
Contents

Acknowledgements	vii
Abstract	ix
Contents	xv
List of Figures	xix
List of Tables	xxiii
1 Introduction	1
1.1 Motivation and Objectives	4
1.1.1 Research contributions and outline of the thesis	5
1.2 List of Publications	6
2 Background	9
2.1 GEMM in Computing	9
2.2 GEMM Algorithms	11
2.2.1 Family of GEMM algorithms	12
2.3 Convolution	13
2.3.1 GEMM-like direct convolution	15
2.3.2 Convolution via IM2COL + GEMM	18
2.4 RISC-V Low-power MCU (GAP8)	19
3 BLIS-like GEMM in the GAP8	21

3.1	Introduction	21
3.2	BLIS-like GEMM in the GAP8 Platform	22
3.2.1	Tailoring the micro-kernel and algorithm for the dot product	22
3.2.2	Pack functions using DMAs	25
3.2.3	Determining Optimal Values for m_c, n_c, k_c in the GAP8	27
3.3	Convolution Aproach	28
3.4	Experimental Results	29
3.5	Comparison with other GEMM Algorithms	31
3.6	Concluding Remarks	32
3.7	Summary	33
4	Performance Modeling of GEMM and Convolution in the GAP8	35
4.1	Introduction	35
4.2	A Performance Simulator	36
4.3	Performance Modeling of GEMM	37
4.3.1	Data transfers inside the micro-kernel	37
4.3.2	Data transfers outside the micro-kernel	38
4.3.3	Calibration	39
4.3.4	Validation	40
4.4	Performance Analysis of GEMM	41
4.5	Performance of the Convolution	43
4.5.1	Performance model	44
4.5.2	Performance analysis	45
4.5.3	Putting it all together: global analysis	47
4.6	Concluding Remarks and Summary	47
5	Parallel GEMM-based Convolution in the GAP8	51
5.1	Introduction	51
5.2	Tailoring GEMM for the GAP8	52
5.2.1	From sequential to parallel GEMM	52
5.2.2	Implementation details	52
5.2.3	Micro-kernel for B3C2A0	52
5.2.4	Data transfers across the memory hierarchy	53
5.2.5	Analysis of parallelization strategies	56
5.2.6	Selection of Optimal Loop for Parallelization	58
5.3	Performance Analysis	62
5.3.1	Experimental configuration	63
5.3.2	Preliminary analysis	63
5.3.3	Global comparison	66
5.3.3.1	Performance enhancements	68

5.3.4	IM2COL vs. IM2ROW transforms	68
5.4	Concluding Remarks	69
5.5	Summary	69
6	Communication-Avoiding Fusion of GEMM-based Convolutions	71
6.1	Introduction	71
6.2	Analysis of the Two-Stage Lowering Approach	72
6.2.1	Micro-kernel analysis	72
6.2.2	Data transfers outside the micro-kernel	74
6.2.3	Data transfers in IM2ROW	75
6.2.4	Validation	76
6.3	Fused Configurations for the Lowering Approach	77
6.3.1	Baseline convolution	79
6.3.2	Fusion IMPA: IM2ROW + Pack A	82
6.3.3	Fusion IMPA_OTF: IM2ROW + Pack A “on-the-fly”	84
6.3.4	Fusion UCIM: Unpack C + IM2ROW	86
6.3.5	Fusion FULL: Unpack C + IM2ROW + Pack A	88
6.3.6	Performance analysis	90
6.4	Concluding Remarks and Summary	92
7	Conclusions, Future Work	95
7.1	Future Work	98
	Acronyms	101
	A Computing with Submatrices	105
	B Vector instructions	107
	B.1 Overview of SIMD	107
	B.2 Application in Matrix Operations	107
	B.3 Benefits and Limitations	108
	C Matrix Multiplication Algorithms	109
	D Linear Algebra Libraries	113
	Bibliography	115