



HAL
open science

Powering AI at the edge: A robust, memristor-based binarized neural network with near-memory computing and miniaturized solar cell

Fadi Jebali, Atreya Majumdar, Clément Turck, Kamel-Eddine Harabi, Mathieu-Coumba Faye, Eloi Muhr, Jean-Pierre Walder, Oleksandr Bilousov, Amadéo Michaud, Elisa Vianello, et al.

► To cite this version:

Fadi Jebali, Atreya Majumdar, Clément Turck, Kamel-Eddine Harabi, Mathieu-Coumba Faye, et al.. Powering AI at the edge: A robust, memristor-based binarized neural network with near-memory computing and miniaturized solar cell. Journées Scientifiques Nationales 2024 du PEPR électronique et PEPR réseaux du futur, Mar 2024, Grenoble, France. hal-04699105

HAL Id: hal-04699105

<https://hal.science/hal-04699105v1>

Submitted on 16 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Powering AI at the Edge: A Robust, Memristor-based Neural Network with Miniaturized Solar Cell



F. Jebali¹, A. Majumdar², C. Turck², K.-E. Harabi², M.-C. Faye^{1,3}, E. Muhr¹, J.-P. Walder¹, O. Bilousov⁴, A. Michaud⁴, E. Vianello³, T. Hirtzlin³, F. Andrieu³, M. Bocquet¹, S. Collin⁴, D. Querlioz³, and J.-M. Portal¹

(1) Aix-Marseille Université, Université de Toulon, CNRS, IM2NP, Marseille, France (3) Université Grenoble Alpes, CEA, LETI, Grenoble, France
(2) Université Paris-Saclay, CNRS, Centre de Nanosciences et de Nanotechnologies, Palaiseau, France (4) Institut Photovoltaïque d'Ile-de-France (IPVF), Palaiseau, France

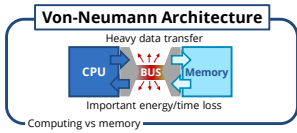


Nature Communications 15, 741 (2024). <https://doi.org/10.1038/s41467-024-44766-6>

Deep neural networks

≈20W **VS.** of self-driving car ~2kW
of language models training GPT-3 ~1.287 GWh

Neuromorphic applications are data intensive:
Meet the Von-Neumann bottleneck

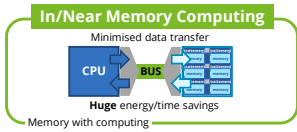


Operation	Energy
Sum (Fixed Point)	1x
Value Access (onchip cache)	60x
Value Access (offchip RAM)	3500x

Pedram, et al., IEEE D&T 2016

Inference on edge device:

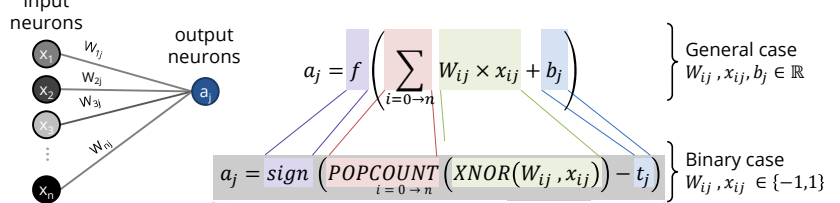
Near Memory Computing for energy harvesting



Our approach for Inference on edge device

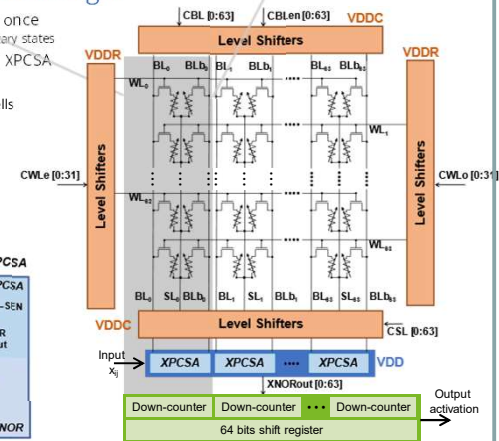
Binary Neural Network based on HfO_x RRAM (Memristor) Technology

⇒ Limit drastically the memory footprint (binary coding)



Desing to a robust solution under low voltages

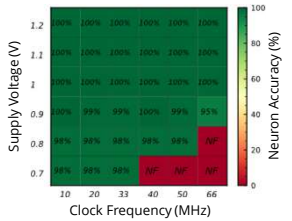
- Weights W_{ij} are stationary and programmed only once
 - Each weight (W_{ij}) is stored in a 2T2R bitcell using complementary states
- Augmented PCSA with XNOR operation is named XPCSA
- Mixed design with digital on top
 - 4x8k RRAM array (32k RRAM) & 10k standard cells
- AI digital controller is embedded
- Up to 128 neurons by 128 synapses
- Manufactured on 130nm + CEA-Leti RRAM
 - RRAM between M4 and M5
 - Area = 7.5 μm²



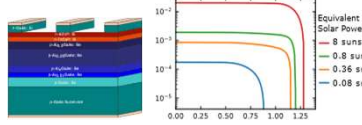
Robustness Assessment

Functional under a wide range of frequency and voltage w/o any calibration.

Measurement of the neuron accuracy for a random pattern



Direct Powering with miniature III-V solar cell (optimized for indoor/low-light conditions)

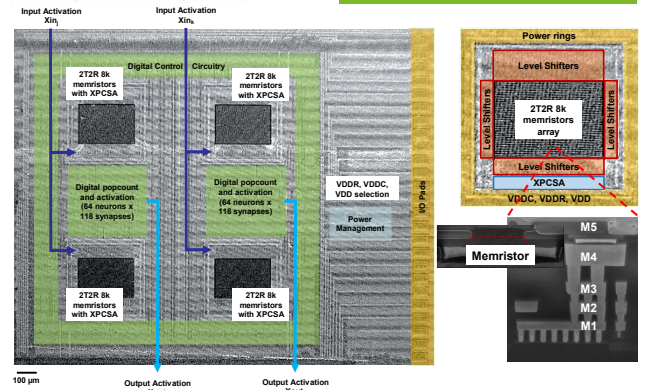
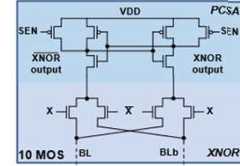


Equivalent Solar Power	MNIST Accuracy	CIFAR-10 Accuracy
Baseline	97.2%	86.6%
8 suns	97.1%	83.6%
0.8 suns	96.9%	78.2%
0.36 suns	96.9%	78.3%
0.08 suns	96.5%	73.4%

Under low illumination, little energy is available. AI does not fail, but gradually becomes less accurate.

	This work	Wan et al. Nature 2022	Xue et al. Nat. Electron. 2021	Jung et al. Nature 2022	Khaddam et al. IEEE SCS 2022
Device	HfO _x /Ti	HfO _x /TaO _x	Proprietary RRAM	MRAM	PCM
CMOS node	130 nm	130 nm	22nm	28nm	14nm
Unit cell	2T2R	1T1R	1T1R	2T2R	8T4R
Levels per cell	SLC	Analog	SLC	1-4	Analog
Weight bit width	1	Analog + ADC	1-4	1	Analog
Read circuit	XPCSA	Analog + ADC	Sense amplifier	Analog + TDC	Analog + ADC
Multi/Accu	Digital/Digital	Analog/Analog	Analog/Analog	Digital/Analog	Analog/Analog
Inference voltage	Flexible (0.7-1.2V)	Predet. Yes	Predet. Yes	Predet. Yes	Predet. Yes
Need for calibration	No	Yes	Yes	Yes	Yes
Reported energy efficiency (TOPS/W)	2.9 (measured) 22.5 (clock-gated) 397 (28nm-projection)	7 to 43	37 to 126	262 to 405	10.5

XPCSA = XNOR + PCSA



Take away message

- AI at the edge needs near-memory (NMC) or in-memory (IMC) computing solutions
- Analog IMC needs calibration and supports a single operating point: incompatible with energy harvesting.
- Our digital NMC is far more stable on a large range of operating points and is compatible with energy harvesters. When little energy is available (e.g., 0.08 suns), it does not fail but gracefully adjusts its accuracy.

