Reinventing Electronics for Energy-Efficient and Trustworthy Embedded AI

Damien Querlioz
Centre de Nanosciences et de Nanotechnologies
Directeur de recherche, CNRS
Professeur attaché, SaclAI-School
damien.querlioz@universite-paris-saclay.fr

How Far Will AI Energy Consumption Grow?

REUTERS®

World Business Markets Sustainability Legal More

Technology

OpenAI CEO Altman says at Davos future AI depends on energy breakthrough

Reuters

January 16, 2024 6:39 PM GMT+1 · Updated 2 months ago

Aa

Sam Altman, CEO of OpenAI, attends the Asia-Pacific Economic Cooperation (APEC) CEO Summit in San Francisco, California, U.S. November 16, 2023. REUTERS/Carlos Barria/File Photo *Purchase Licensing Rights*

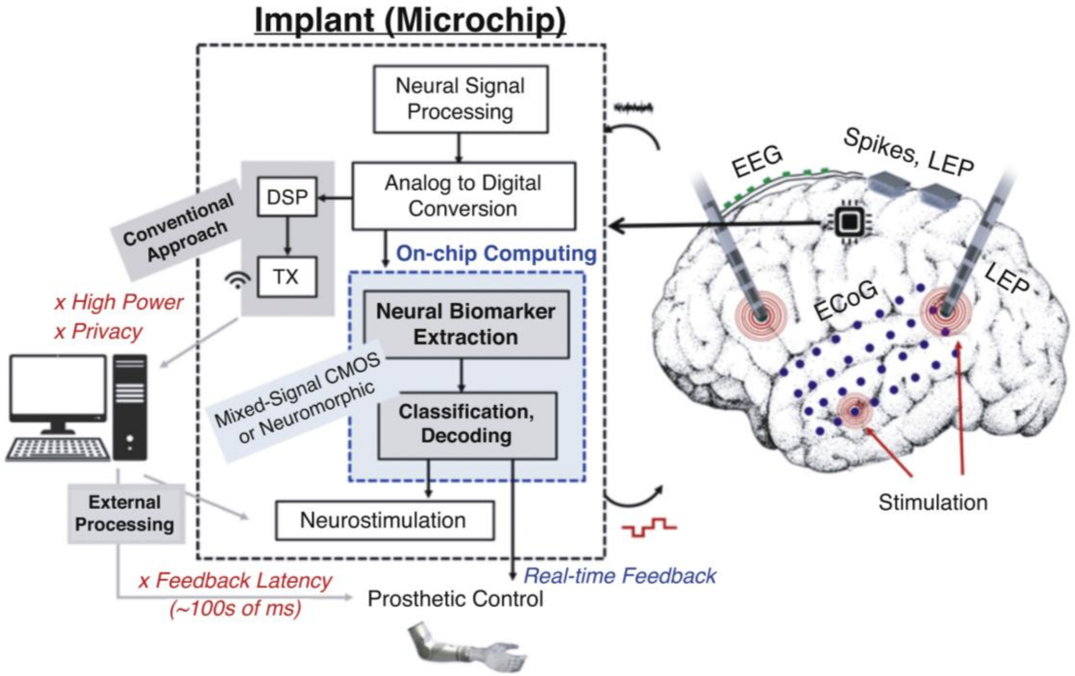
DAVOS, Switzerland, Jan 16 (Reuters) - OpenAI's CEO Sam Altman on Tuesday said an energy breakthrough is necessary for future artificial intelligence, which will consume vastly more power than people have expected.

Speaking at a Bloomberg event on the sidelines of the World Economic Forum's annual meeting in Davos, Altman said the silver lining is that more climate-friendly sources of energy, particularly nuclear fusion or cheaper solar power and storage, are the way forward for AI.

Or can we make AI *dramatically* more energy efficient?

cnrs université PARIS-SACLAY

AI Energy Inefficiency Limits Its Best Applications



Yoo and Shoaran,
Current Opinion Biotechnol., 2021

- **Edge AI implemented in a medical implant could allow epilepsy prediction, advanced BMI…**
- **But currently limited to *elementary* AI**

The AI Energy Problem Is a *Memory* Problem

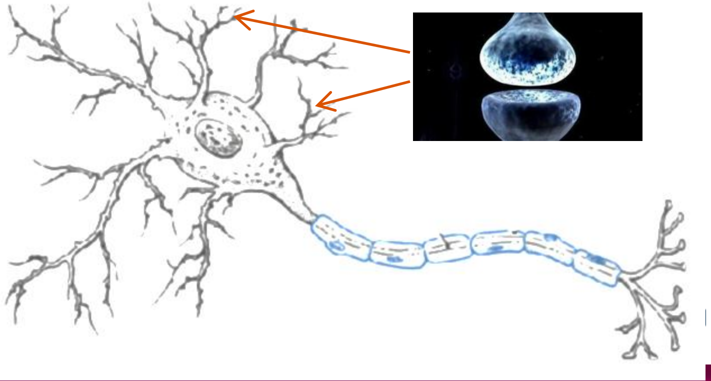
- Neural networks: not a lot of arithmetics,
  but **Huge volume of parameters**

- **In computers, GPUs, and AI accelerators developed by industry (Google TPU, Apple NPU…), memory access is extremely costly**

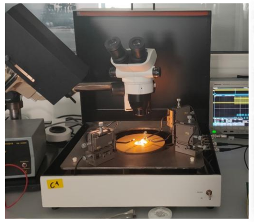
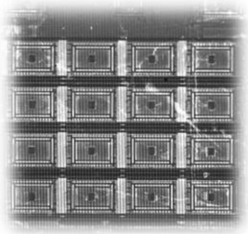
Operation	Energy
Addition of data (fixed point)	1x
~~Access data (onchip cache)~~	~~60x~~
~~Access data (offchip main memory)~~	3500x

Pedram et al , IEEE D&T 2016

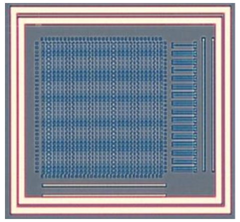
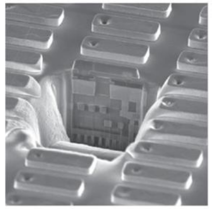
Absent in the brain!



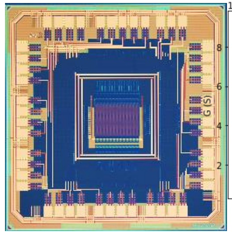
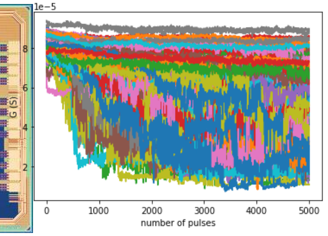
Toward Energy-Efficient and Trustworthy Embedded AI: The In-Memory Computing Approach

- Cointegrating Logic & Memory for Low-Energy AI

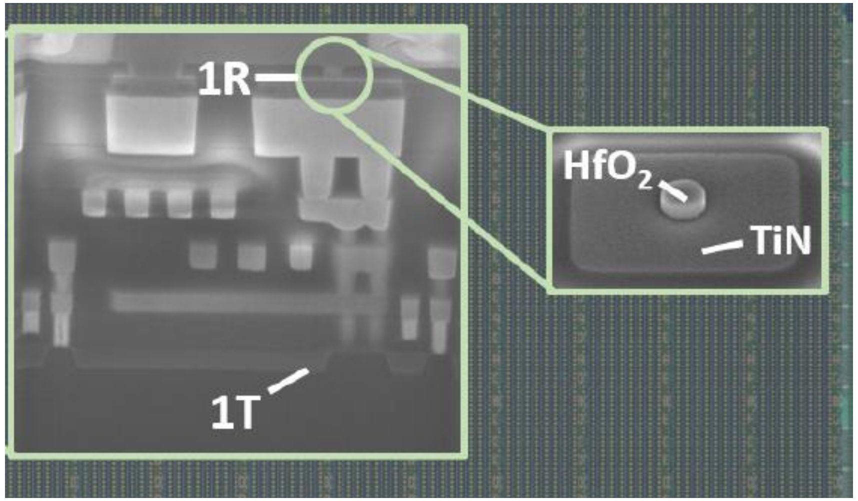



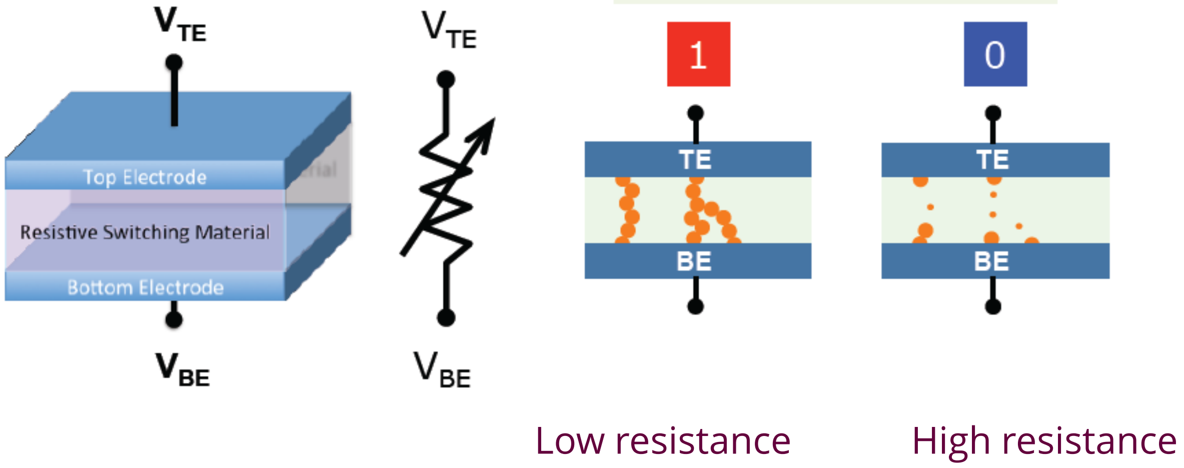
- The Analog & Stochastic Paths

- Learning with Physics

Memristor/RRAM: an Artificial Synapse!





- **TiN/HfO$_x$/Ti/TiN stack**

High voltage: move atoms to switch memristor between low/high resistance

Low voltage: allows reading the resistance

These New Memories Already Have a Market: Microcontroller Applications (NOR Flash replacement)

Products Applications Design Support Community About Infineon Careers

Press General Information Press Releases Market News Press Kits Media Pool Events Contacts

› Home › About Infineon › Press › Market News › Infineon and TSMC to introduce RRAM technology for automotive AURIX™ TC4x product family

Infineon and TSMC to introduce RRAM technology for automotive AURIX™ TC4x product family

Nov 25, 2022 | Market News

Munich, Germany – 25 November, 2022 – Infineon Technologies AG (FSE: IFX / OTCQX: IFNNY) and TSMC today announced the companies are preparing to introduce TSMC's Resistive RAM (RRAM) Non-Volatile Memory (NVM) technology into Infineon's next generation AURIX™ microcontrollers (MCU).

Embedded Flash microcontrollers have been the main building blocks of automotive electronic control units (ECU) since the introduction of the first engine management systems. They are essential components for clean, safe and smart cars, used in propulsion systems, vehicle dynamics control, driver assistance and body applications. They enable major innovations in the automotive space with regards to electrification, new E/E architectures and automated driving. Currently, the majority of MCU families in the market are based on embedded Flash memory technology. RRAM is a next step in embedded memories that allows to further scale to 28nm and beyond.

The Infineon AURIX TC4x microcontroller products combine performance extension with latest trends in virtualization, security, and networking features to enable the next generation of software-defined vehicles and new E/E architectures. TSMC and Infineon successfully created the basis for introduction of RRAM in the automotive domain. This will put AURIX microcontrollers on a broader

ElectronicsWeekly.com

NEWS BUSINESS MARKETS DESIGN PRODUCTS BLOGS

Home » News » Business Add to Bookmarks

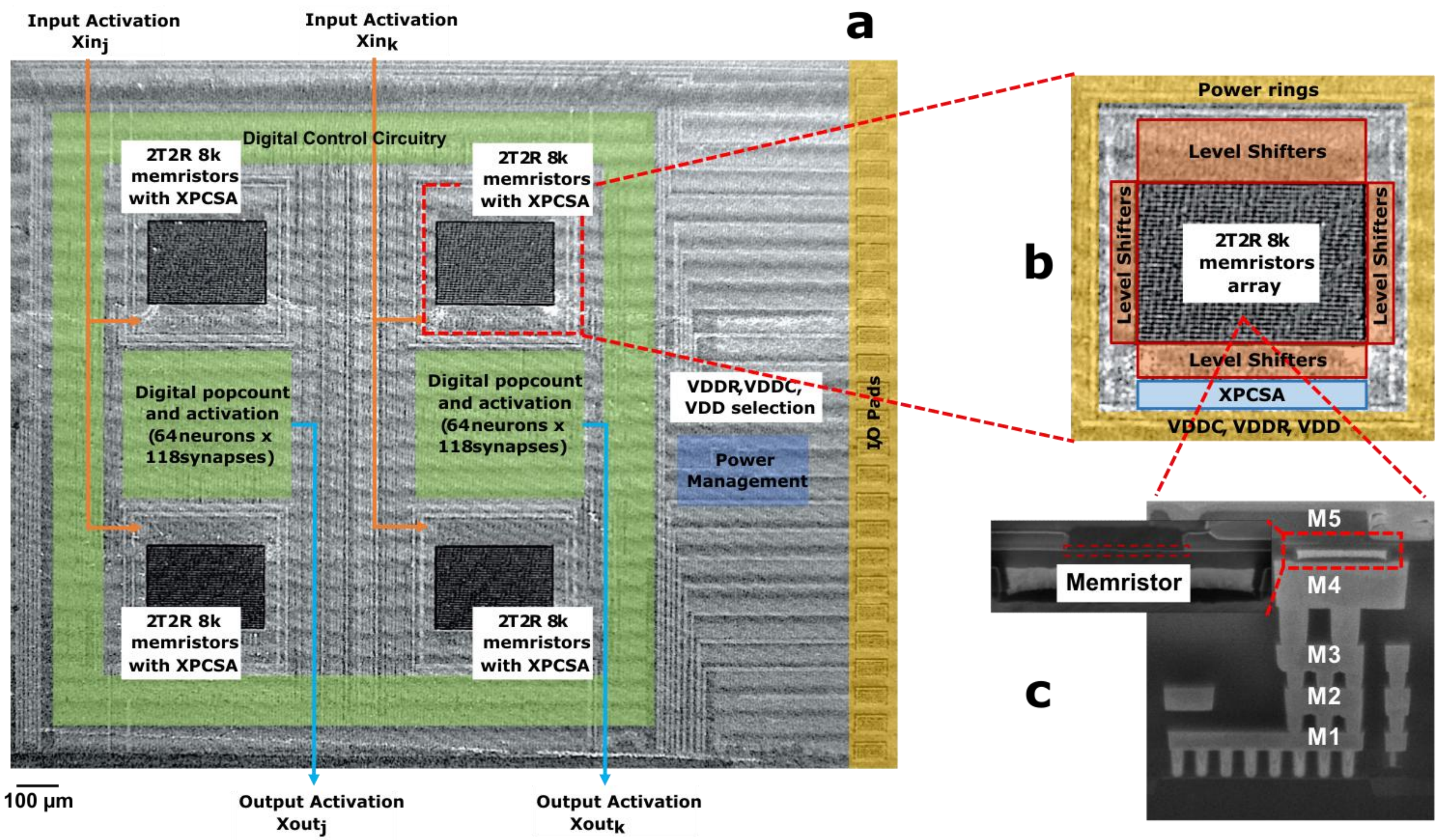
By David Manners 20th March 2024

ST to add FD-SOI/ePCM 32bit MCU for embedded applications

In H2, ST will sample an STM32 MCU for industrial applications made on an 18nm Samsung FD-SOI process with embedded phase change (ePCM) memory. Volume production is planned for H2 2025.

We repurpose them as *magic memory* for AI!

A Super-Robust Digital Binarized Neural Network

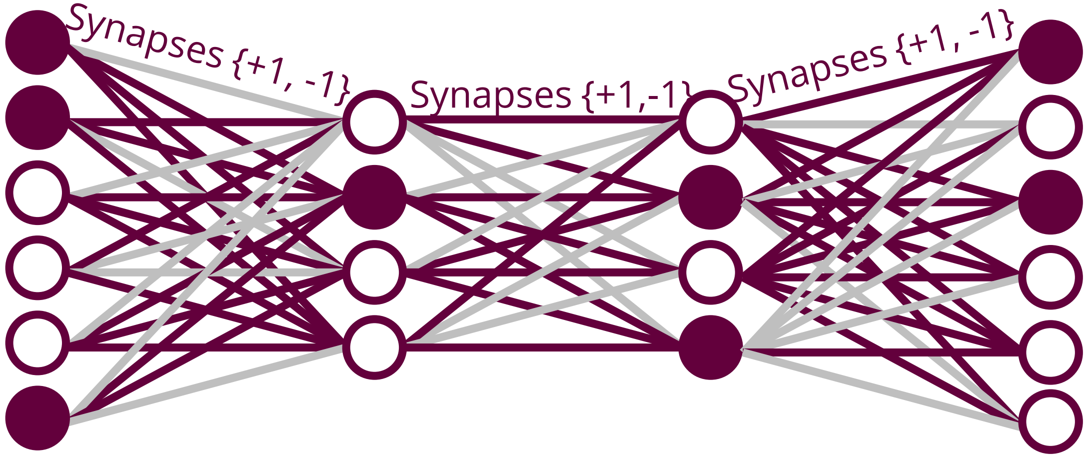


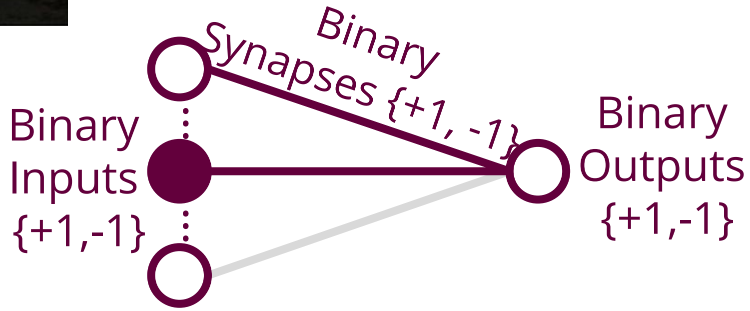
Collaboration

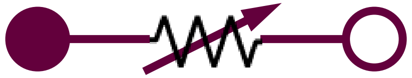
Jebali et al, Nature Communications 15, 741 (2024)

Binarized Neural network:
A Low Precison Neural Networks

Hubara, Courbariaux et al. NIPS 2016
Yoshua Bengio's group

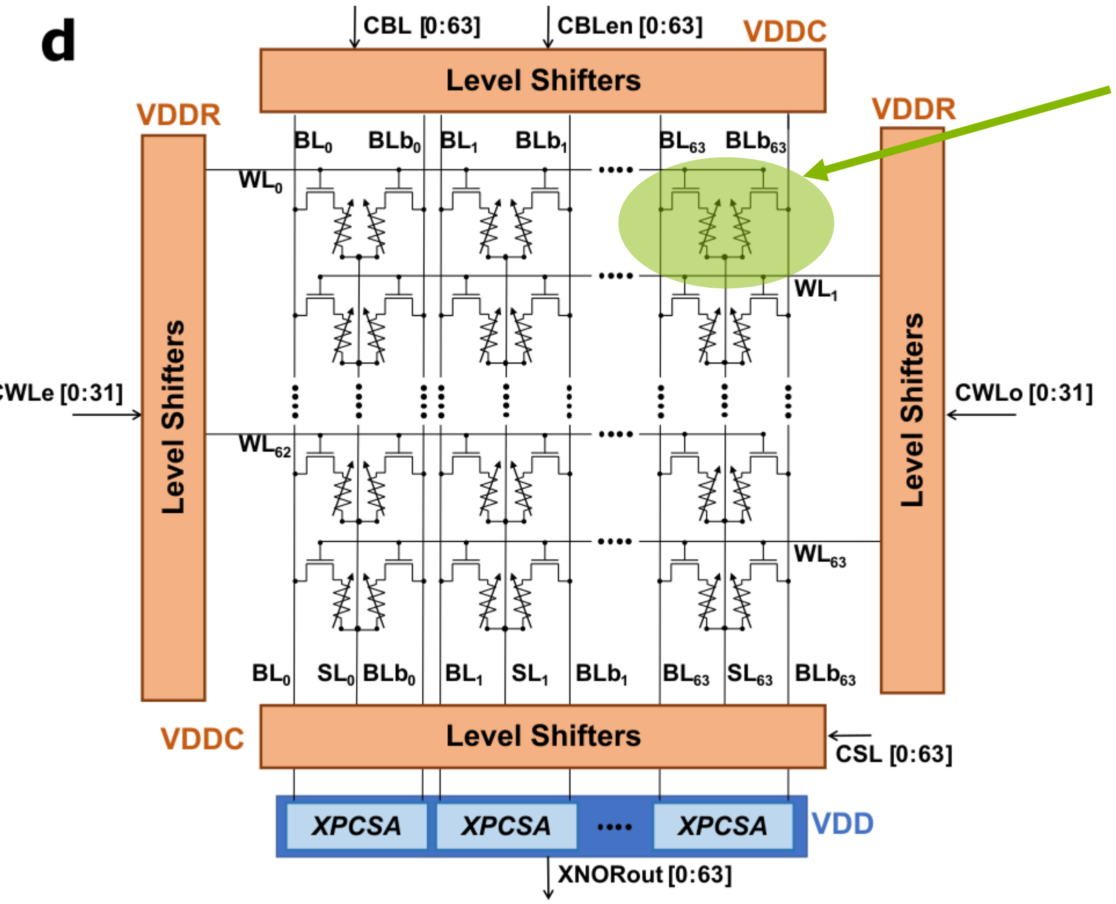




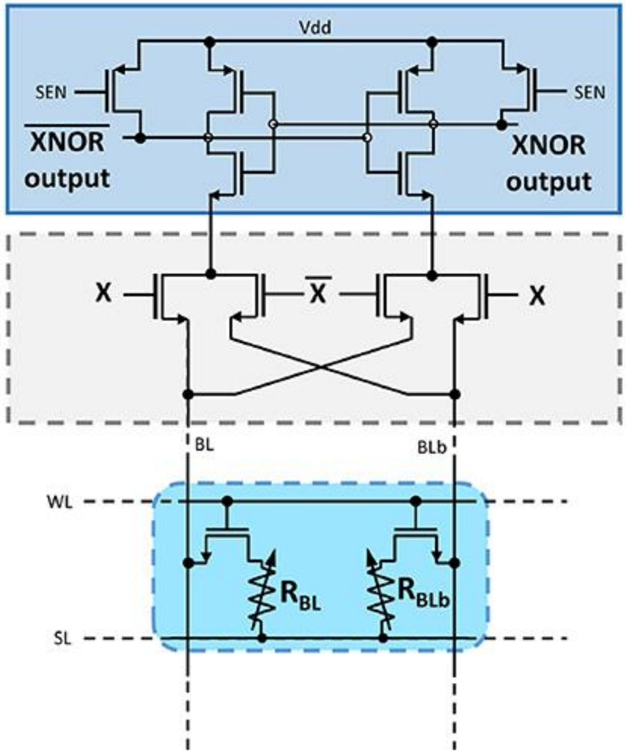


Our work :
Memristor as binary synapse

Robust Logic-In-Memory Feature: Multiplication Within the Sense Amplifier



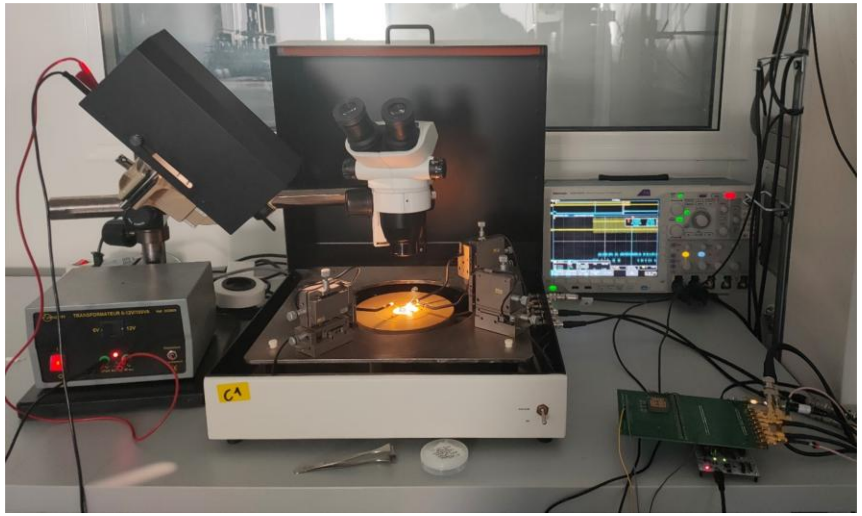
One weight: Programmed as 0/1 or 1/0



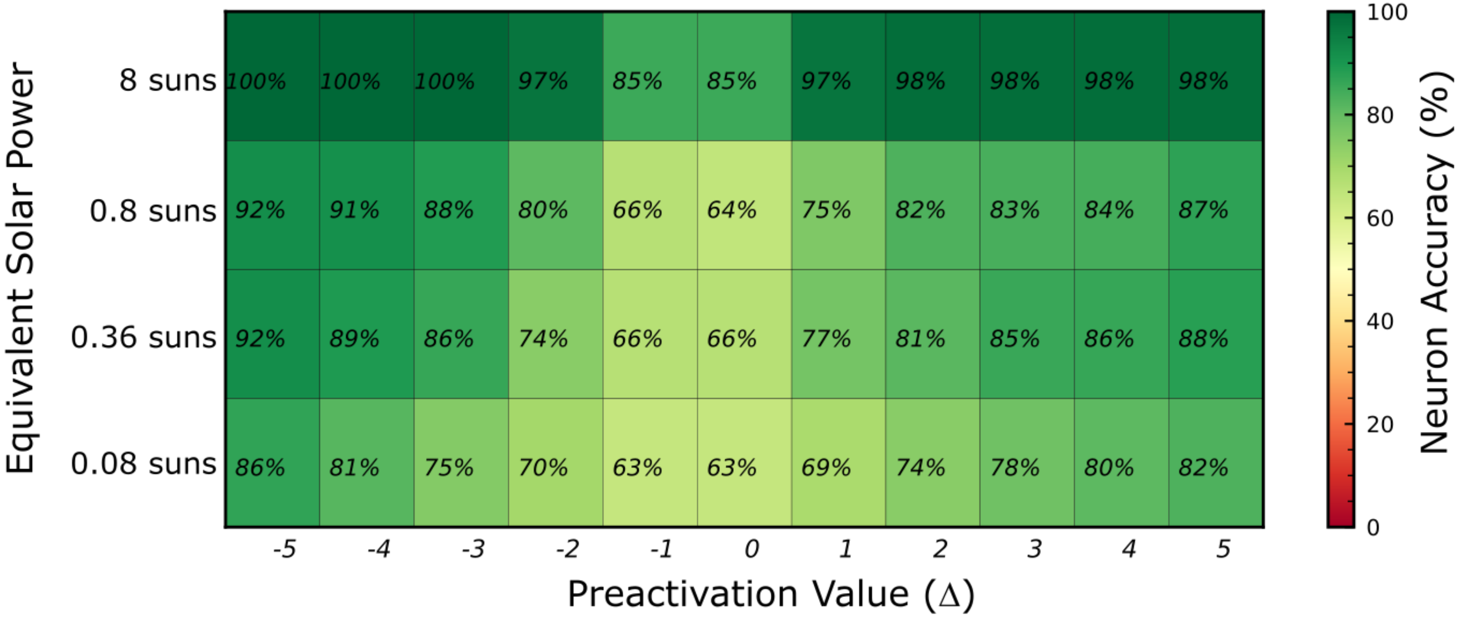
Weisheng Zhao, … Klein, Querlioz …., *IEEE TCAS I:* 2014.
IEEE Guillemin-Cauer Best Paper Award

Jebali et al, Nature Communications 15, 741 (2024)

Self-Powered Operation



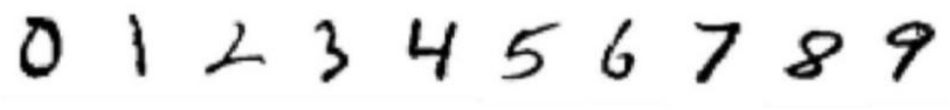
Miniature wide-bandgap solar cell



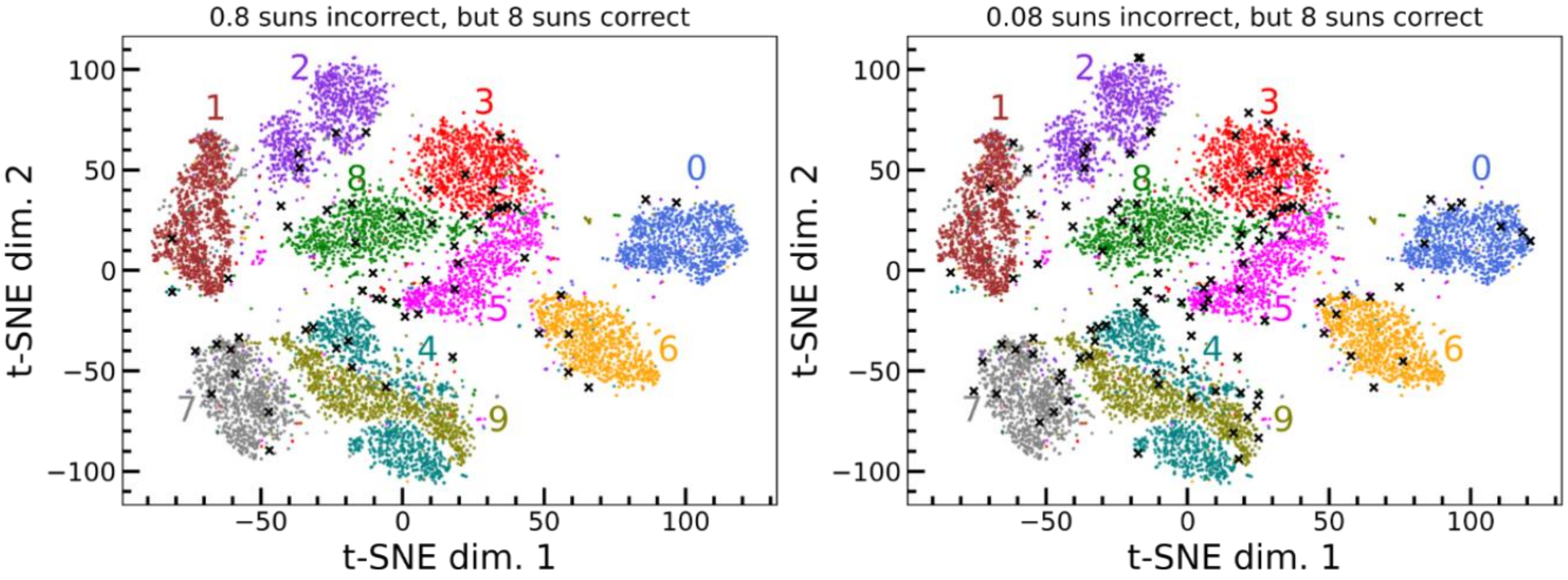
Neuron value $t_j = f\left(\sum w_{ij} s_i\right)$

Under low illumination the circuit naturally switches to approximate computation

Self-Powered Operation



Illumination	8 suns	0.8 suns	0.08 suns
Accuracy	97.2%	96.9%	96.5%



Jebali et al, Nature Communications 15, 741 (2024).

Computing In-Near Memory Is Applicable to Other AI Approaches

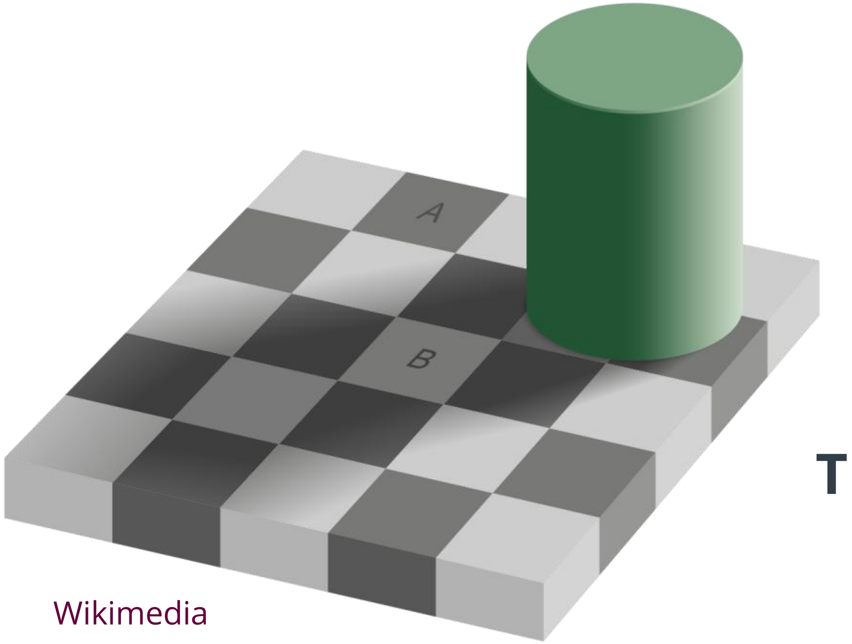
Thomas Bayes

Likelihoods *Prior*

$$p(Disease \mid Observations) \propto \underbrace{p(Observations \mid Disease) \times p(Disease)}$$

Constructed with expert knowledge+Data

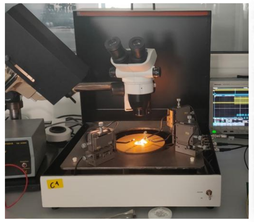
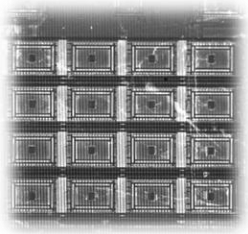
Hard to Compute



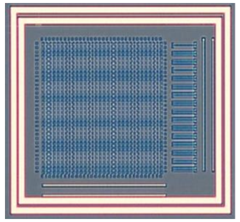
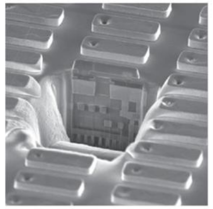
Wikimedia

The Brain Might Be Using Bayesian Reasoning

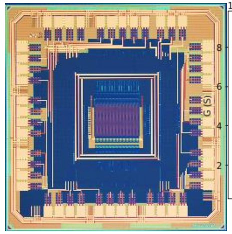
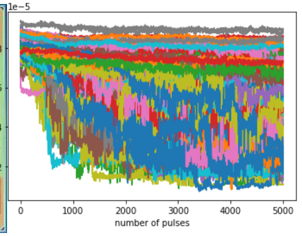
Toward Energy-Efficient and Trustworthy Embedded AI: The In-Memory Computing Approach

➢ Cointegrating Logic & Memory for Low-Energy AI

➢ The Analog & Stochastic Paths

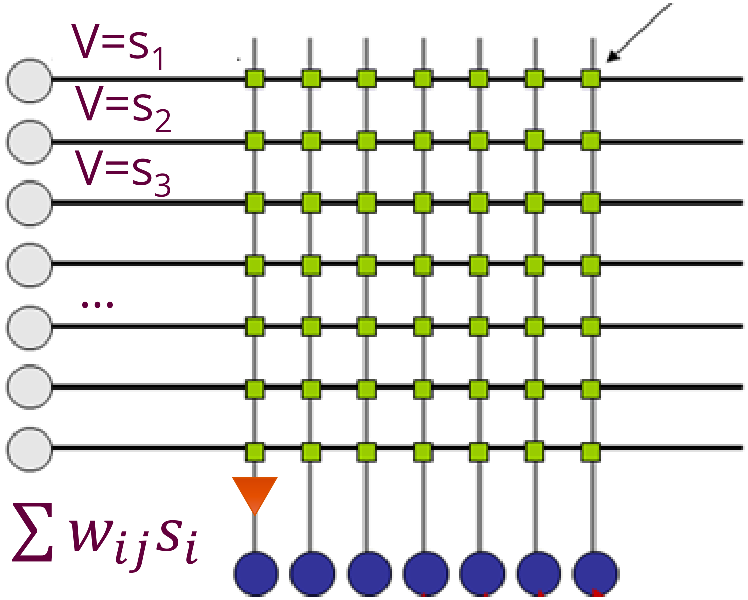



➢ Learning with Physics

Analog In-Memory Computing Performs Neural Network Inference Very Naturally

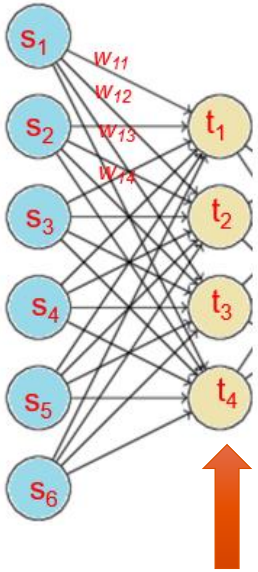
A matrix of analog memristors naturally implements a layer of neural network with **Ohm’s and Kirchhoff’s laws**!

Memristor conductance G = synaptic weight w



$I_1 = t_1 = \sum w_{ij} s_i$

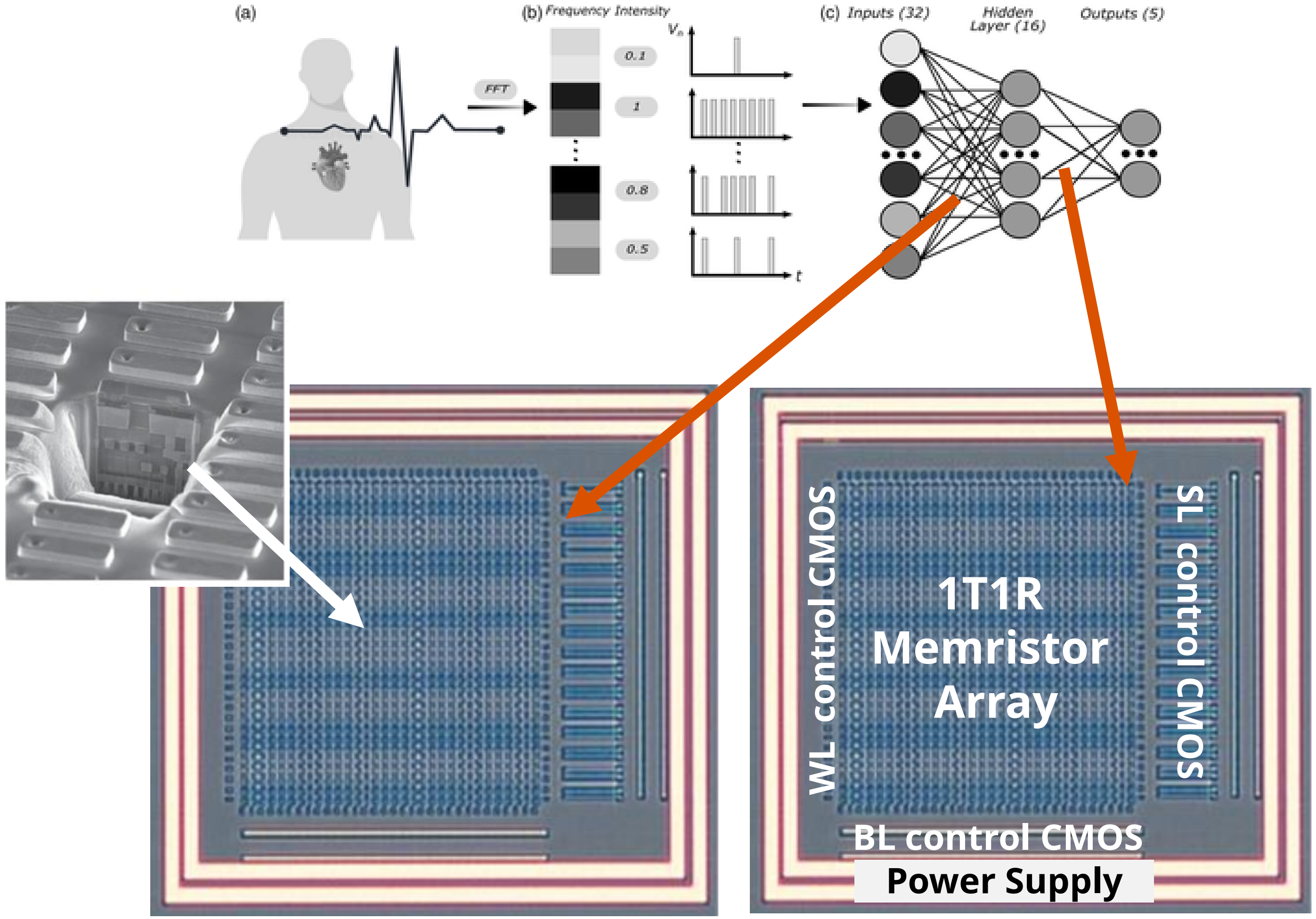
$\equiv$



Neuron value

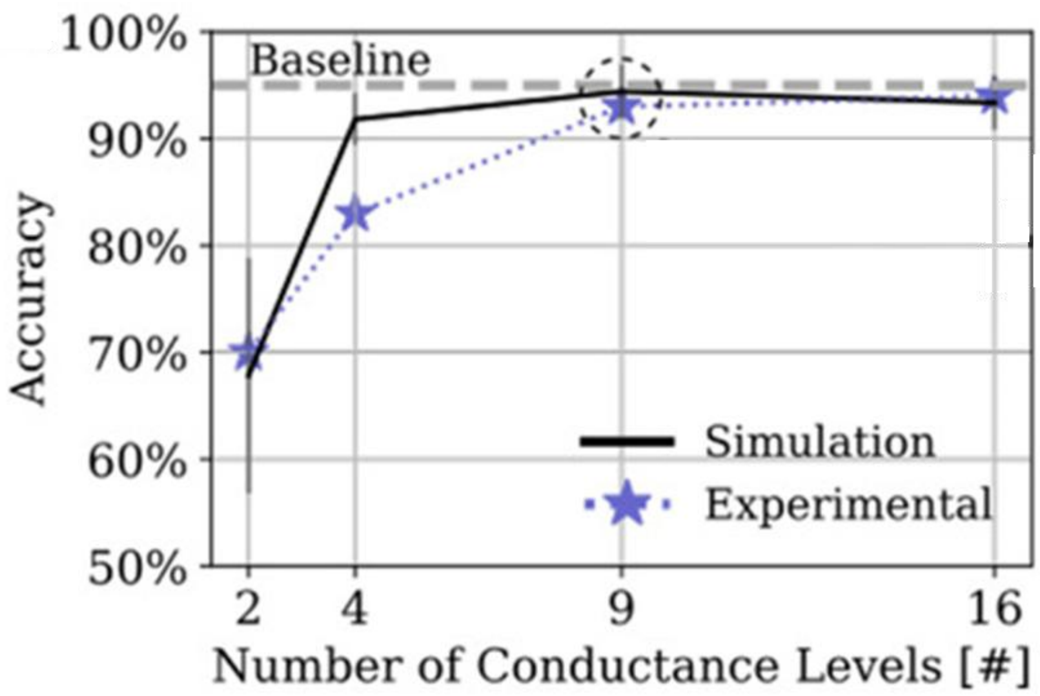
$t_j = f\left(\sum w_{ij} s_i\right)$

Analog In-Memory Computing with Memristors



HfO_x memristor integrated in BEOL of 130 nm-CMOS

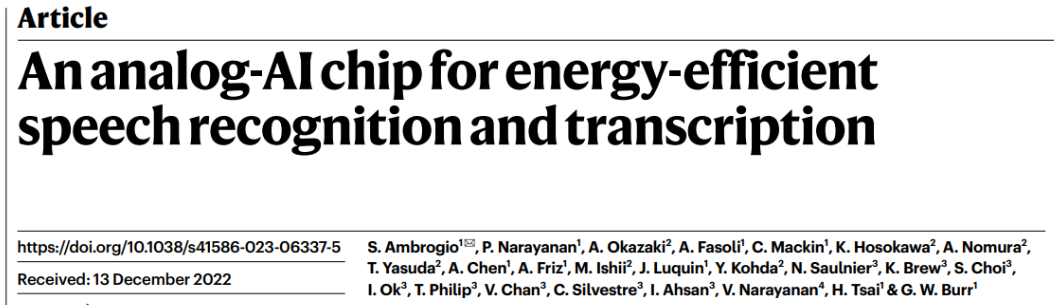
Accuracy on arrhythmia identification



Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

Collaboration

Current State of the Art (IBM Research, Almaden)

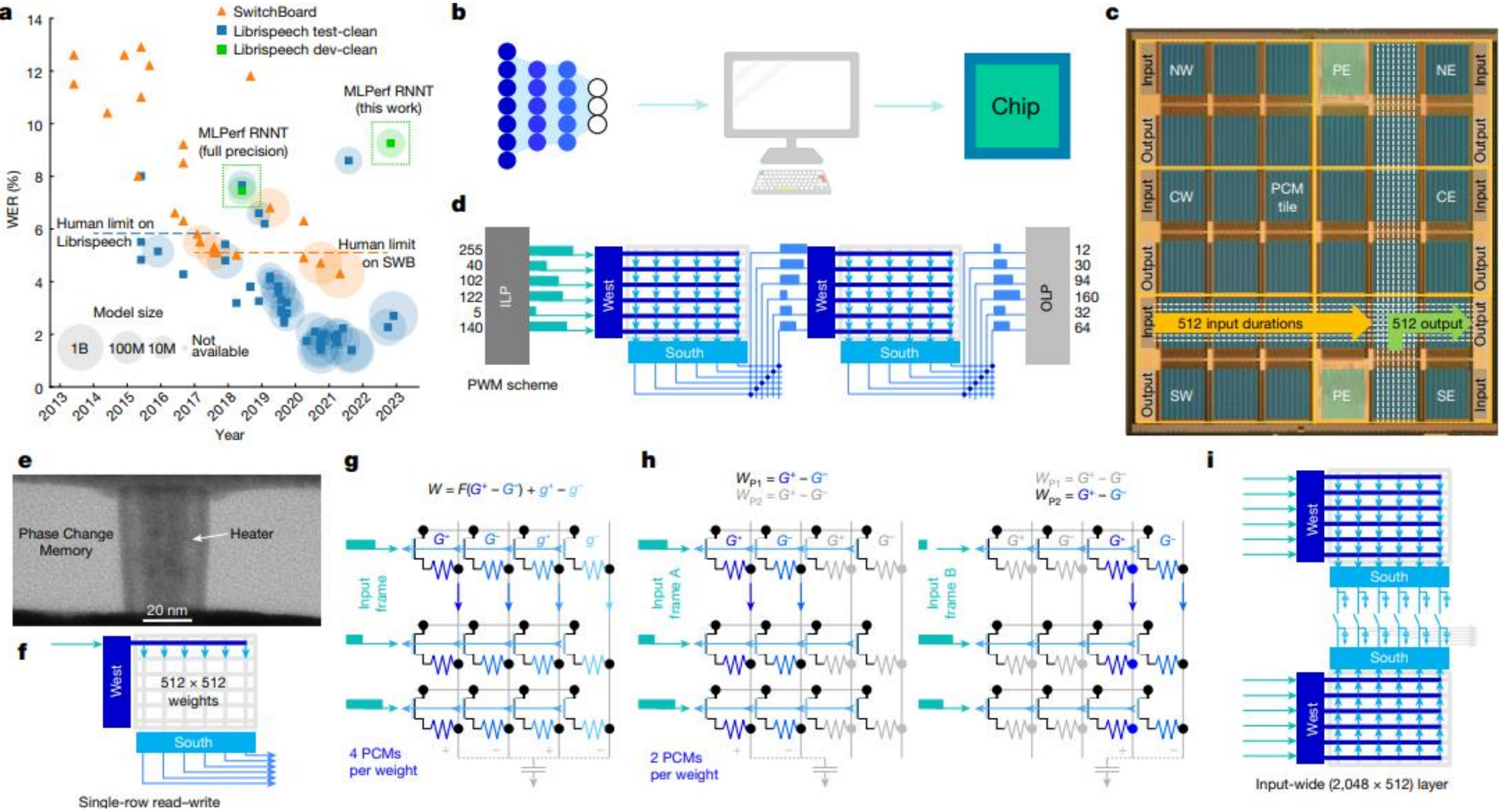

Article

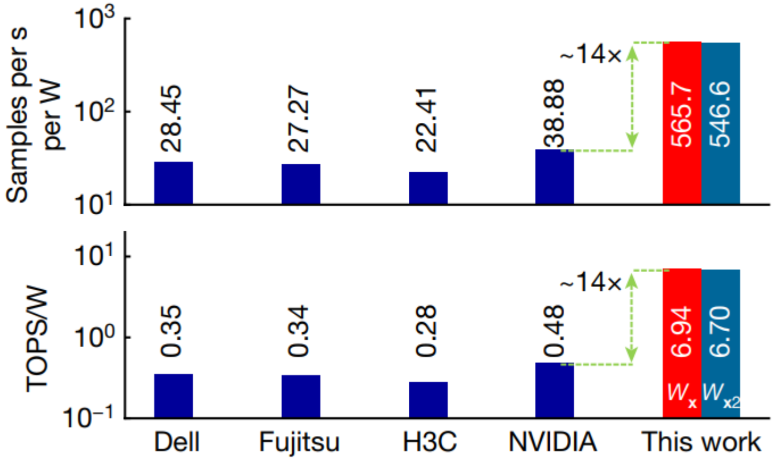
An analog-AI chip for energy-efficient speech recognition and transcription

https://doi.org/10.1038/s41586-023-06337-5

Received: 13 December 2022

S. Ambrogio[1✉], P. Narayanan[1], A. Okazaki[2], A. Fasoli[1], C. Mackin[1], K. Hosokawa[2], A. Nomura[2], T. Yasuda[2], A. Chen[1], A. Friz[1], M. Ishii[2], J. Luquin[1], Y. Kohda[2], N. Saulnier[3], K. Brew[3], S. Choi[3], I. Ok[3], T. Philip[3], V. Chan[3], C. Silvestre[3], I. Ahsan[3], V. Narayanan[4], H. Tsai[1] & G. W. Burr[1]

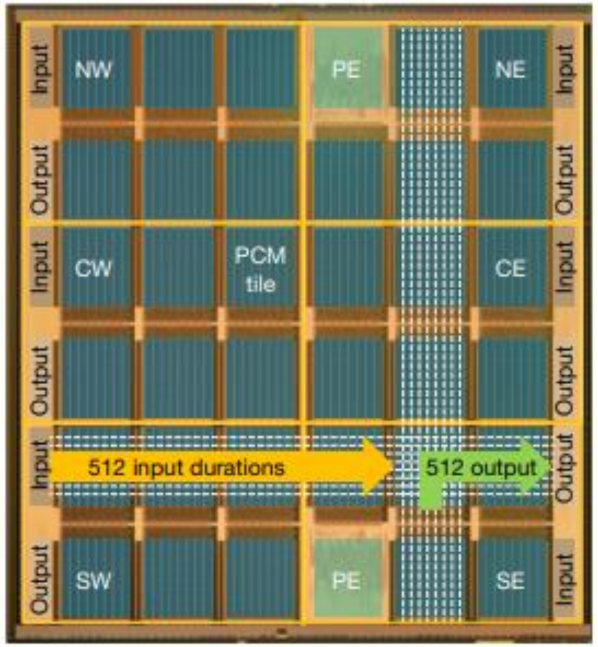


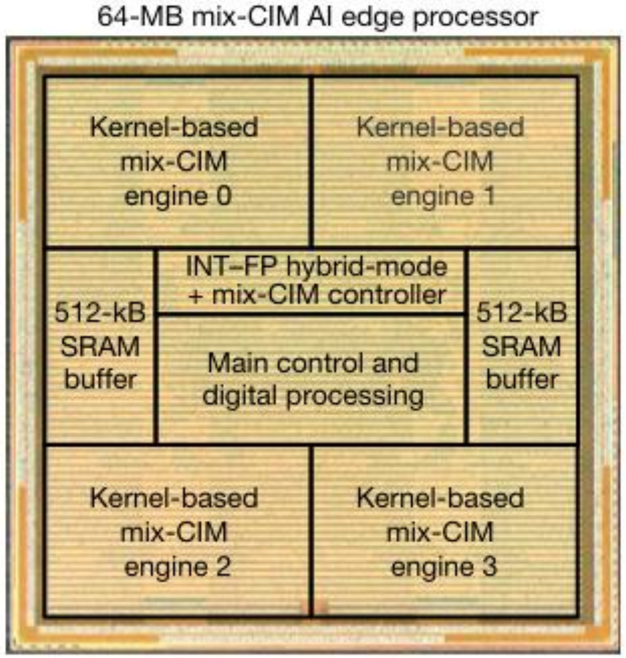


Despite prototype nature, 14x more energy efficient than any AI accelerator

- 35M phase change memories

Current State of the Art





IBM Research
LSTM Language model
35M memristors (phase change)
Ambrogio et al, Nature 2023

TSMC Research/NTHU
MobileNet-v2
64M memristors
Khwa et al, Nature 2025

7 TOPS/W
(140 fJ/operation)

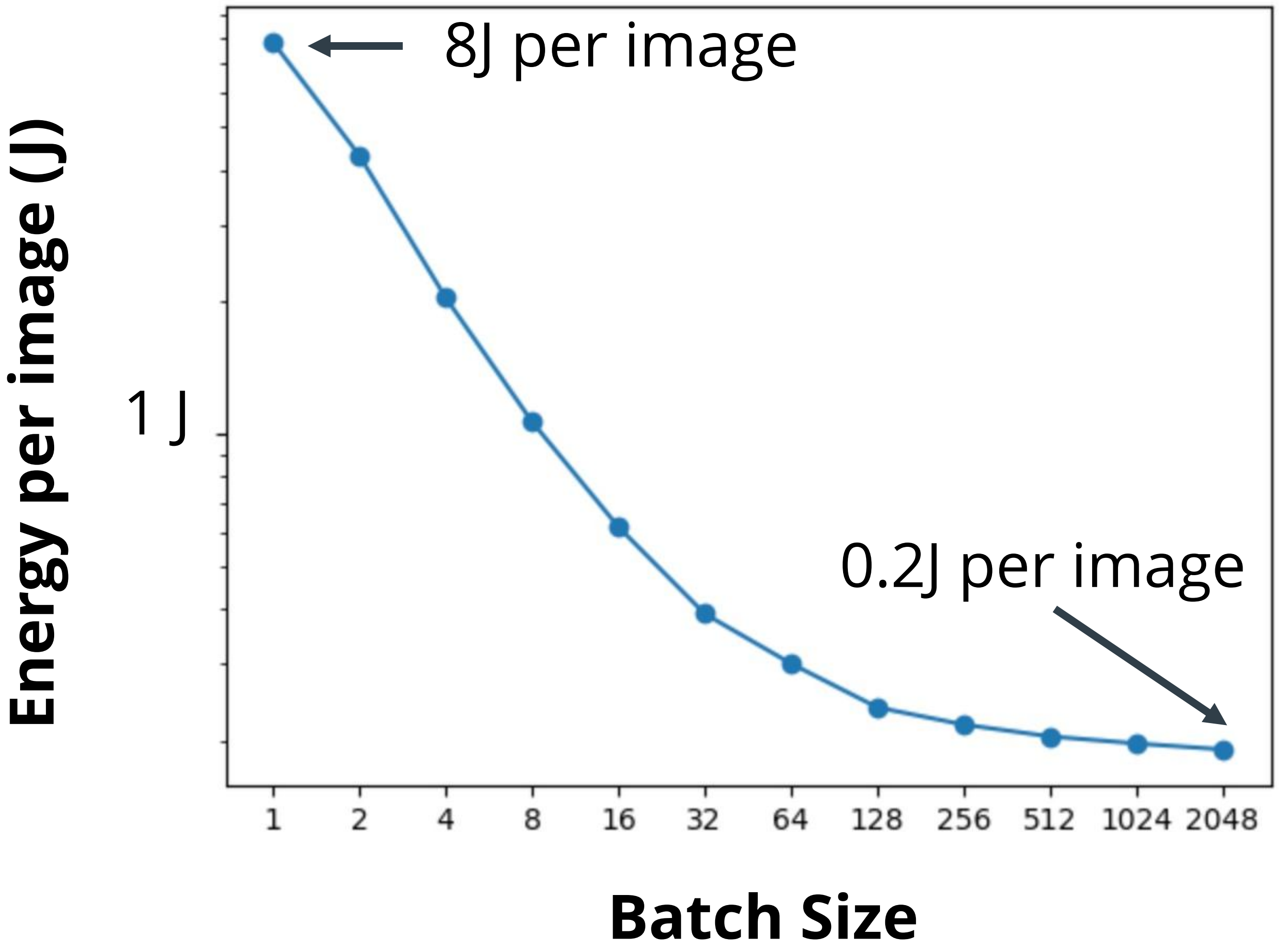
40 TOPS/W
(25 fJ/operation)

« These TOPS/W Are Not That High »

- Only prototypes

- **TOPS/W is a problem metric**

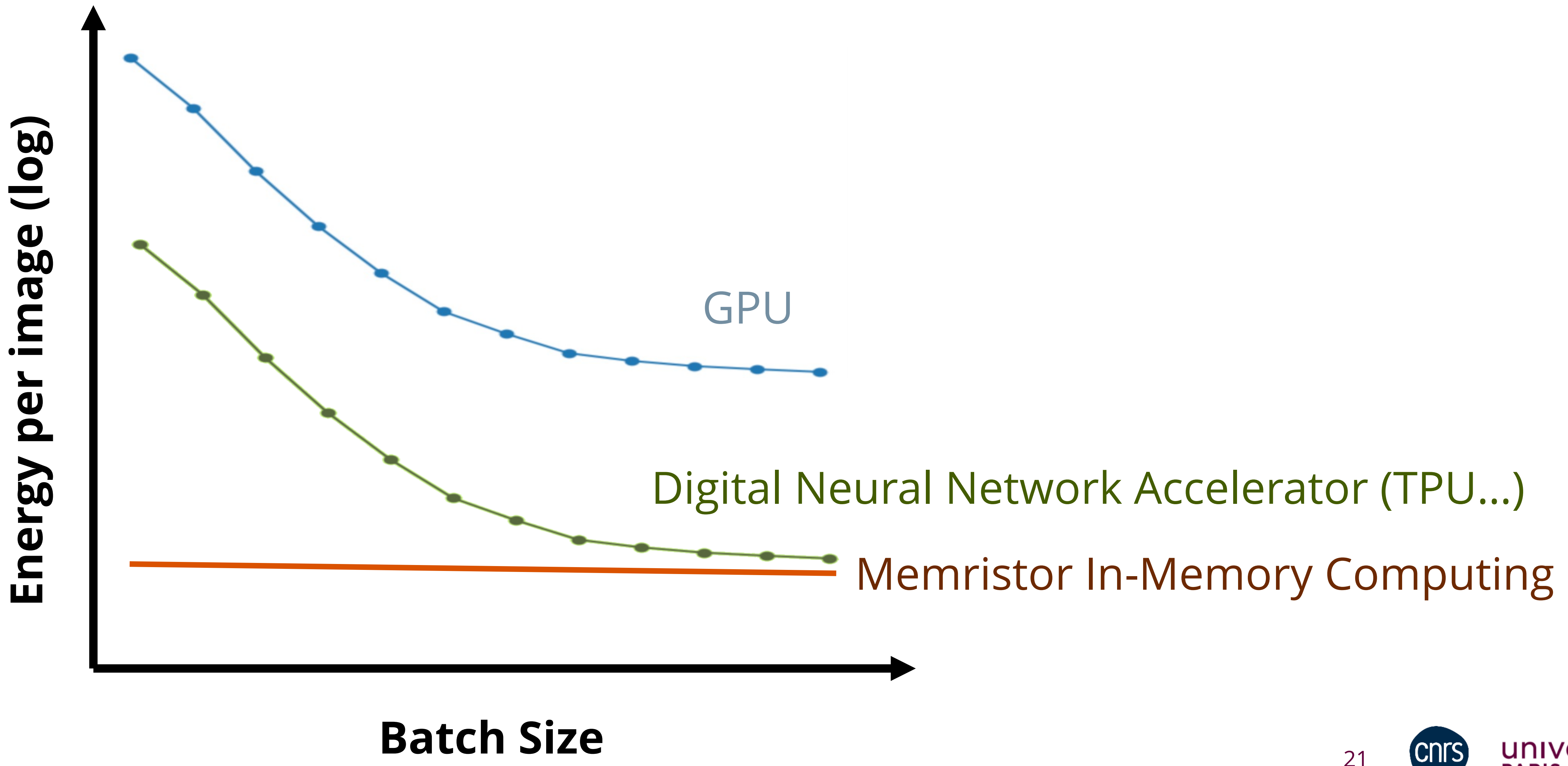
Real Energy Consumption Can Be Dramatically Higher than Peak



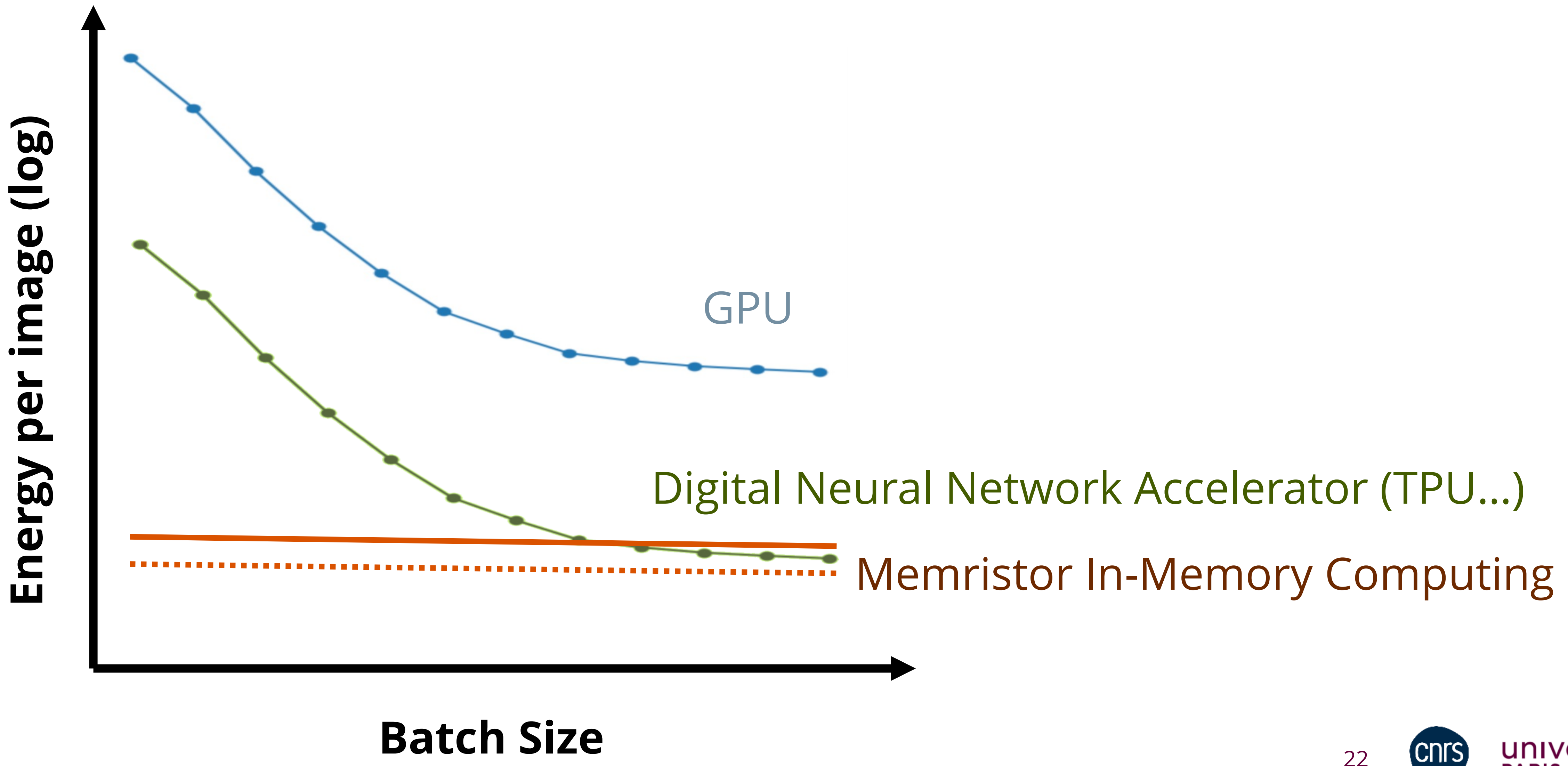
NVIDIA A100 GPU
ResNet50

Courtesy of David Novo (Univ Montpellier/CNRS)

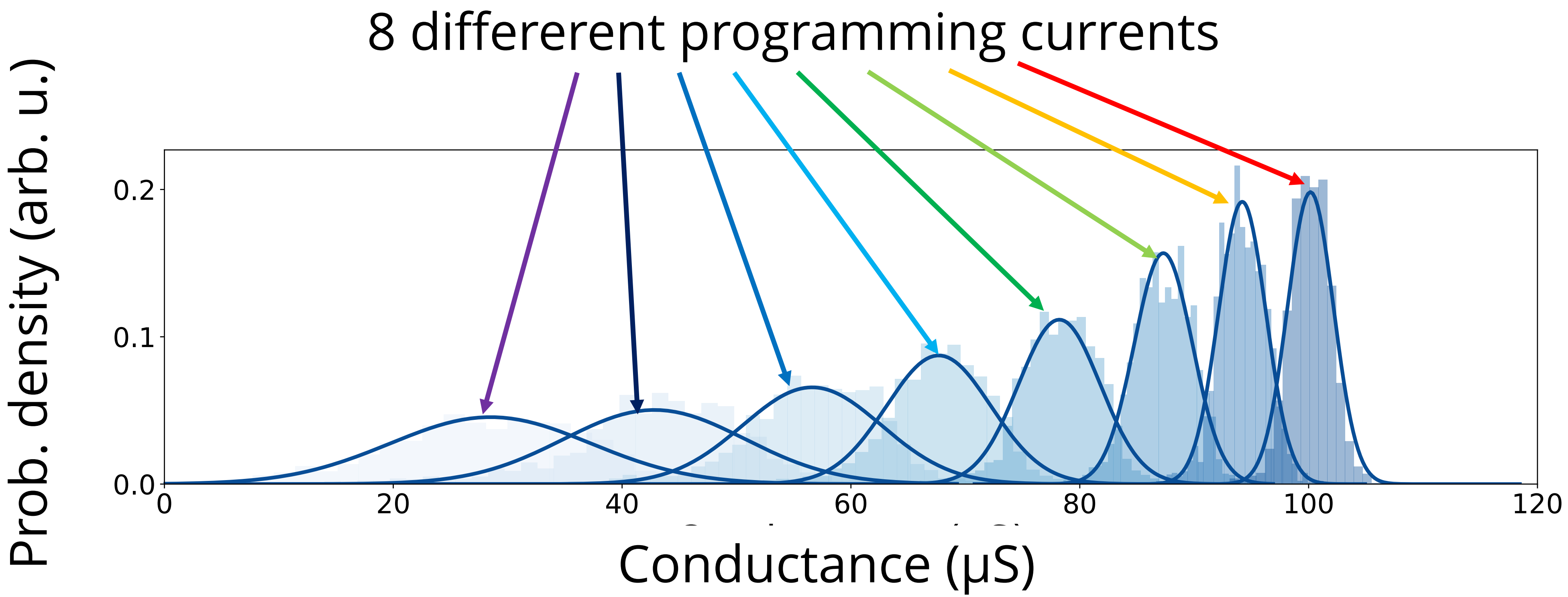
What Memristors Can Do



What Memristors Can Do

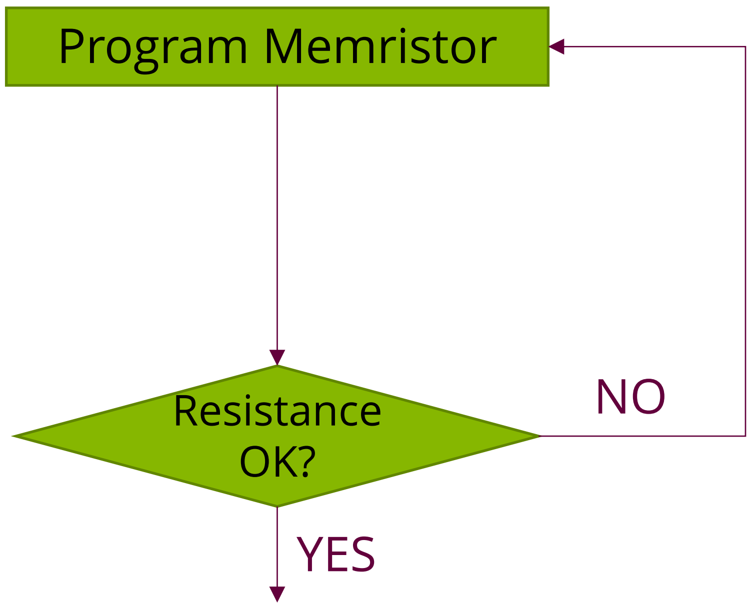


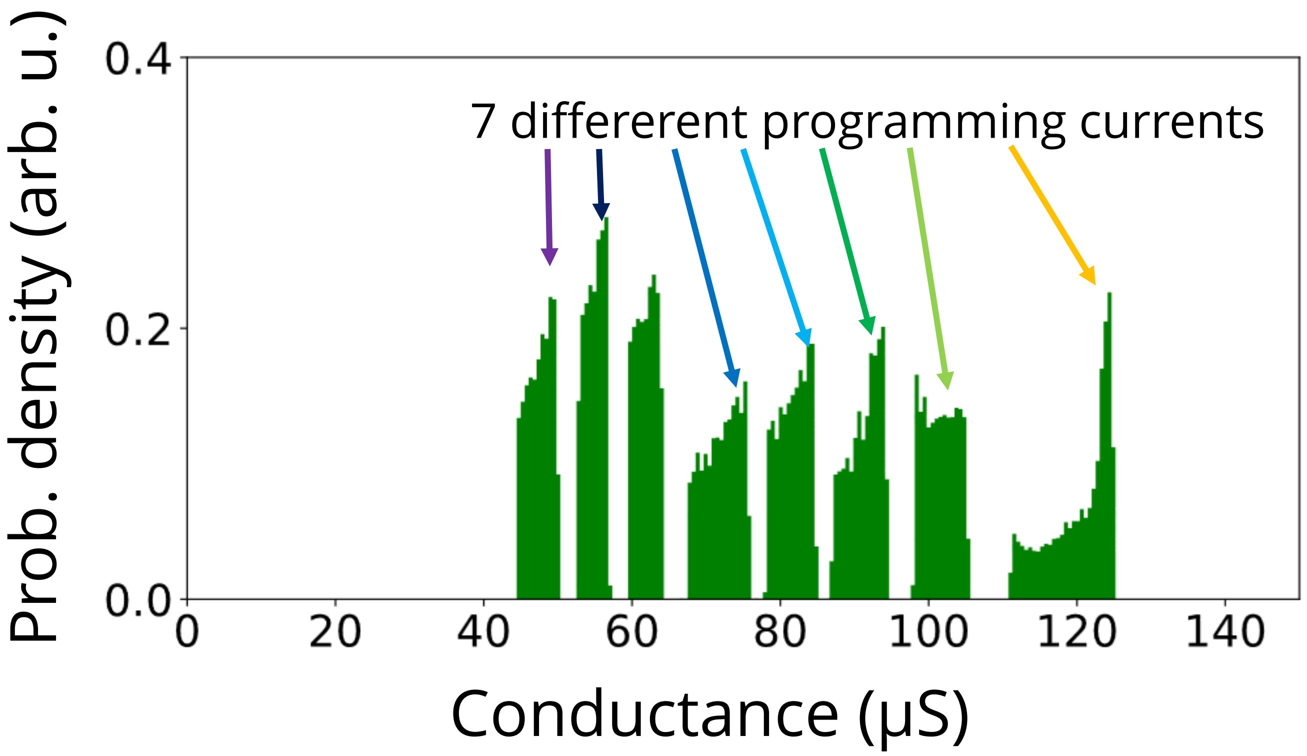
The Challenge of Analog In-Memory Computing: Memristors Act as Random Variables



Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

Fighting the Random Character of Memristors: *Program and Verify*



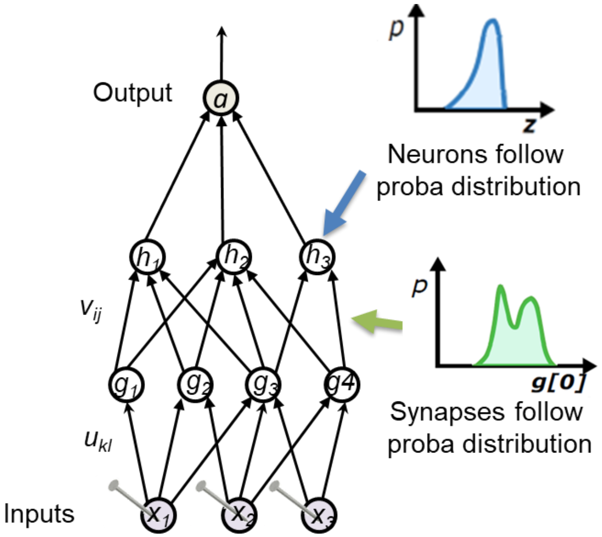


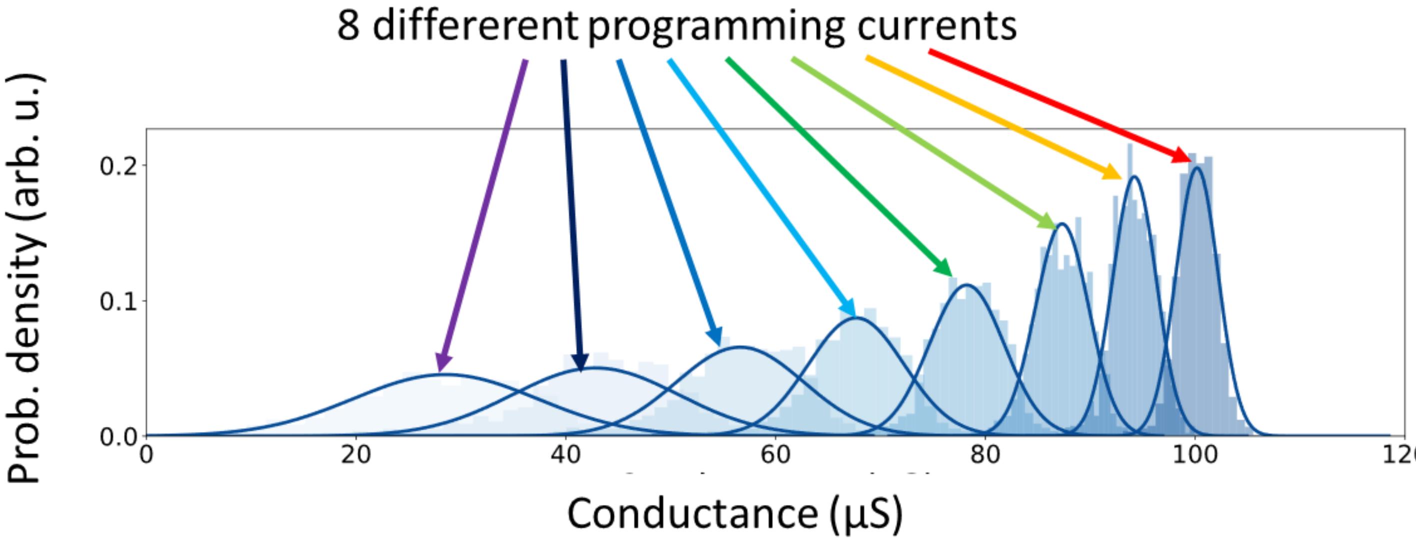
Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

Memristor Imperfections Can Naturally Produce a *Bayesian* Neural Network!

In Bayesian models, everything is a random variable that follows specific probability distributions

Memristors actually act as a random variable that follow specific probability distributions!





Our concept: Bayesian models can be a "better" way to exploit memristors

Bonnet et al, Nature Communications 14, 7530 (2023)

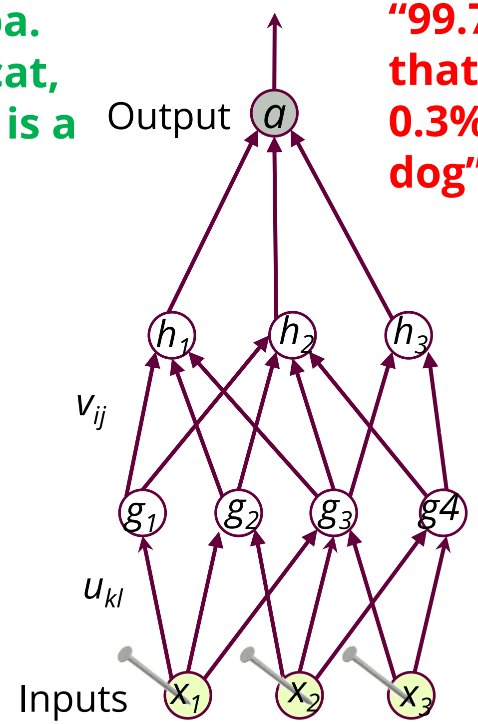
Collaboration

Bayesian Neural Networks

Assuming NN was trained to recognize « cats » and « dogs »

Conventional Neural Network

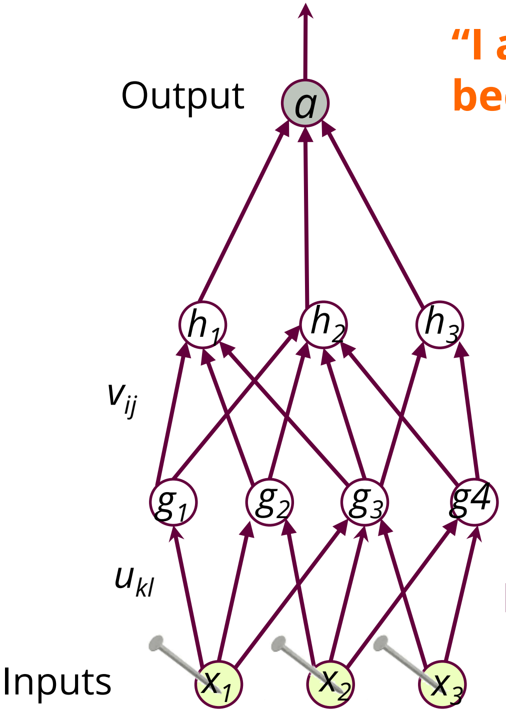
“99.9% proba. that this a cat, 0.1% that it is a dog”



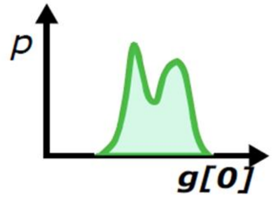
“99.7% proba. that this a cat, 0.3% that it is a dog”

0.7
Each synapse has one weight

Bayesian Neural Network



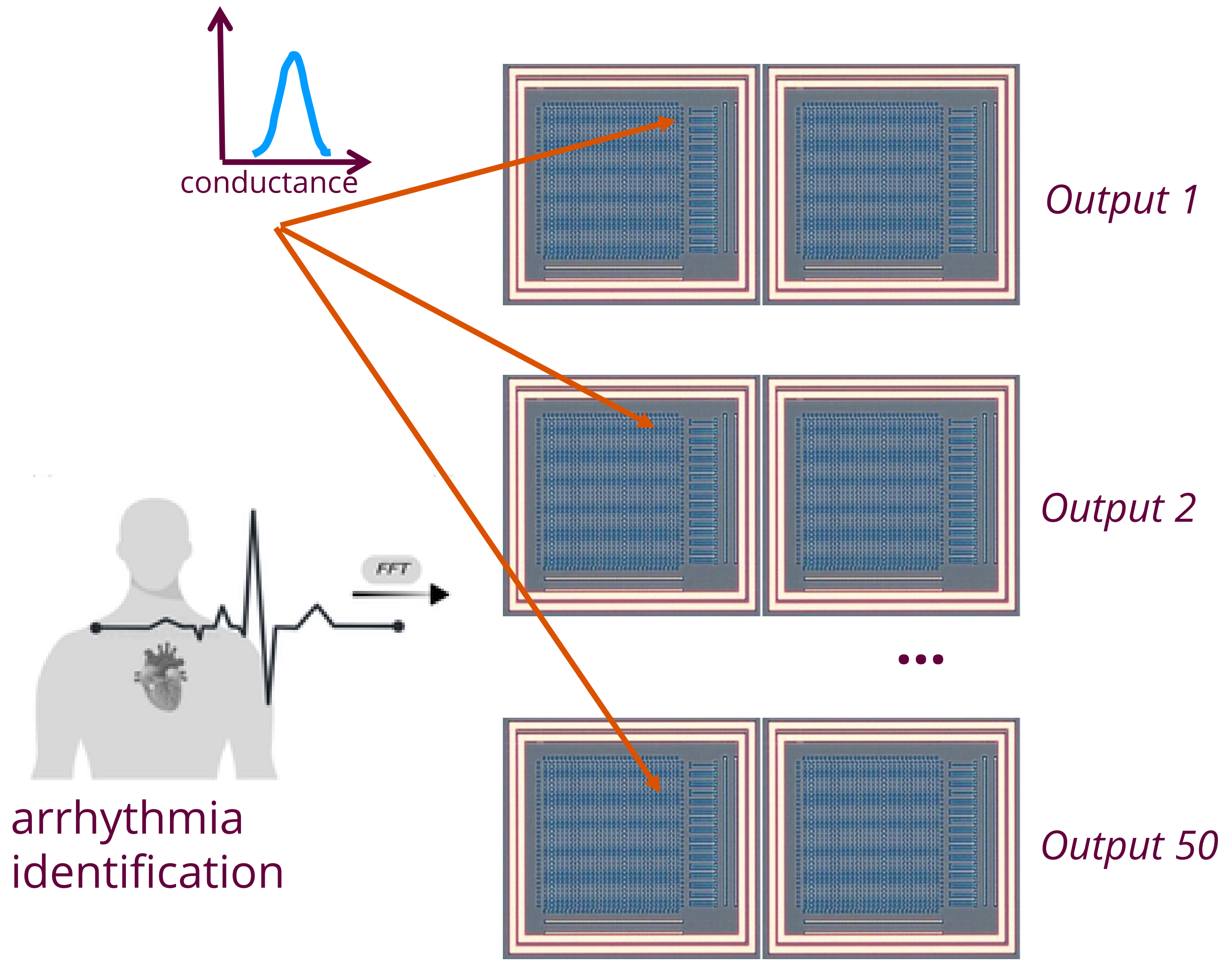
“I am not really sure what this is, because I have never seen it”



Synapses follow probability distribution

Memristor-Based Bayesian Neural Networks

We program 50 memristor-neural networks (each with two layers). We apply same input to them



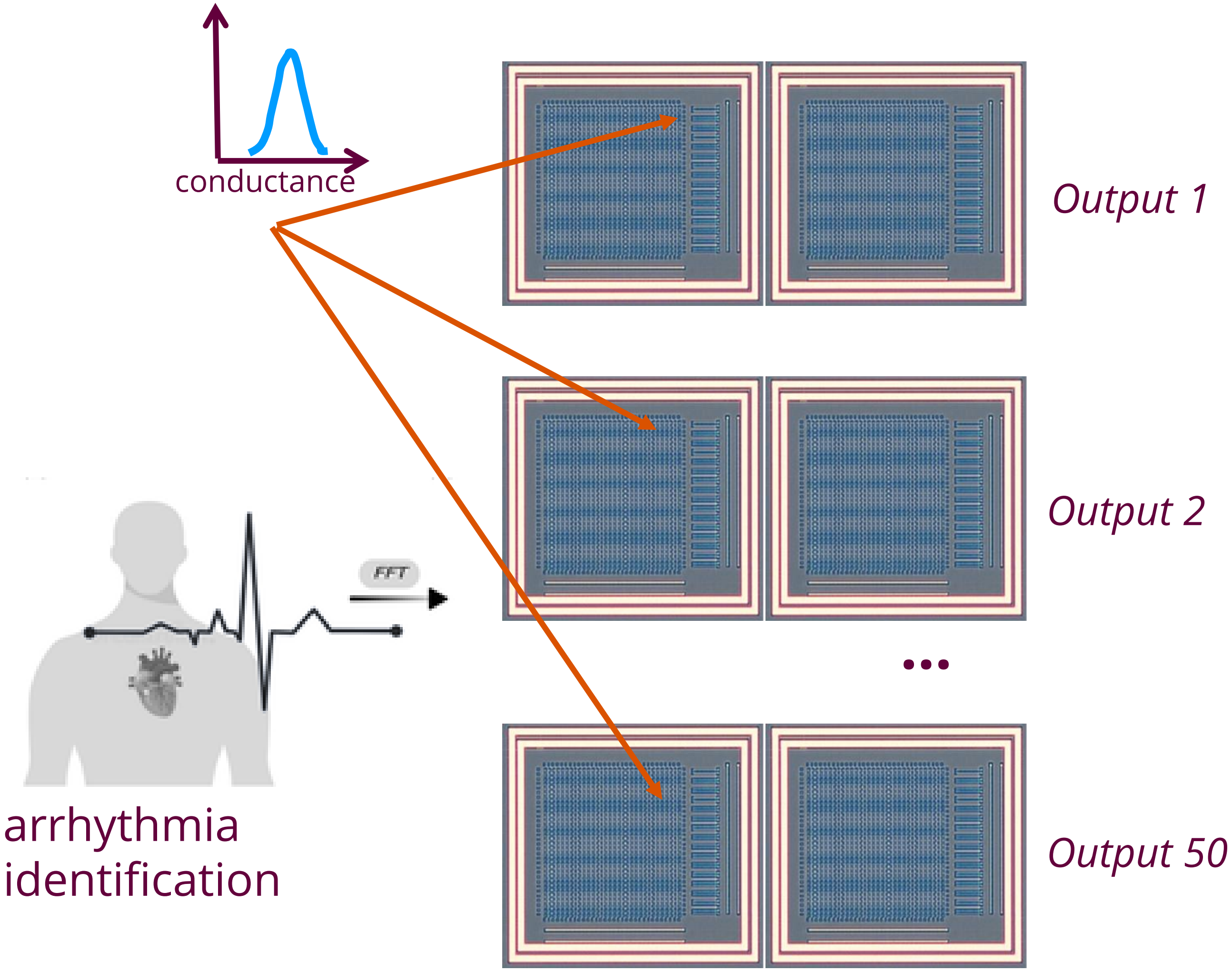
We get 50 outputs:
their dispersion tells about the *certainty* of the neural network

Neural network trained with a "technological loss"

Bonnet et al, Nature Communications 14, 7530 (2023)

Memristor-Based Bayesian Neural Networks

We program 50 memristor-neural networks (each with two layers). We apply same input to them



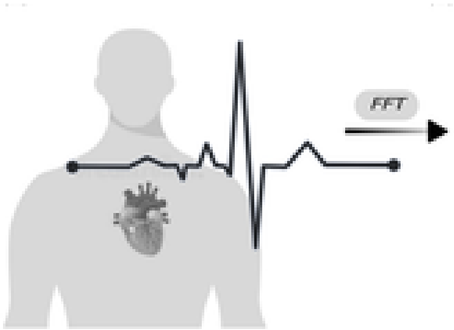
We get 50 outputs:
their dispersion tells about the *certainty* of the neural network

Neural network trained with Variational Inference incorporating a specific "technological loss"

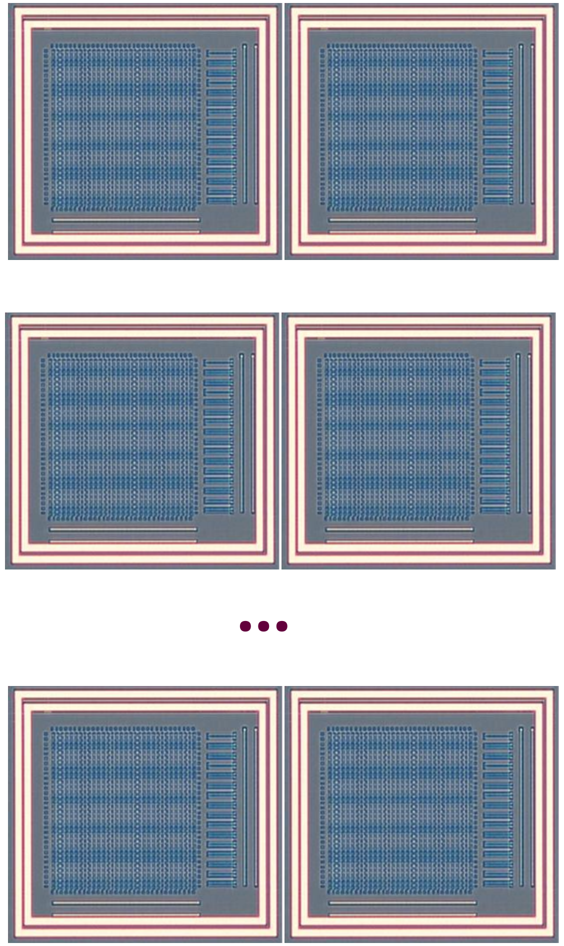
Bonnet et al, Nature Communications 14, 7530 (2023)

Memristor-Based Bayesian Neural Networks

We program 50 memristor-neural networks. We apply same input to them



arrhythmia
identification



Arrythmia type 1

Arrythmia type 1

...

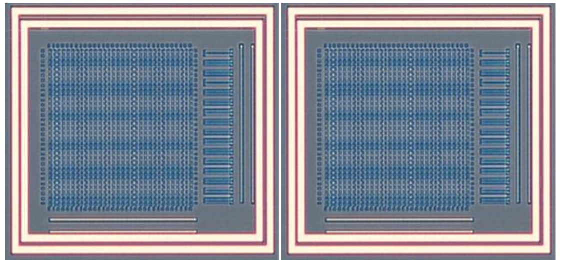
Arrythmia type 1

LOW UNCERTAINTY

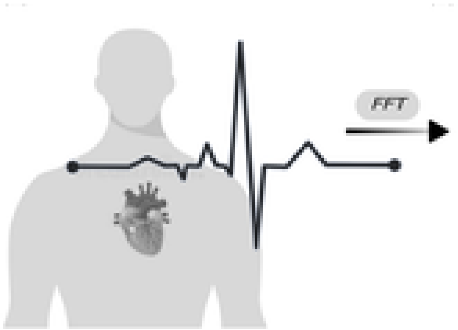
Bonnet et al, Nature Communications 14, 7530 (2023)

Memristor-Based Bayesian Neural Networks

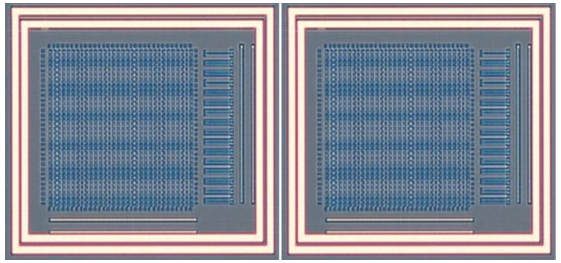
We program 50 memristor arrays, and we apply the same input to them



Arrythmia type 1

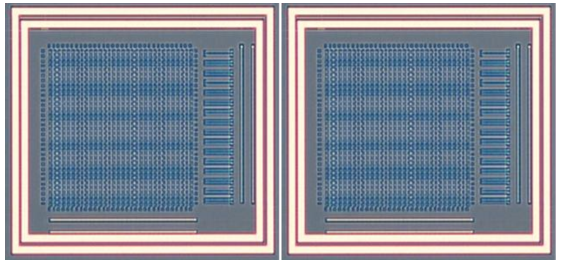


arrhythmia
identification



Hesitates
arrythmia type 1 or 2

...



Hesitates
arrythmia type 3 or 4

Unknown data:

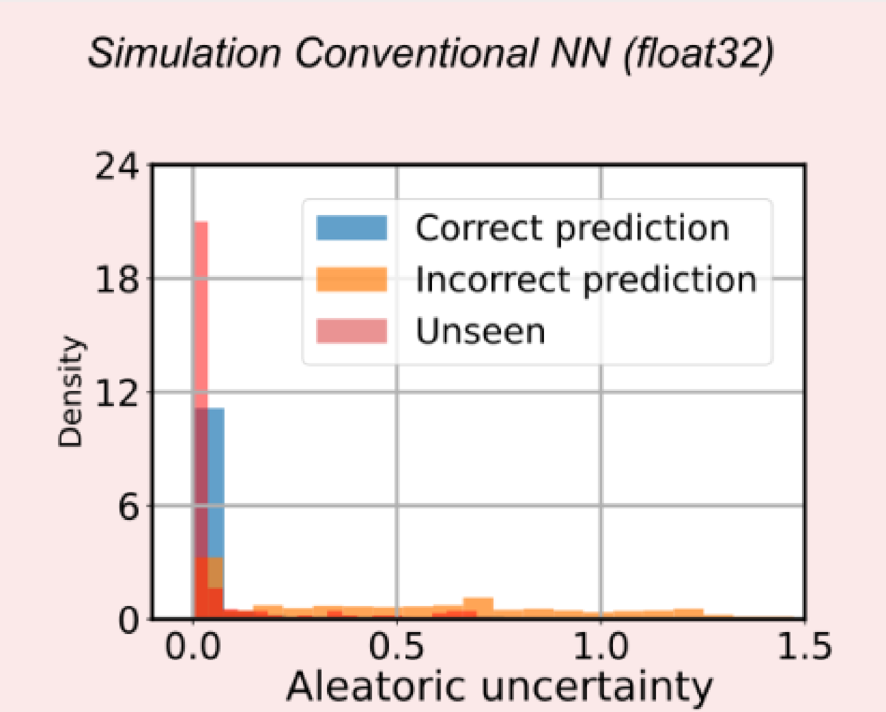
EPISTEMIC UNCERTAINTY

Bonnet et al, Nature Communications 14, 7530 (2023)

Fully Experimental Arrythmia Recogniton

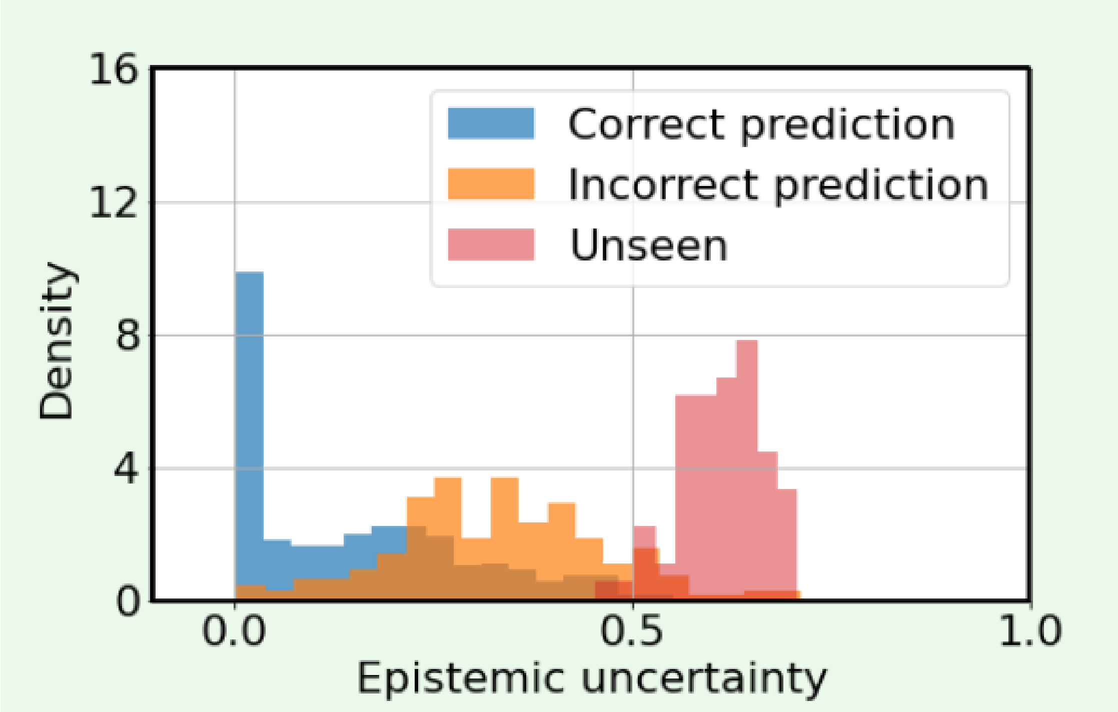
Traditional Neural Network

80% classification accuracy



Bayesian Neural Network

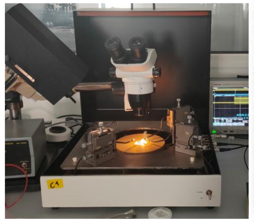
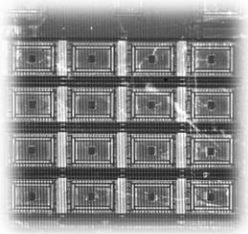
79% classification accuracy



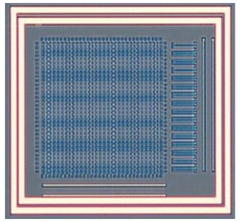
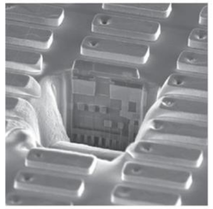
Easily recognizes
unknown types of arrythmia

Bonnet et al, Nature Communications 14, 7530 (2023)

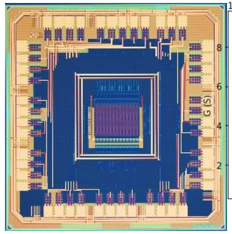
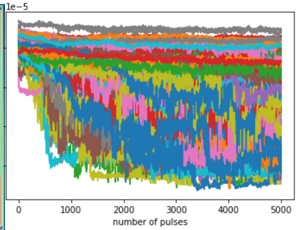
Toward Energy-Efficient and Trustworthy Embedded AI: The In-Memory Computing Approach

➢ Cointegrating Logic & Memory for Low-Energy AI

➢ The Analog & Stochastic Paths

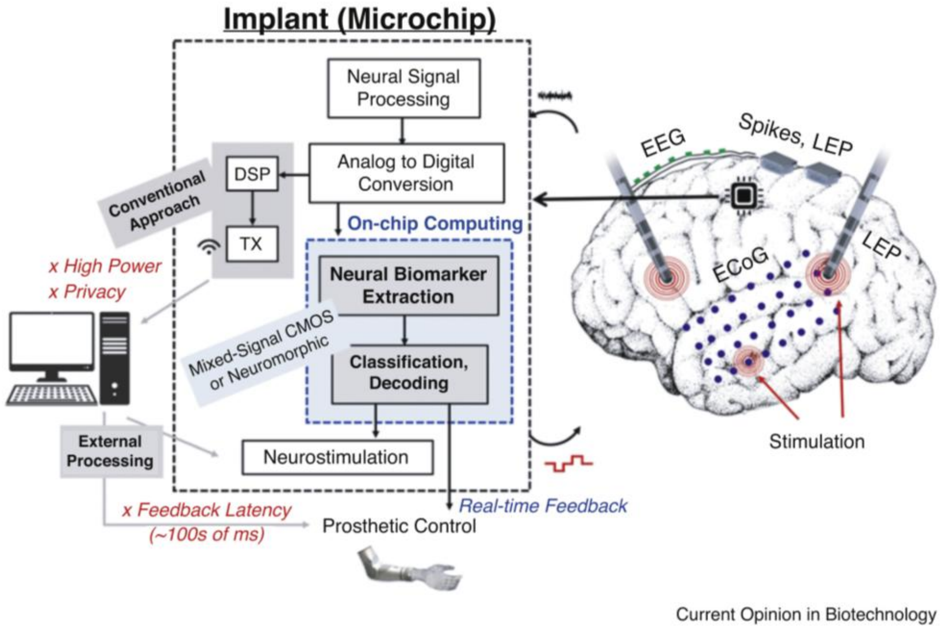



➢ **Learning with Physics**

The Grand Challenge of On-Chip Learning

- Current in-memory computing AI accelerators are focused on inference
- On-chip learning/adaptation would also have tremendous prospects

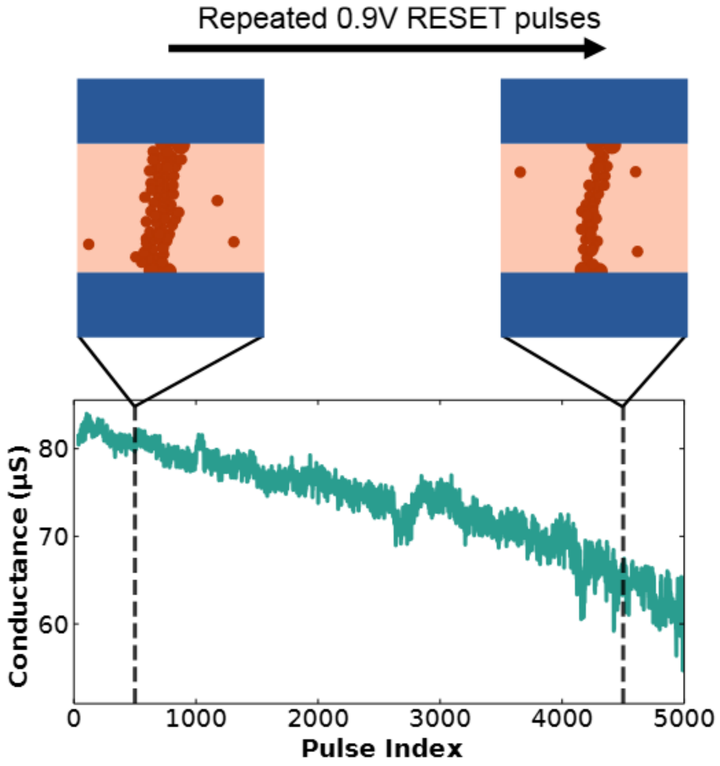
Medical applications

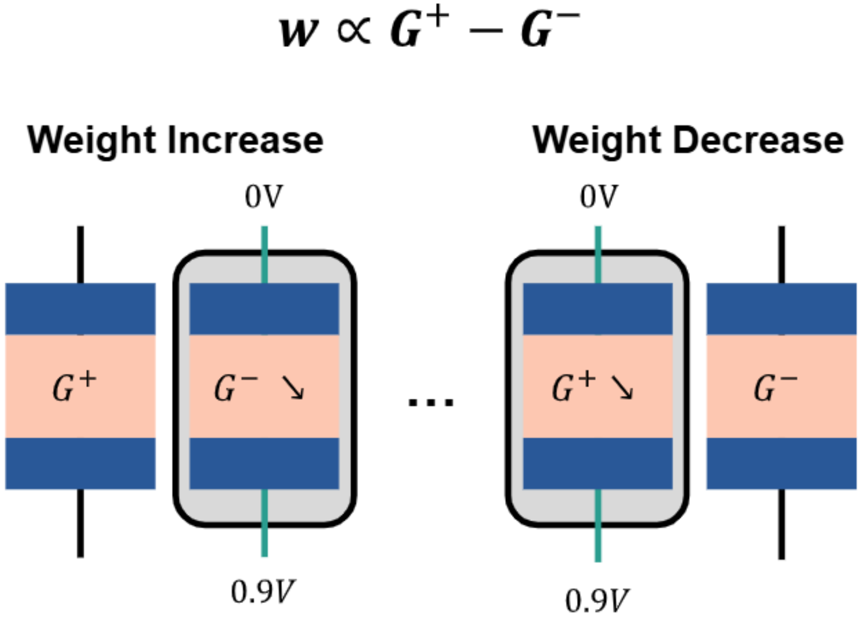


Yoo and Shoaran,
Current Opinion Biotechnol., 2021

But on-Chip learning requires reprogramming the weights repeatedly

Learning With Memristor Low-Voltage Reset



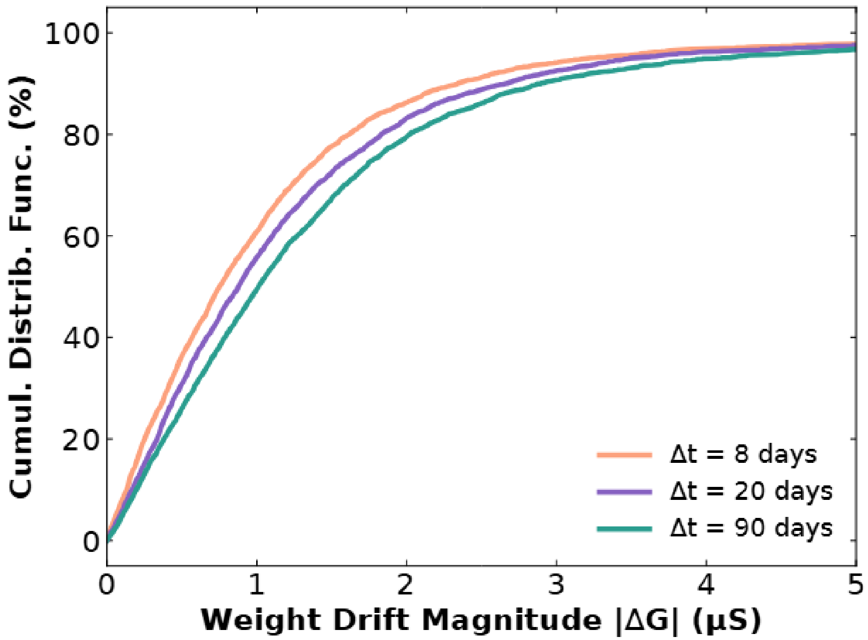


Imbert, Renaudineau, Diallo et al, IEEE IEDM (2024)

Learning With Memristor Low-Voltage Reset

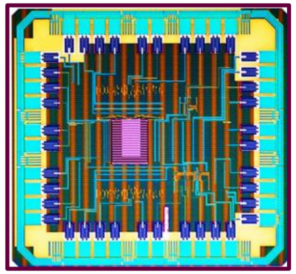
Low voltage Reset is noisy, but:

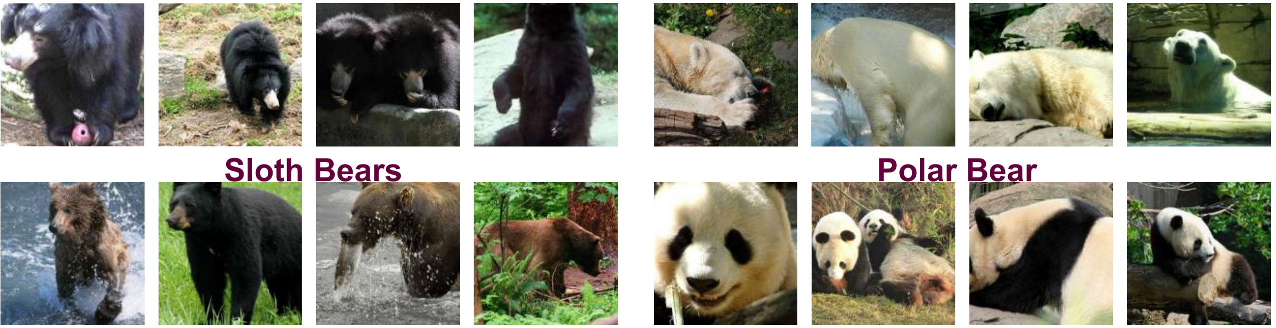
- Energy efficient (<1V, <100ns)
- Outstanding endurance
- Month-long retention



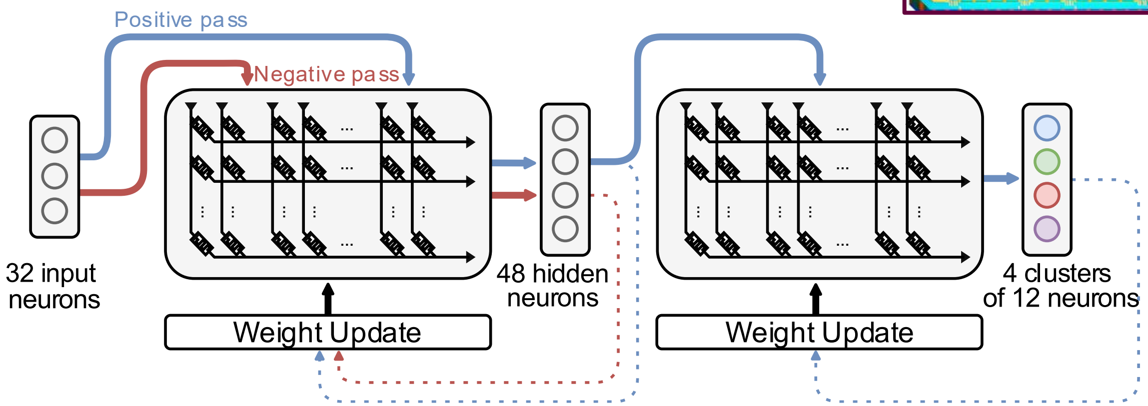
Imbert, Renaudineau, Diallo et al, IEEE IEDM (2024)

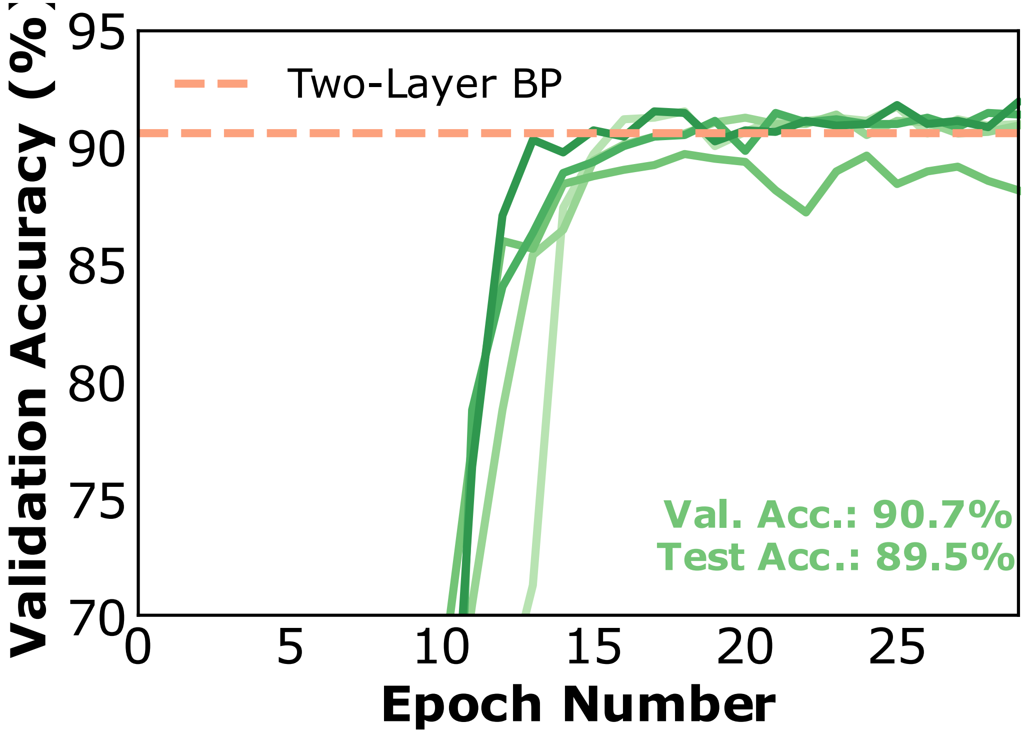
Learning to Classify Bears with 8,064 Memristors





Supervised Forward-Forward (SFF)



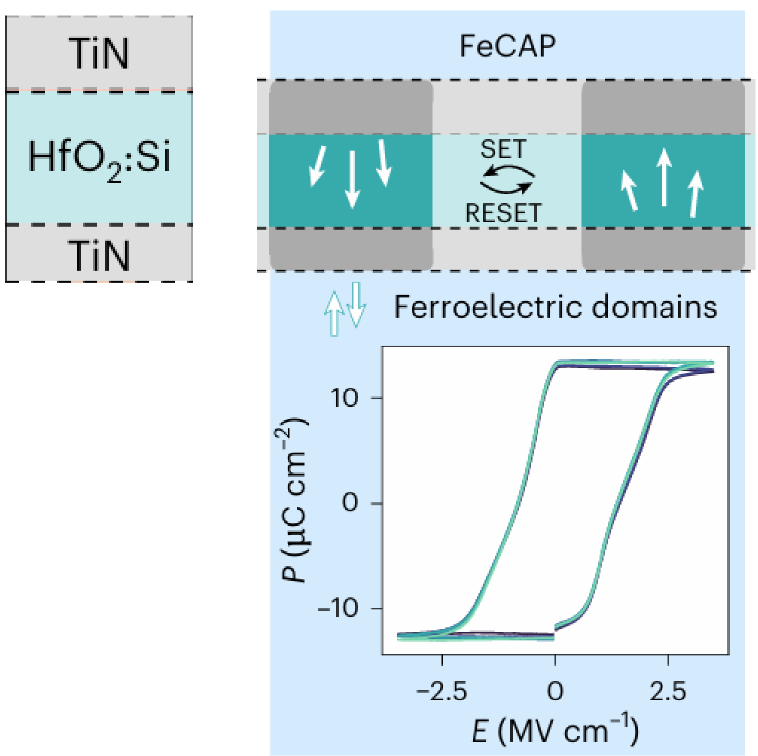


Forward-only learning: learning using the same memristor-level operations than inference

Renaudineau et al, submitted (2025)

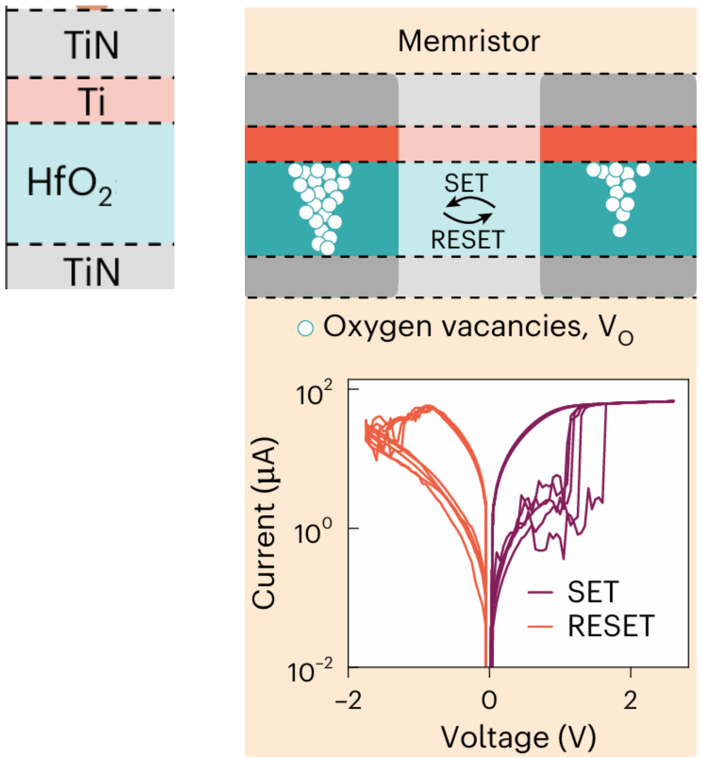
Hafnium Oxide Can Also Provide a Short Memory

Ferroelectric HfO_2



Ultra-low energy writes, outstanding endurance
Destructive read

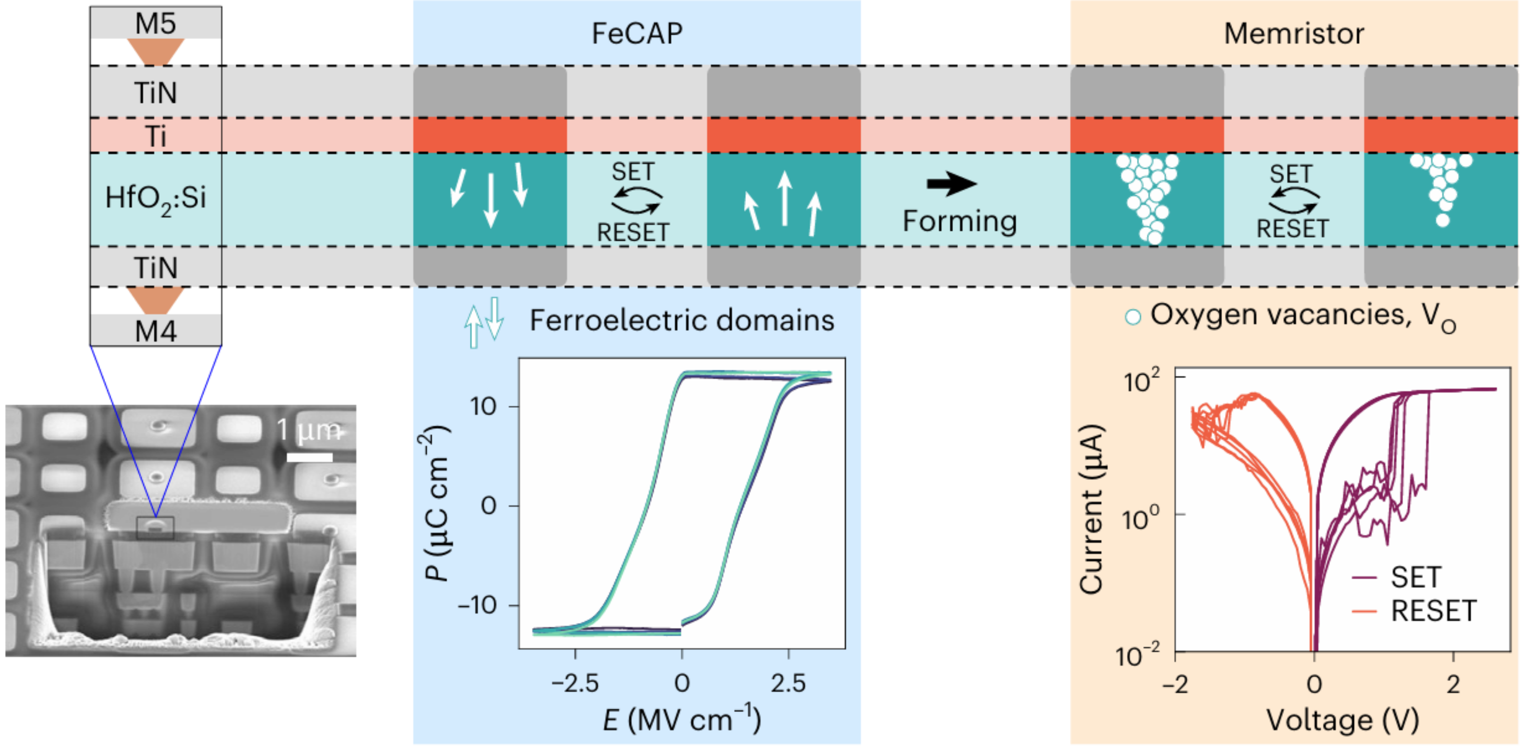
Memristor HfO_2

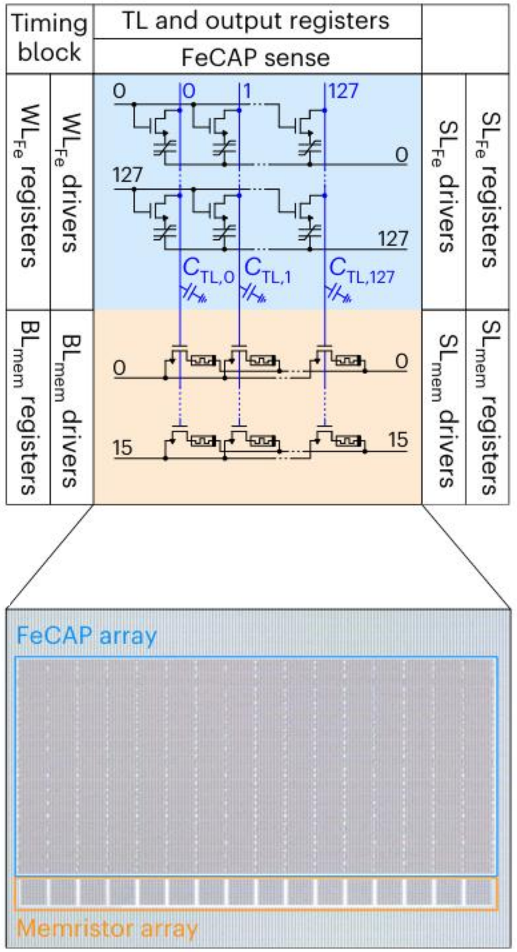


Higher energy writes, lower endurance
Destructive read

Martemucci et al, Nature Electronics 8, 921 (2025)

The Ferroelectric-Memristor Memory





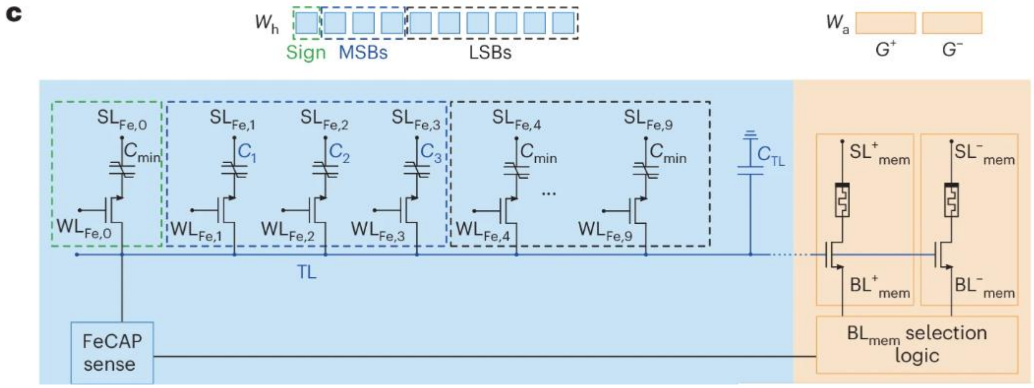
Collaboration

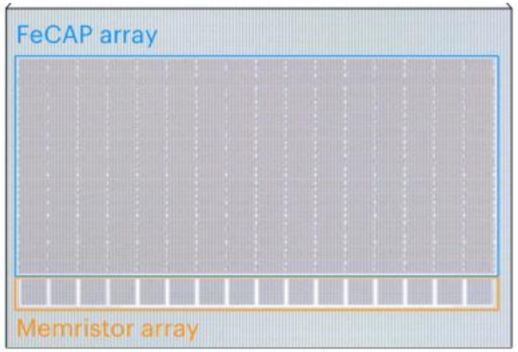
cea leti · cea list · ims Bordeaux · C2N

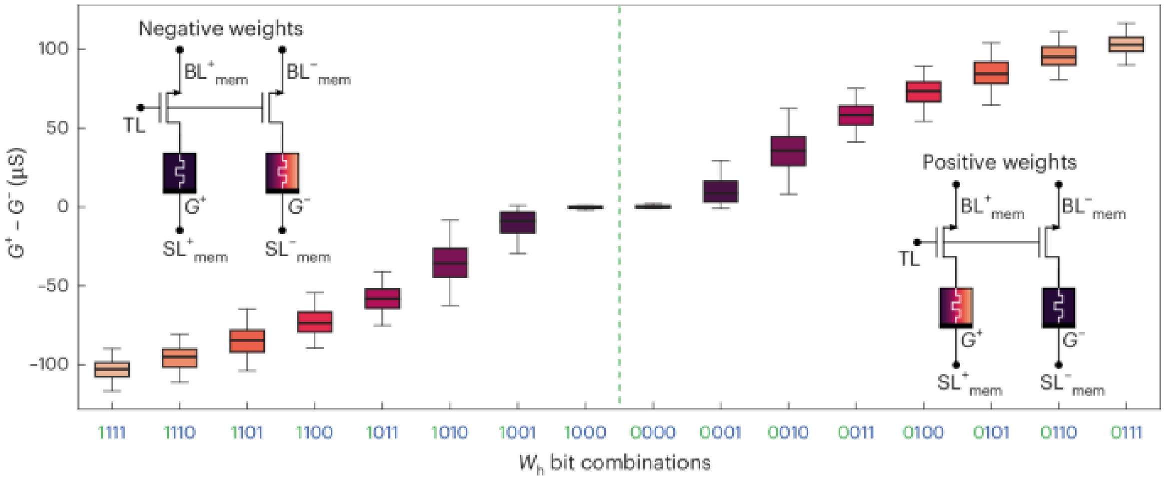
Martemucci et al, Nature Electronics 8, 921 (2025)

cnrs université PARIS-SACLAY

Information Can Be Transferred From Ferroelectric Short-Term to Memristor Long-Term Memory







Martemucci et al, Nature Electronics 8, 921 (2025)

Conclusion

- Computing with nanoelectronics devices: huge potential to reinvent electronics for low-power & trustworthy AI

- Interdisciplinary approach to electronics: codevelopment/codesign of technology, circuits, computer architecture and algorithms

- Current challenges: going 3D and moving toward genAI

Acknowledgments

- Clément Turck
- Atreya Majumdar
- Marie Drouhin
- Jacques-Olivier Klein
- Maxence Ernoult
- Axel Laborieux
- Adrien Renaudineau
- Théo Ballet
- Théo Dupuis
- Mohammed Akib Iftakher
- Kellian Cottart
- Filip Osana
- Benjamin Walker

- **Elisa Vianello**
- Laurent Grenouillet
- Louis Hutin
- Tifenn Hirtzlin
- Eduardo Esmanhotto
- Djohan Bonnet
- Michele Martemucci
- Niccolo Castellani
- François Andrieu
- Kamel-Eddine Harabi

- Stéphane Collin
- Oleksandr Bilousov
- Amadéo Michaud

- Julie Grollier

- Jacques Droulez
- Pierre Bessière

- Adrien Vincent
- Sylvain Saïghi

- Thomas Dalgaty
- François Rummens
- Yannick Malot

- **Jean-Michel Portal**
- Marc Bocquet
- Fadi Jebali
- Mamadou Hawa Diallo

- Jean-Pierre Walder
- Eloi Muhr
- Mathieu-Coumba Faye

Raphaël Laurent

Thank you for your attention!

damien.querlioz@universite-paris-saclay.fr
https://sites.google.com/site/damienquerlioz/