Stage 1)	139
8.3	Kernel-wise Optimization (Stage 2)	139
8.4	Accuracy Evaluation	143
Acronyms		147
Bibliography		149
List of Figures		163
List of Tables		167
List of Algorithms		169