

# Contents

<b>Acknowledgements</b>	<b>v</b>
<b>Abstract (English/French)</b>	<b>vii</b>
<b>List of figures</b>	<b>xvi</b>
<b>List of tables</b>	<b>xvii</b>
<b>Nomenclature</b>	<b>xix</b>
<b>1 Introduction and related work</b>	<b>1</b>
1.1 Introduction	1
1.2 State of the art	4
1.2.1 Approximate circuits using technological knobs	5
1.2.2 Circuit-level design techniques	7
1.2.3 Software-level approximations	9
1.2.4 Discussion	10
<b>2 Speculative and pruned arithmetic circuits</b>	<b>11</b>
2.1 Inexact speculative adder (ISA)	12
2.1.1 Proposed architecture	12
2.1.2 COMP block implementation	15
2.1.3 Analysis of error compensation	16
2.1.4 The ISA design strategy	17
2.1.5 Results and comparison	19
2.2 Gate-level pruning (GLP)	24
2.2.1 Proposed framework	24
2.2.2 Results	26
2.3 Combination of speculation and pruning	28
2.3.1 Proposed method	28
2.3.2 Results and comparison	28
2.4 Overclocking of speculative adders	31
2.4.1 Proposed method for timing-error prediction	31
2.4.2 Combining structural and timing errors	32

2.4.3	Experimental study	34
2.4.4	Results	35
2.5	Conclusion	40
<b>3</b>	<b>Approximate circuits by fabrication of false timing paths</b>	<b>41</b>
3.1	Approximate circuit design and optimization by fabrication of false paths	42
3.1.1	False-path fabrication	42
3.1.2	Significance-driven cuts	43
3.2	Carry cut back adder (CCBA)	45
3.2.1	State-of-the-art approximate adders	45
3.2.2	Proposed architecture	46
3.2.3	Circuit timing	49
3.2.4	Arithmetic and errors	50
3.2.5	Worst-case error and floating-point precision	52
3.2.6	Design and implementation strategy	54
3.2.7	Design-space minimization	56
3.3	Results and comparison	58
3.3.1	Methodology	58
3.3.2	CCBA results	59
3.3.3	Comparative study	62
3.4	Conclusion	66
<b>4</b>	<b>Approximate accelerators and applications</b>	<b>67</b>
4.1	Approximate floating-point units	68
4.1.1	Inexact speculative multiplier (ISM)	69
4.1.2	Approximate FPU architecture	70
4.1.3	Measurement and results	71
4.2	Application to HDR image tone-mapping	74
4.2.1	Tone-mapping algorithm	74
4.2.2	Results	75
4.3	Conclusion and perspectives	78
<b>5</b>	<b>Precision-scalable multiply-accumulate units for neural-network processing</b>	<b>79</b>
5.1	Introduction	80
5.2	Considerations and methodology	81
5.2.1	Scalability by design and by data gating	81
5.2.2	Scope and methodology of the study	81
5.3	Multiply-accumulate units	82
5.3.1	Data-gated conventional MAC	82
5.3.2	DVAFS MAC	83
5.3.3	Divide-and-conquer strategy	84
5.3.4	Bit serial design	84

5.3.5	Multi-bit serial design	85
5.4	Results and comparative study	86
5.4.1	Precision-scaling energy breakdown	86
5.4.2	Comparative study	88
5.5	Conclusion	90
<b>Conclusions and perspectives</b>		<b>91</b>
<b>Appendix</b>		<b>95</b>
A	Characterization with adaptive sample-size inferential statistics (CASSIS)	95
A.1	CASSIS method	95
A.2	Experimental study	96
A.3	Results	97
A.4	Conclusion	99
B	Neural networks using logarithmic quantization	101
B.1	Log-quantized neural-network framework	101
B.2	Considerations	102
B.3	Results	102
B.4	Remarks and perspectives	103
<b>Bibliography</b>		<b>105</b>
<b>List of publications</b>		<b>113</b>
<b>Curriculum vitae</b>		<b>115</b>