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.

BEP

L'électronique pour le calcul

Projet ciblé BEP

FRANCE 2030 — PROGRAMME DE RECHERCHE ÉLECTRONIQUE

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

F. Jebali[1], A. Majumdar[2], C. Turck[2], K.-E. Harabi[2], M.-C. Faye[1,3], E. Muhr[1], J.-P. Walder[1], O. Bilousov[4], A. Michaud[4], E. Vianello[3], T. Hirtzlin[3], F. Andrieu[3], M. Bocquet[1], S. Collin[4], D. Querlioz[3], and J.-M. Portal[1]

(1) Aix-Marseille Université, Université de Toulon, CNRS, IM2NP, Marseille, France
(2) Université Paris-Saclay, CNRS, Centre de Nanosciences et de Nanotechnologies, Palaiseau, France
(3) Université Grenoble Alpes, CEA, LETI, Grenoble, 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

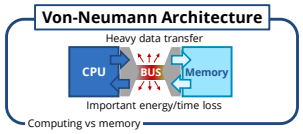
Deep Neural Networks

≈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

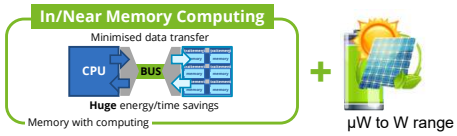


Operation	Energy
Sum (Fixed Point)	1x
Value Access (onchip cache)	60x
Value Access (offchip RAM)	3500x

Pedram, *et al.*, IEEE D&T 2016

Inference on edge device:

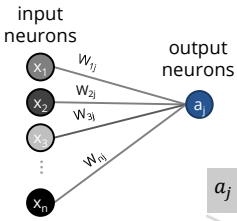
Near Memory Computing for energy harvesting



Our Approach for Inference on edge device

Binary Neural Network based on HfO_x RRAM (Memristor) Technology

➪ *Limit drastically the memory footprint (binary coding)*

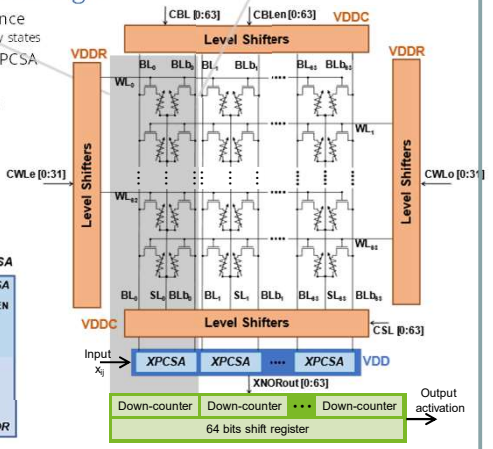


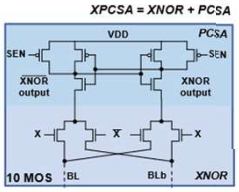
$$a_j = f\left(\sum_{i=0\to n} W_{ij} \times x_{ij} + b_j\right)$$ General case $W_{ij}, x_{ij}, b_j \in \mathbb{R}$

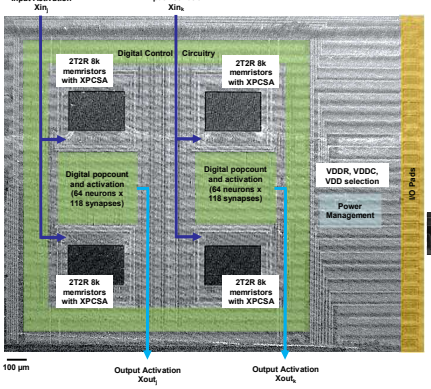
$$a_j = sign\left(POPCOUNT_{i=0\to n}\left(XNOR\left(W_{ij}, x_{ij}\right)\right) - t_j\right)$$ Binary case $W_{ij}, x_{ij} \in \{-1,1\}$

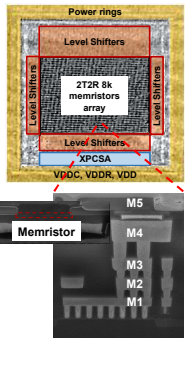
Desing to a robust solution under low voltages

- 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 mm²





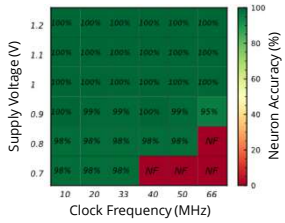




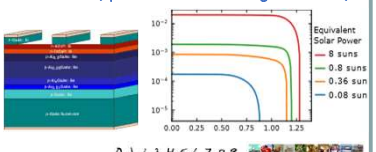
Robustness Assessment

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 JSSC 2022
Device CMOS node	HfOx/Ti 130 nm	HfOx/TaOx 130 nm	Proprietary RRAM 22nm	MRAM 28nm	PCM 14nm
Unit cell Levels per cell Weight bit width Read circuit	2T2R SLC 1 XPCSA	1T1R Analog Analog Analog + ADC	1T1R SLC 1-4 Sense amplifier	2T2R SLC 1 Analog + TDC	8T4R Analog Analog Analog + ADC
Multi/Accu	Digital/Digital	Analog/Analog	Analog/Analog	Digital/Analog	Analog/Analog
Inference voltage Need for calibration	**Flexible (0.7-1.2V) No**	**Predet. Yes**	**Predet. Yes**	**Predet. Yes**	**Predet. 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

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.

ANR-22-PEEL-00010