

1. Record Nr.	UNINA9911007148703321
Autore	Yu Hao
Titolo	ReRAM-Based Machine Learning
Pubbl/distr/stampa	Stevenage : , : Institution of Engineering & Technology, , 2021 ©2021
ISBN	1-83724-558-4 1-5231-3657-X 1-83953-082-0
Edizione	[1st ed.]
Descrizione fisica	1 online resource (261 pages)
Collana	Computing and Networks
Altri autori (Persone)	NiLeibin DinakarraoSai Manoj Pudukotai
Disciplina	006.31
Soggetti	Machine learning
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Cover -- Contents -- Acronyms -- Preface -- About the authors -- Part I. Introduction -- 1 Introduction -- 1.1 Introduction -- 1.1.1 Memory wall and powerwall -- 1.1.2 Semiconductor memory -- 1.1.2.1 Memory technologies -- 1.1.2.2 Nanoscale limitations -- 1.1.3 Nonvolatile IMC architecture -- 1.2 Challenges and contributions -- 1.3 Book organization -- 2 The need of in-memory computing -- 2.1 Introduction -- 2.2 Neuromorphic computing devices -- 2.2.1 Resistive random-access memory -- 2.2.2 Spin-transfer-torque magnetic random-access memory -- 2.2.3 Phase change memory -- 2.3 Characteristics of NVM devices for neuromorphic computing -- 2.4 IMC architectures for machine learning -- 2.4.1 Operating principles of IMC architectures -- 2.4.1.1 In-macro operating schemes -- 2.4.1.2 Architectures for operating schemes -- 2.4.2 Analog and digitized fashion of IMC -- 2.4.3 Analog IMC -- 2.4.3.1 Analog MAC -- 2.4.3.2 Cascading IMC macros -- 2.4.3.3 Bitcell and array design of analog IMC -- 2.4.3.4 Peripheral circuitry of analog IMC -- 2.4.3.5 Challenges of analog IMC -- 2.4.3.6 Trade-offs of analog IMC devices -- 2.4.4 Digitized IMC -- 2.4.5 Literature review of IMC -- 2.4.5.1 DRAM-based IMCs -- 2.4.5.2 NAND-Flash-based IMCs -- 2.4.5.3 SRAM-based IMCs -- 2.4.5.4 ReRAM-based IMCs -- 2.4.5.5 STT-MRAM-based IMCs --

2.4.5.6 SOT-MRAM-based IMCs -- 2.5 Analysis of IMC architectures -- 3 The background of ReRAM devices -- 3.1 ReRAM device and SPICE model -- 3.1.1 Drift-type ReRAM device -- 3.1.2 Diffusive-type ReRAM device -- 3.2 ReRAM-crossbar structure -- 3.2.1 Analog and digitized ReRAM crossbar -- 3.2.1.1 Traditional analog ReRAM crossbar -- 3.2.1.2 Digitalized ReRAM crossbar -- 3.2.2 Connection of ReRAM crossbar -- 3.2.2.1 Direct-connected ReRAM -- 3.2.2.2 One-transistor-one-ReRAM -- 3.2.2.3 One-selector-one-ReRAM -- 3.3 ReRAM-based oscillator. 3.4 Write-in scheme for multibit ReRAM storage -- 3.4.1 ReRAM data storage -- 3.4.2 Multi-threshold resistance for data storage -- 3.4.3 Write and read -- 3.4.3.1 Write-in method -- 3.4.3.2 Readout method -- 3.4.4 Validation -- 3.4.5 Encoding and 3-bit storage -- 3.4.5.1 Exploration of the memristance range -- 3.4.5.2 Uniform input encoding -- 3.4.5.3 Nonuniform encoding -- 3.5 Logic functional units with ReRAM -- 3.5.1 OR gate -- 3.5.2 AND gate -- 3.6 ReRAM for logic operations -- 3.6.1 Simulation settings -- 3.6.2 ReRAM-based circuits -- 3.6.2.1 Logic operations -- 3.6.2.2 Readout circuit -- 3.6.3 ReRAM as a computational unit-cum-memory -- Part II. Machine learning accelerators -- 4 The background of machine learning algorithms -- 4.1 SVM-based machine learning -- 4.2 Single-layer feedforward neural network-based machine learning -- 4.2.1 Single-layer feedforward network -- 4.2.1.1 Feature extraction -- 4.2.1.2 Neural network-based learning -- 4.2.1.3 Incremental LS solver-based learning -- 4.2.2 L2-norm-gradient-based learning -- 4.2.2.1 Multilayer neural network -- 4.2.2.2 Direct-gradient-based L2-norm optimization -- 4.3 DCNN-based machine learning -- 4.3.1 Deep learning for multilayer neural network -- 4.3.2 Convolutional neural network -- 4.3.3 Binary convolutional neural network -- 4.3.3.1 Bitwise convolution -- 4.3.3.2 Bitwise batch normalization -- 4.3.3.3 Bitwise pooling and activation functions -- 4.3.3.4 Bitwise CNN model overview -- 4.4 TNN-based machine learning -- 4.4.1 Tensor-train decomposition and compression -- 4.4.2 Tensor-train-based neural network -- 4.4.3 Training TNN -- 5 XIMA: the in-ReRAM machine learning architecture -- 5.1 ReRAM network-based ML operations -- 5.1.1 ReRAM-crossbar network -- 5.1.1.1 Mapping of ReRAM crossbar for matrix-vector multiplication -- 5.1.1.2 Performance evaluation. 5.1.2 Coupled ReRAM oscillator network -- 5.1.2.1 Coupled-ReRAM-oscillator network for L2-norm calculation -- 5.1.2.2 Performance evaluation -- 5.2 ReRAM network-based in-memory ML accelerator -- 5.2.1 Distributed ReRAM-crossbar in-memory architecture -- 5.2.1.1 Memory-computing integration -- 5.2.1.2 Communication protocol and control bus -- 5.2.2 3D XIMA -- 5.2.2.1 3D single-layer CMOS-ReRAM architecture -- 5.2.2.2 3D multilayer CMOS-ReRAM architecture -- 6 The mapping of machine learning algorithms on XIMA -- 6.1 Machine learning algorithms on XIMA -- 6.1.1 SLFN-based learning and inference acceleration -- 6.1.1.1 Step 1. Parallel digitizing -- 6.1.1.2 Step 2. XOR -- 6.1.1.3 Step 3. Encoding -- 6.1.1.4 Step 4. Adding and shifting for inner-product result -- 6.1.2 BCNN-based inference acceleration on passive array -- 6.1.2.1 Mapping bitwise convolution -- 6.1.2.2 Mapping bitwise batch normalization -- 6.1.2.3 Mapping bitwise pooling and binarization -- 6.1.2.4 Summary of mapping bitwise CNN -- 6.1.3 BCNN-based inference acceleration on 1S1R array -- 6.1.3.1 Mapping unsigned bitwise convolution -- 6.1.3.2 Mapping batch normalization, pooling and binarization -- 6.1.4 L2-norm gradient-based learning and inference acceleration -- 6.1.4.1 Mapping matrix-vector multiplication on ReRAM-crossbar network -- 6.1.4.2 Mapping L2-norm calculation on coupled ReRAM oscillator network --

6.1.4.3 Mapping flow of multilayer neural network on ReRAM network -- 6.1.5 Experimental evaluation of machine learning algorithms on XIMA architecture -- 6.1.5.1 SLFN-based learning and inference acceleration -- 6.1.5.2 L2-norm gradient-based learning and inference acceleration -- 6.1.5.3 BCNN-based inference acceleration on passive array -- 6.1.5.4 BCNN-based inference acceleration on 1S1R array -- 6.2 Machine learning algorithms on 3D XIMA -- 6.2.1 On-chip design for SLFN. 6.2.1.1 Data quantization -- 6.2.1.2 ReRAM layer implementation for digitized matrix-vector multiplication -- 6.2.1.3 CMOS layer implementation for decoding and incremental least-squares -- 6.2.2 On-chip design for TNNs -- 6.2.2.1 Mapping on multilayer architecture -- 6.2.2.2 Mapping TNN on single-layer architecture -- 6.2.3 Experimental evaluation of machine learning algorithms on 3D CMOS-ReRAM -- 6.2.3.1 On-chip design for SLFN-based face recognition -- 6.2.3.2 Results of TNN-based on-chip design with 3D multilayer architecture -- 6.2.3.3 TNN-based distributed on-chip design on 3D single-layer architecture -- Part III. Case studies -- 7 Large-scale case study: accelerator for ResNet -- 7.1 Introduction -- 7.2 Deep neural network with quantization -- 7.2.1 Basics of ResNet -- 7.2.2 Quantized convolution and residual block -- 7.2.3 Quantized BN -- 7.2.4 Quantized activation function and pooling -- 7.2.5 Quantized deep neural network overview -- 7.2.6 Training strategy -- 7.3 Device for in-memory computing -- 7.3.1 ReRAM crossbar -- 7.3.2 Customized DAC and ADC circuits -- 7.3.3 In-memory computing architecture -- 7.4 Quantized ResNet on ReRAM crossbar -- 7.4.1 Mapping strategy -- 7.4.2 Overall architecture -- 7.5 Experiment result -- 7.5.1 Experiment settings -- 7.5.2 Device simulations -- 7.5.3 Accuracy analysis -- 7.5.3.1 Peak accuracy comparison -- 7.5.3.2 Accuracy under device variation -- 7.5.3.3 Accuracy under approximation -- 7.5.4 Performance analysis -- 7.5.4.1 Energy and area -- 7.5.4.2 Throughput and efficiency -- 8 Large-scale case study: accelerator for compressive sensing -- 8.1 Introduction -- 8.2 Background -- 8.2.1 Compressive sensing and isometric distortion -- 8.2.2 Optimized near-isometric embedding -- 8.3 Boolean embedding for signal acquisition front end -- 8.3.1 CMOS-based Boolean embedding circuit. 8.3.2 ReRAM crossbar-based Boolean embedding circuit -- 8.3.3 Problem formulation -- 8.4 IH algorithm -- 8.4.1 Orthogonal rotation -- 8.4.2 Quantization -- 8.4.3 Overall optimization algorithm -- 8.5 Row generation algorithm -- 8.5.1 Elimination of norm equality constraint -- 8.5.2 Convex relaxation of orthogonal constraint -- 8.5.3 Overall optimization algorithm -- 8.6 Numerical results -- 8.6.1 Experiment setup -- 8.6.2 IH algorithm on high-D ECG signals -- 8.6.2.1 Algorithm convergence and effectiveness -- 8.6.2.2 ECG recovery quality comparison -- 8.6.3 Row generation algorithm on low-D image patches -- 8.6.3.1 Algorithm effectiveness -- 8.6.3.2 Image recovery quality comparison -- 8.6.4 Hardware performance evaluation -- 8.6.4.1 Hardware comparison -- 8.6.4.2 Impact of ReRAM variation -- 9 Conclusions: wrap-up, open questions and challenges -- 9.1 Conclusion -- 9.2 Future work -- References -- Index -- Back Cover.

Sommario/riassunto

Serving as a bridge between researchers in the computing domain and computing hardware designers, this book presents ReRAM techniques for distributed computing using IMC accelerators, ReRAM-based IMC architectures for machine learning (ML) and data-intensive applications, and strategies to map ML designs onto hardware accelerators.