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

Damien Querlioz
Université Paris-Saclay, CNRS, Palaiseau, France
Centre de Nanosciences et de Nanotechnologies
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

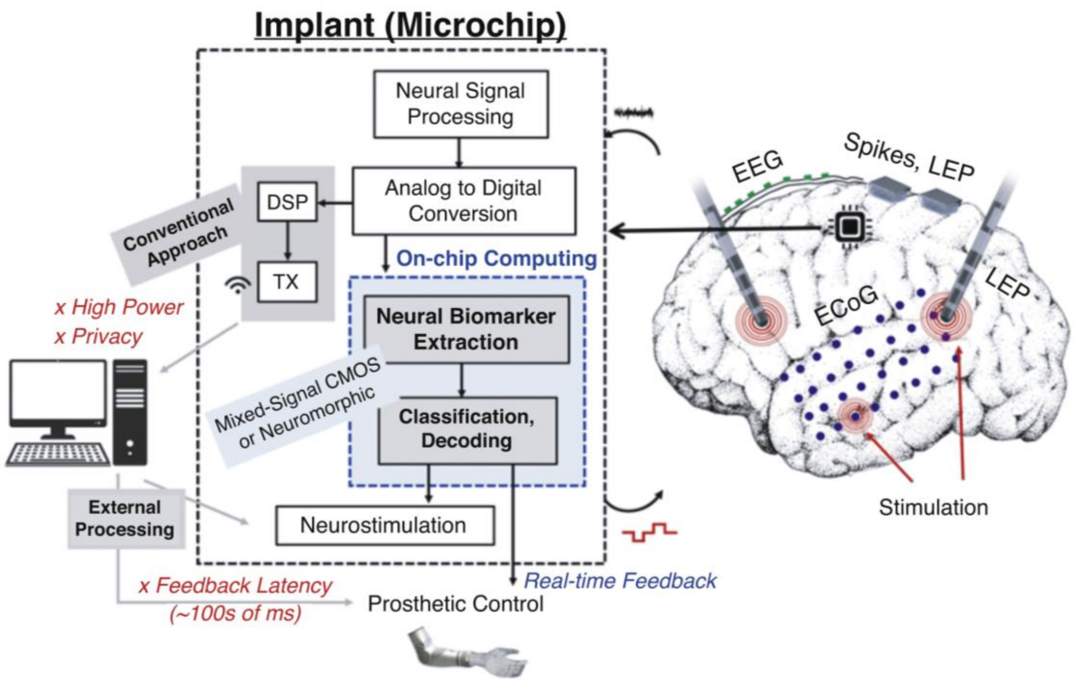
Sam Altman, CEO of OpenAI, attends the Asia-Pacific Economic Cooperation (APEC) CEO Summit in San Francisco, California, U.S. November 16, 2023. REUTERS/Carlos Barria/File Photo *Purchase Licensing Rights*

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?

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**

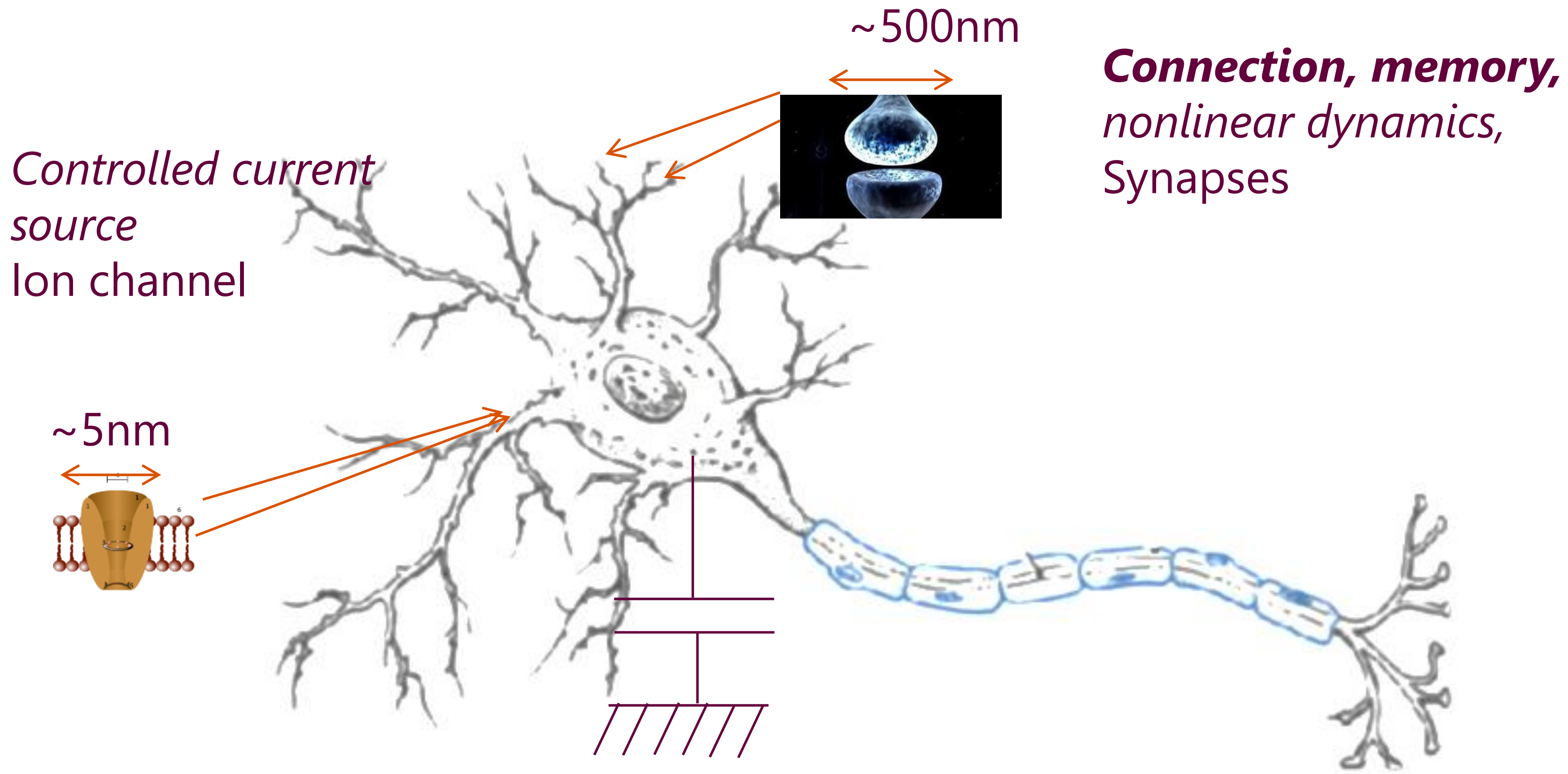
- **In computers, GPUs, and AI accelerators developed by industry (Google TPU, Apple NPU…), memory access is extremely costly**

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

Absent in the brain!

Pedram et al , IEEE D&T 2016

The Brain Achieves High Energy Efficiency by Computing « In Memory »

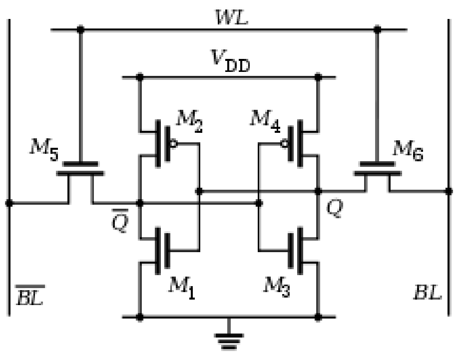


10,000 times more synapses than neurons!

Brain = a gigantic memory with some computation in the middle
No shared arithmetic unit

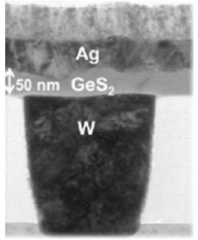
Integrating Logic & Memory Is a Considerable Challenge

- Only possible technology: Static RAM
  - *Even in "5-nanometer CMOS", SRAM bit is 150x150 nanometers*
  - *Leaky*
  - *Volatile*



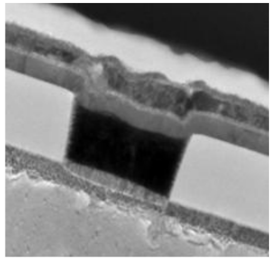
1 bit of SRAM

- **New Memories are coming to the rescue!**



Phase Change Memory

CEA-LETI



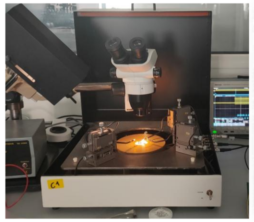
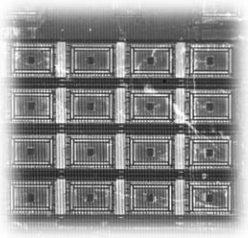
Memristors, Oxide Resistive Memory

CEA-LETI

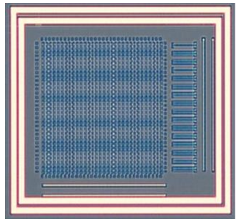
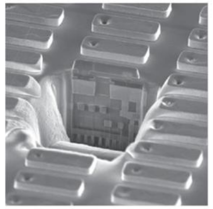
Spin Torque Memory

Samsung

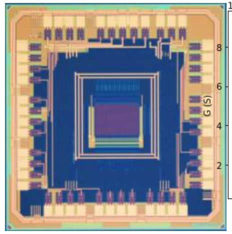
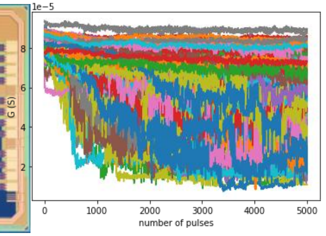
Toward Energy-Efficient and Trustworthy Embedded AI: The In-Memory Computing Approach

- Cointegrating Logic & Memory for Low-Energy AI

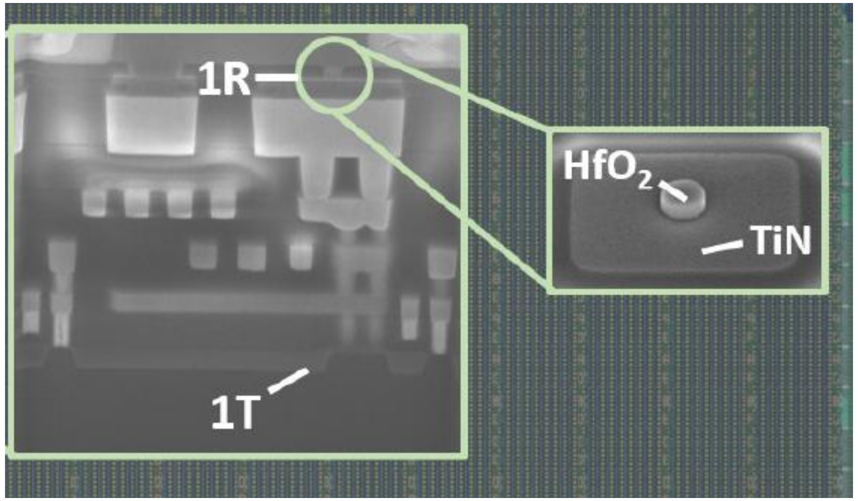



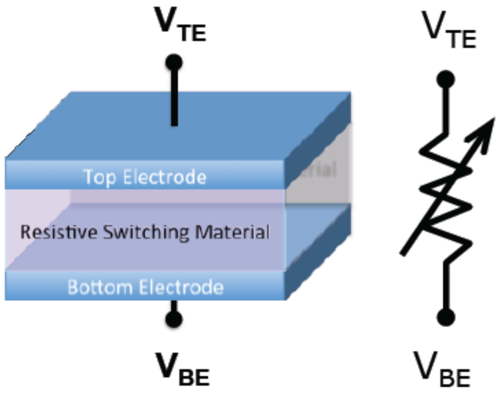
- The Analog & Stochastic Paths

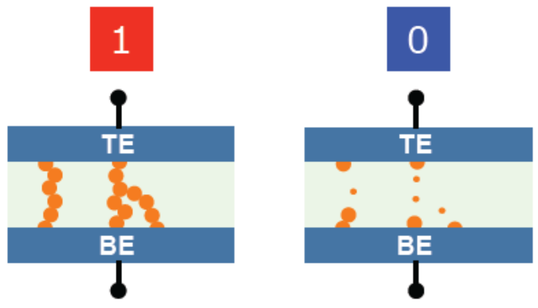



- Learning with Physics

Memristor/RRAM: an Artificial Synapse!







Low resistance

High resistance

- **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: Microcontroler Applications (NOR Flash replacement)

Infineon

All | Search

Newsletter Contact Wl

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

Claim your FREE
Electronics

NEWS BUSINESS MARKETS DESIGN PRODUCTS BLOGS R

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.

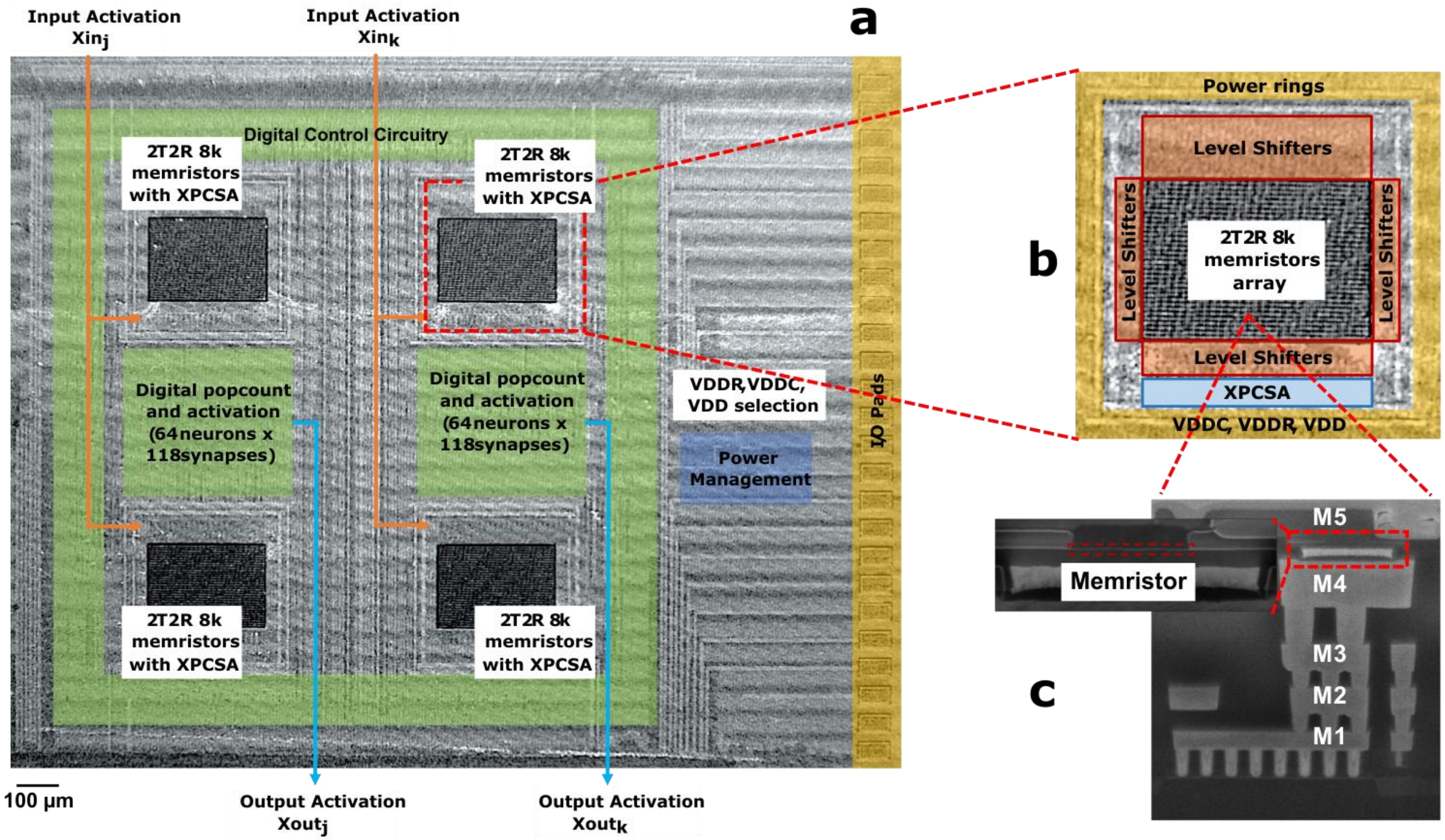
We repurpose them as *magic memory* for AI!

cnrs
université PARIS-SACLAY

Memristors Are Very Different From SRAM/DRAM

- Read is just as good as SRAM/DRAM
- Write is slower and write endurance is limited
  - ***You need to move atoms!***
- BER is typically ~ 10^{-6}

A Super-Robust Digital Binarized Neural Network



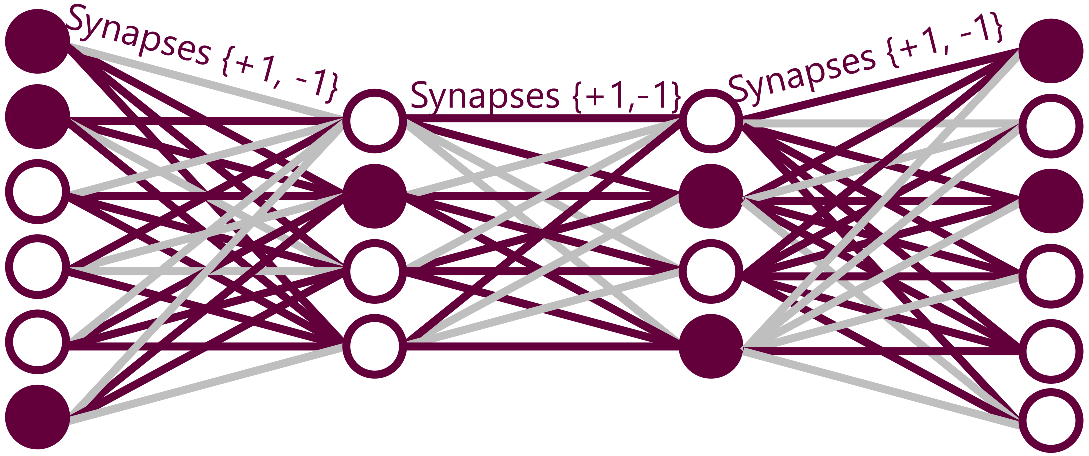
Collaboration

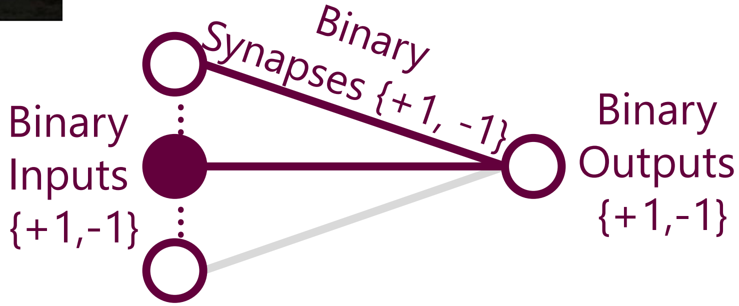
Jebali et al, Nature Communications 15, 741 (2024)

Binarized Neural network:
A Low Precison Neural Networks

Hubara, Courbariaux et al. NIPS 2016

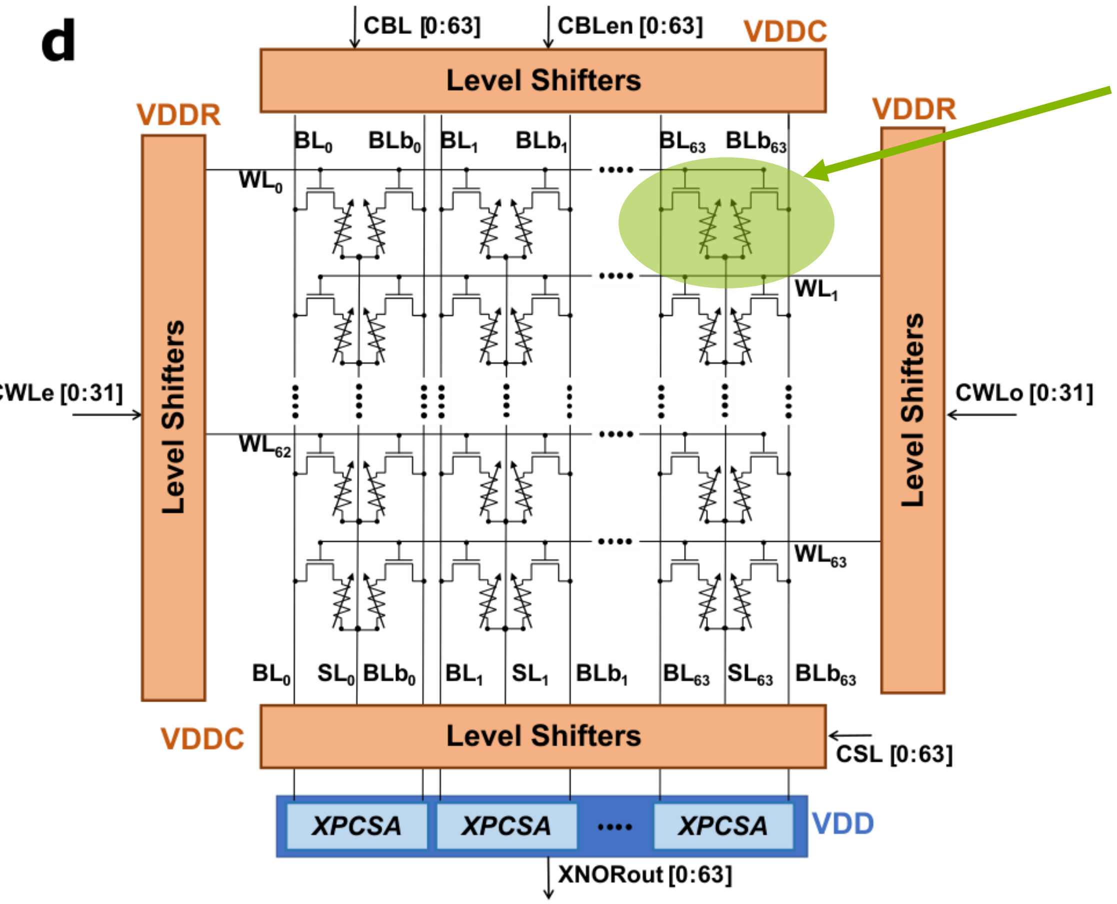
Yoshua Bengio's group



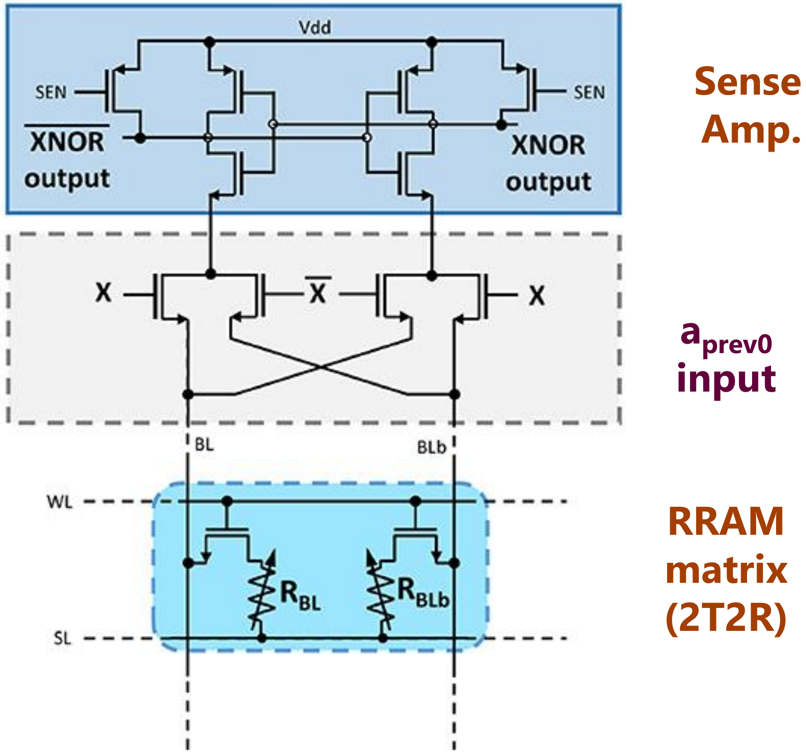


Our work :
Memristor as binary synapse

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



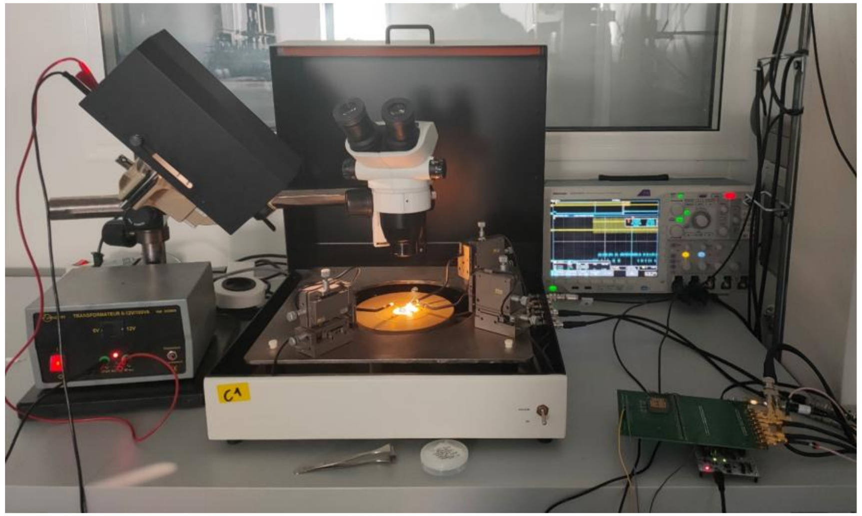
One weight: Programmed as 0/1 or 1/0



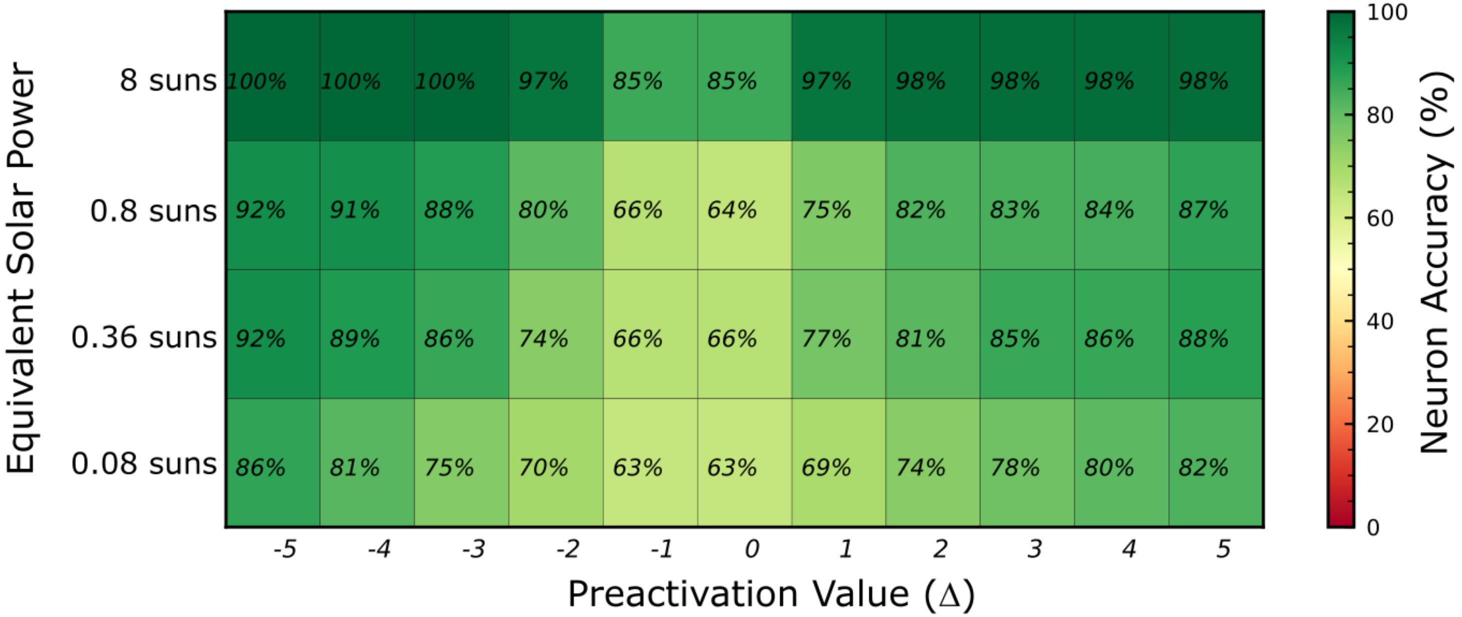
Weisheng Zhao, … Klein, Querlioz …., *IEEE TCAS I:* 2014.
IEEE Guillemin-Cauer Best Paper Award

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

Self-Powered Operation



Miniature wide-bandgap solar cell

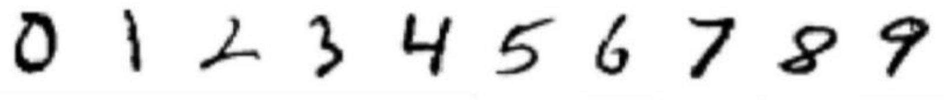


Neuron value $t_j = f\left(\sum w_{ij} s_i\right)$

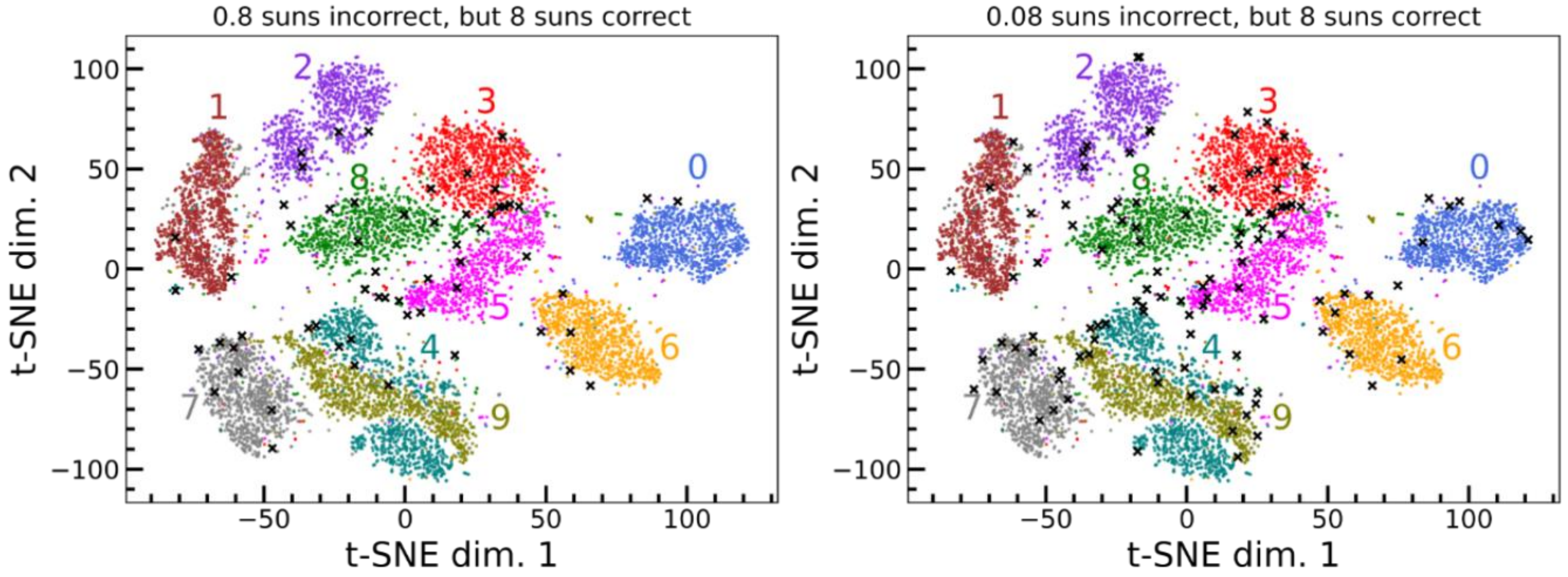
Under low illumination the circuit naturally switches to approximate computation

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

Self-Powered Operation



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



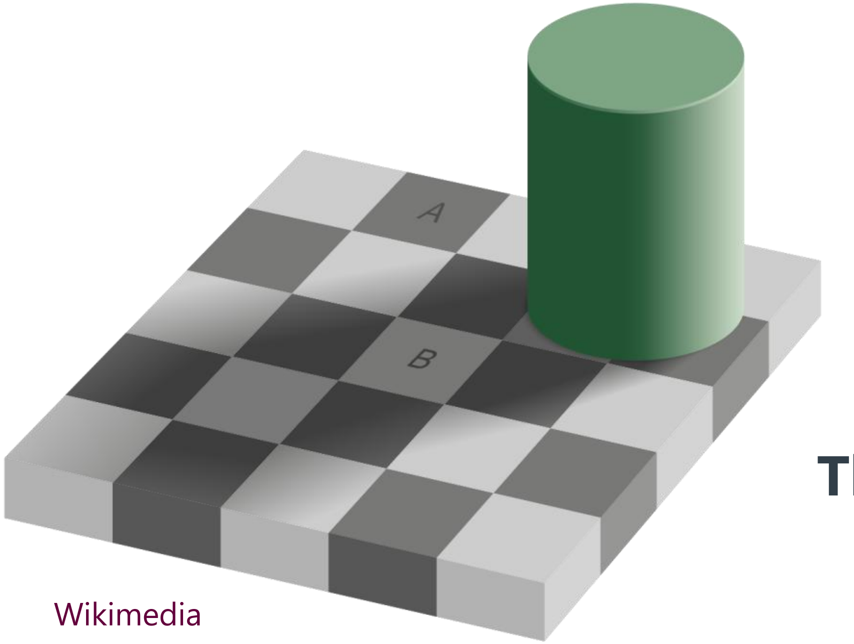
Jebali et al, Nature Communications 15, 741 (2024).

Computing In-Near Memory Is Applicable to Other AI Approaches

Thomas Bayes

$$p(Disease \mid Observations) \propto \underbrace{\overbrace{p(Observations \mid Disease)}^{Likelihoods} \times \overbrace{p(Disease)}^{Prior}}_{\text{Constructed with expert knowledge+Data}}$$

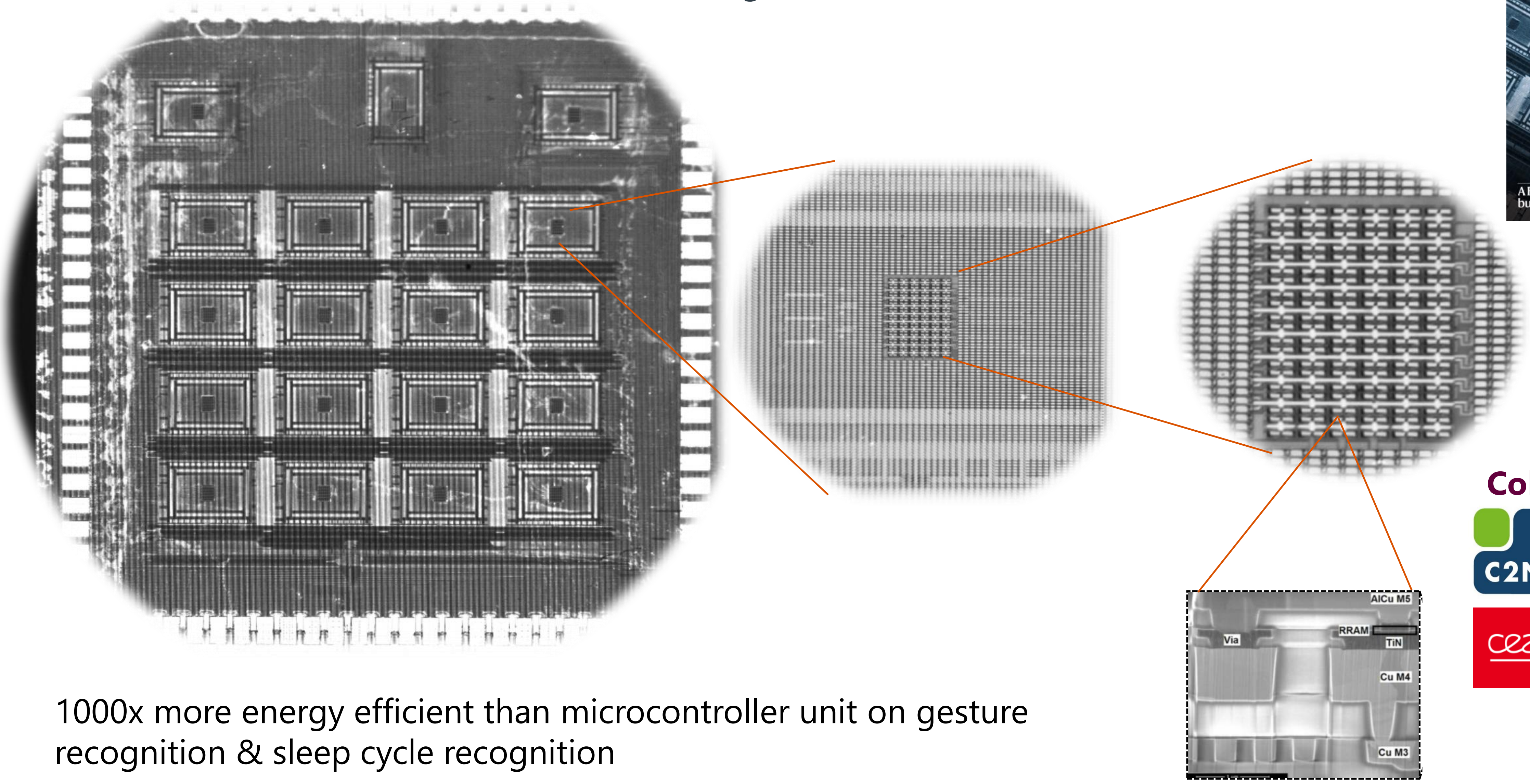
Hard to Compute



Wikimedia

The Brain Might Be Using Bayesian Reasoning

The Memristor-Based Bayesian Machine

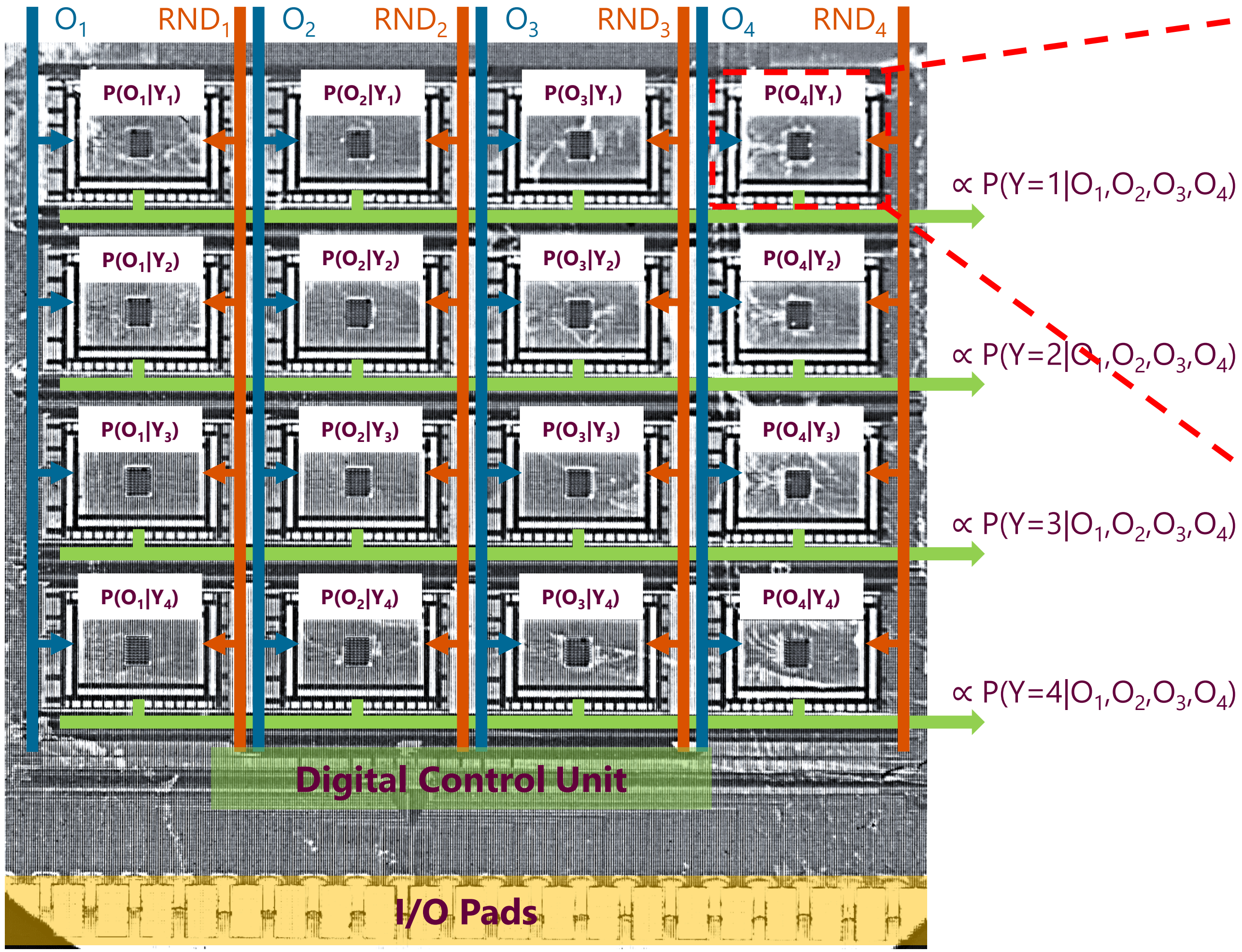


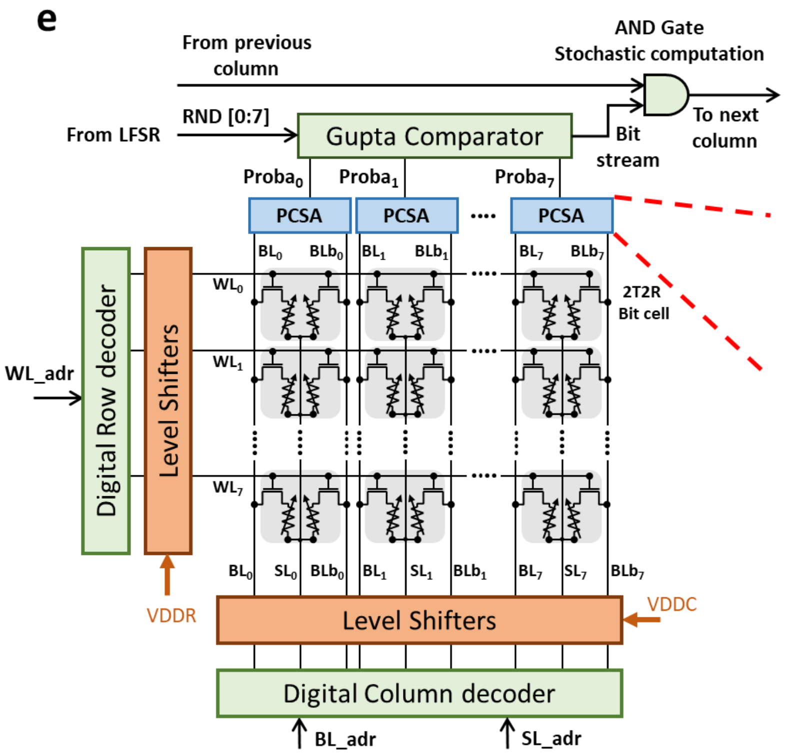
1000x more energy efficient than microcontroller unit on gesture recognition & sleep cycle recognition

Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Collaboration

The Memristor-Based Bayesian Machine

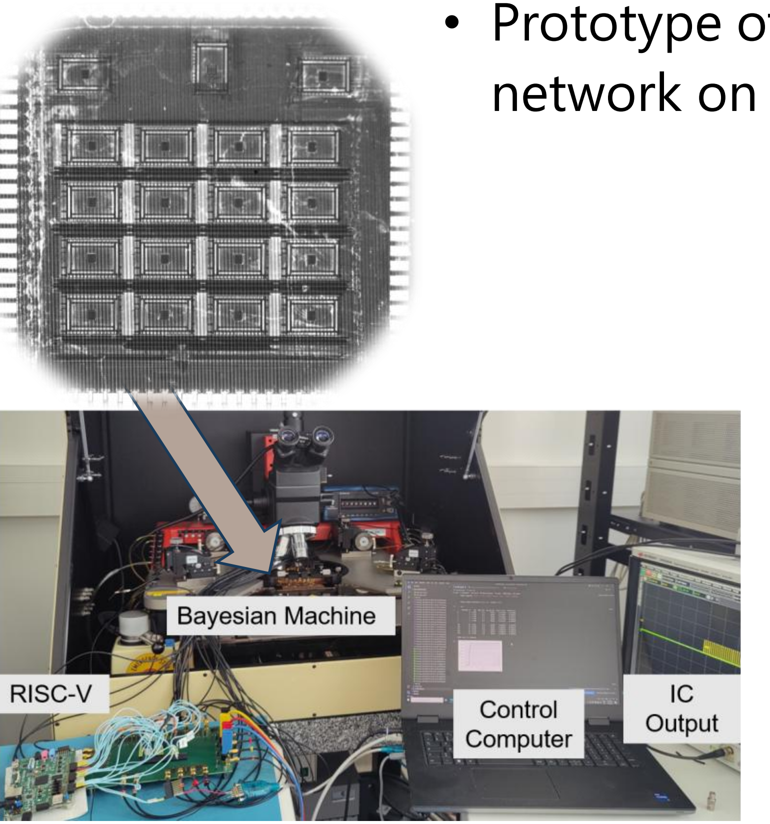




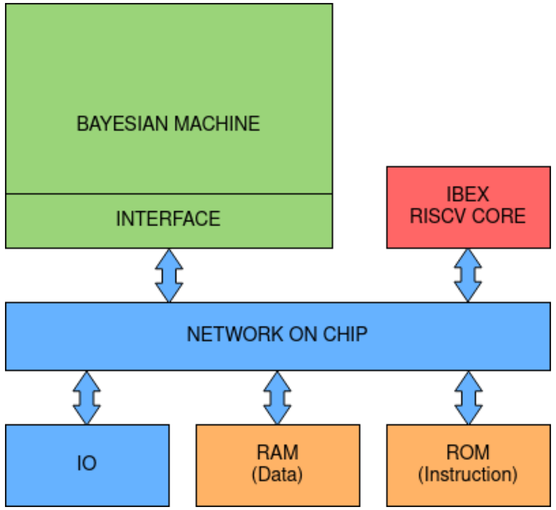
$$P(Y=4|O_1,O_2,O_3,O_4) \propto P(O_1|Y=4) \times \cdots \times P(O_4|Y=4)$$

Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

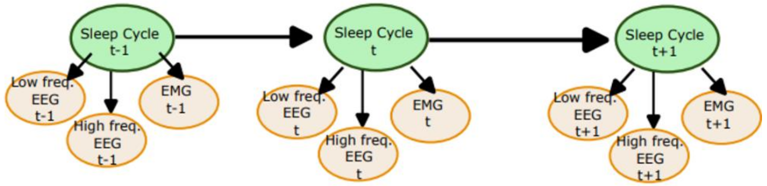
Integration of a Memristor Bayesian Machine with a RISC-V Processor

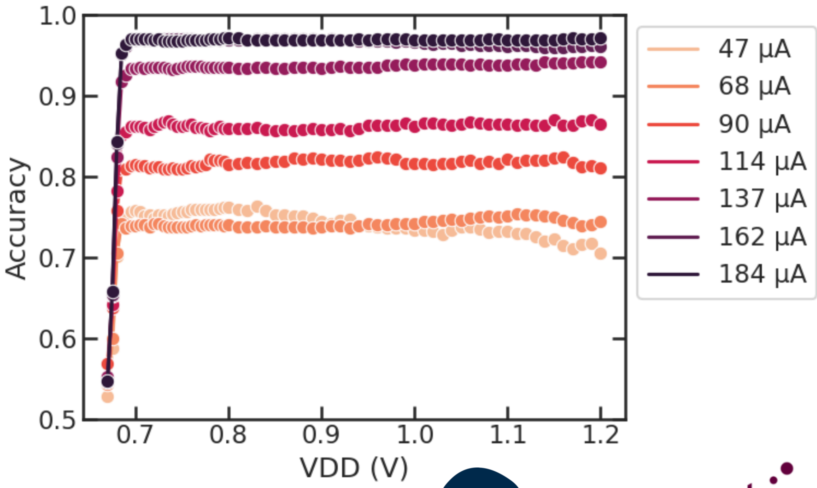


- Prototype of a SoC with Bayesian machine, network on chip and RISC-V core



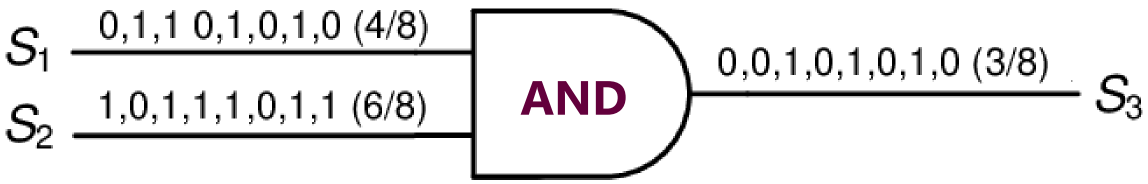
Allows testing real-life application (e.g., sleep cycle classification)





Turck et al, Communications Engineering (2025)

Stochastic Computing



$p(S_3)=p(S_1)\times p(S_2)$

A AND gate implements the multiplication of two probabilities!

Some resemblance with neurons
(one wire = one real number)

Logarithmic Computing

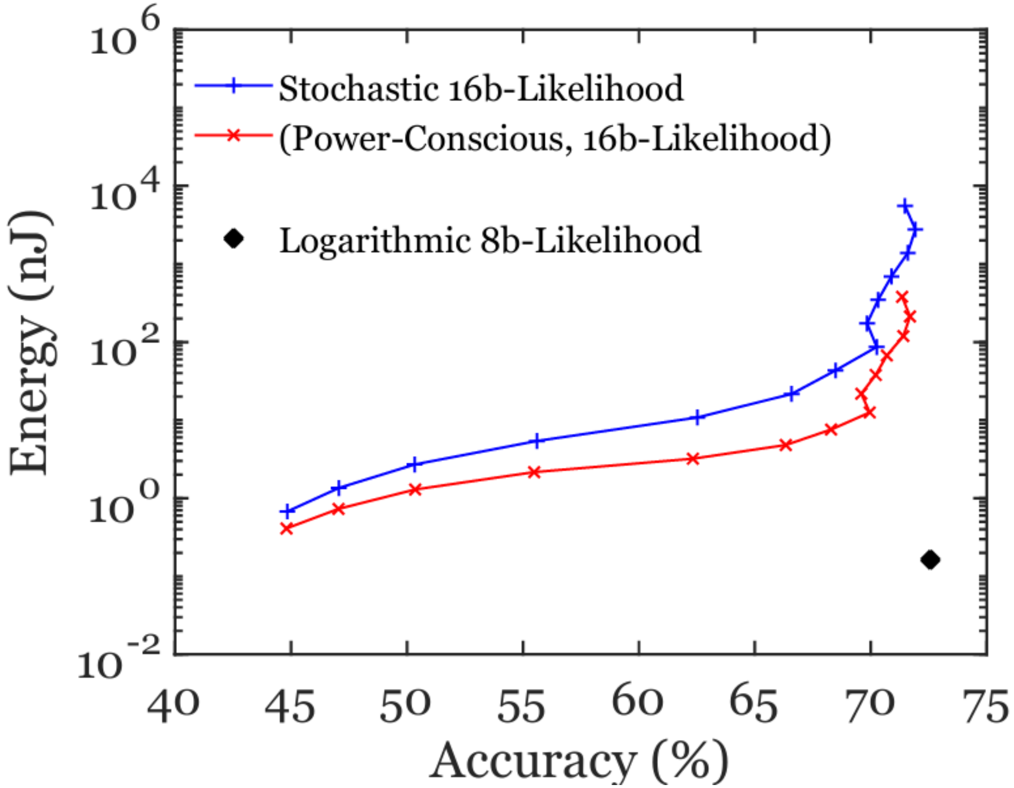
$$p_1 \times p_2 = 0.5^{(n_1+n_2)/m}$$

Probabilities represented logarithmically with integers

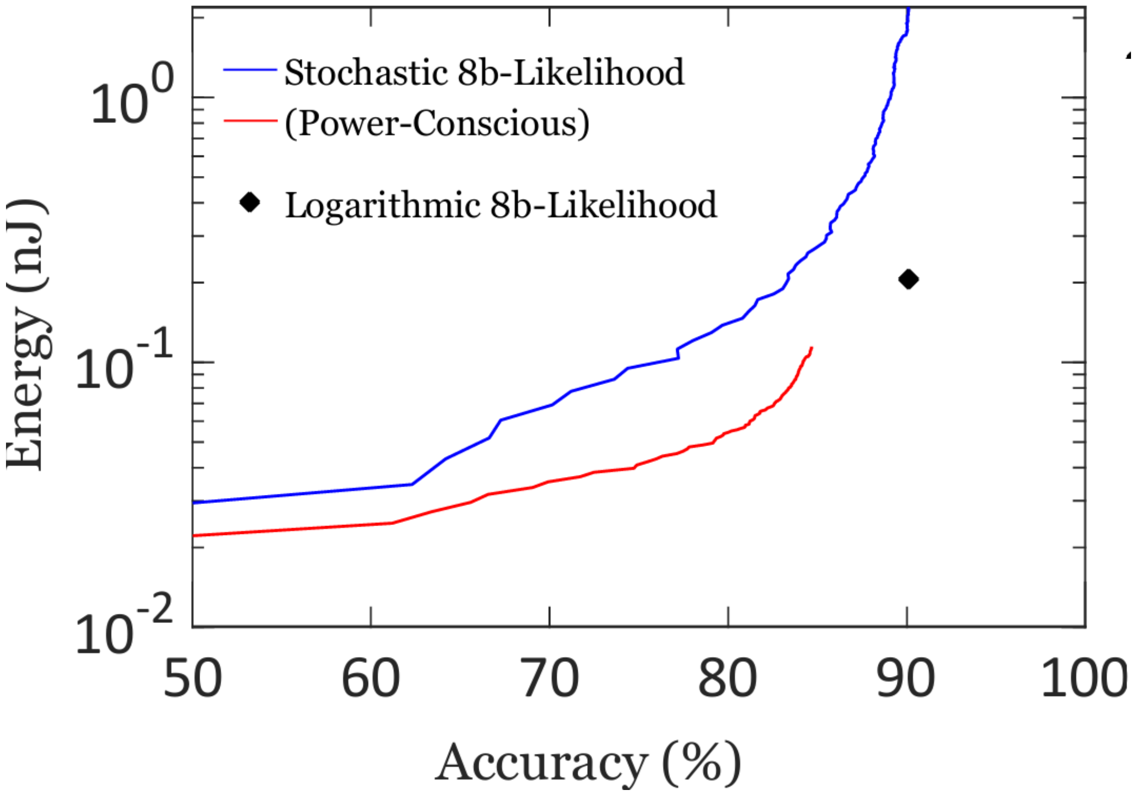
Products becomes integer sum

Energy Efficiency of Bayesian Machines

Sleep Cycle Classification



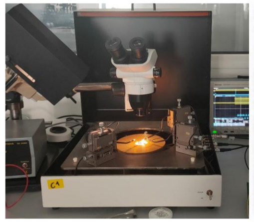
Gesture Recognition

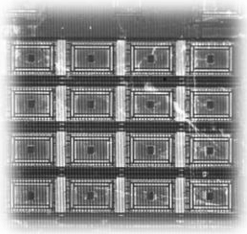


STM32 Microcontroller units: a few microjoules

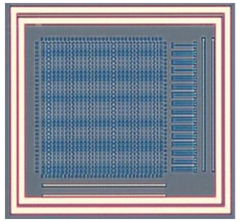
Turck et al, Communications Engineering (2025)

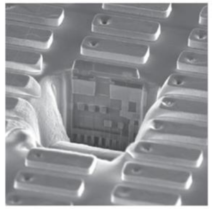
Toward Energy-Efficient and Trustworthy Embedded AI: The In-Memory Computing Approach



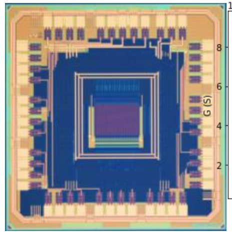


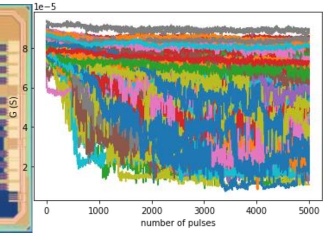
- Cointegrating Logic & Memory for Low-Energy AI





- The Analog & Stochastic Paths



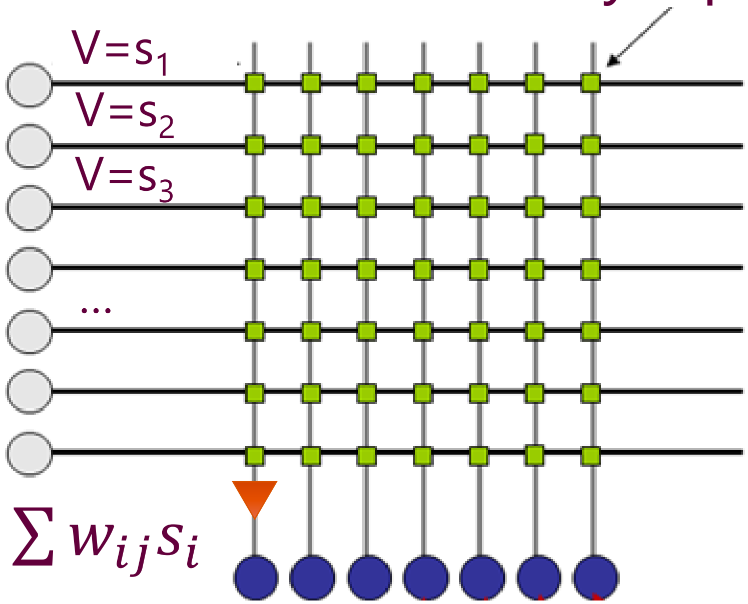


- Learning with Physics

Analog In-Memory Computing Performs Neural Network Inference Very Naturally

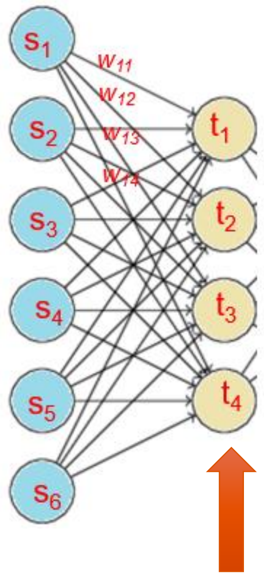
A matrix of analog memristors naturally implements a layer of neural network with **Ohm's and Kirchhoff's laws**!

Memristor conductance G = synaptic weight w



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

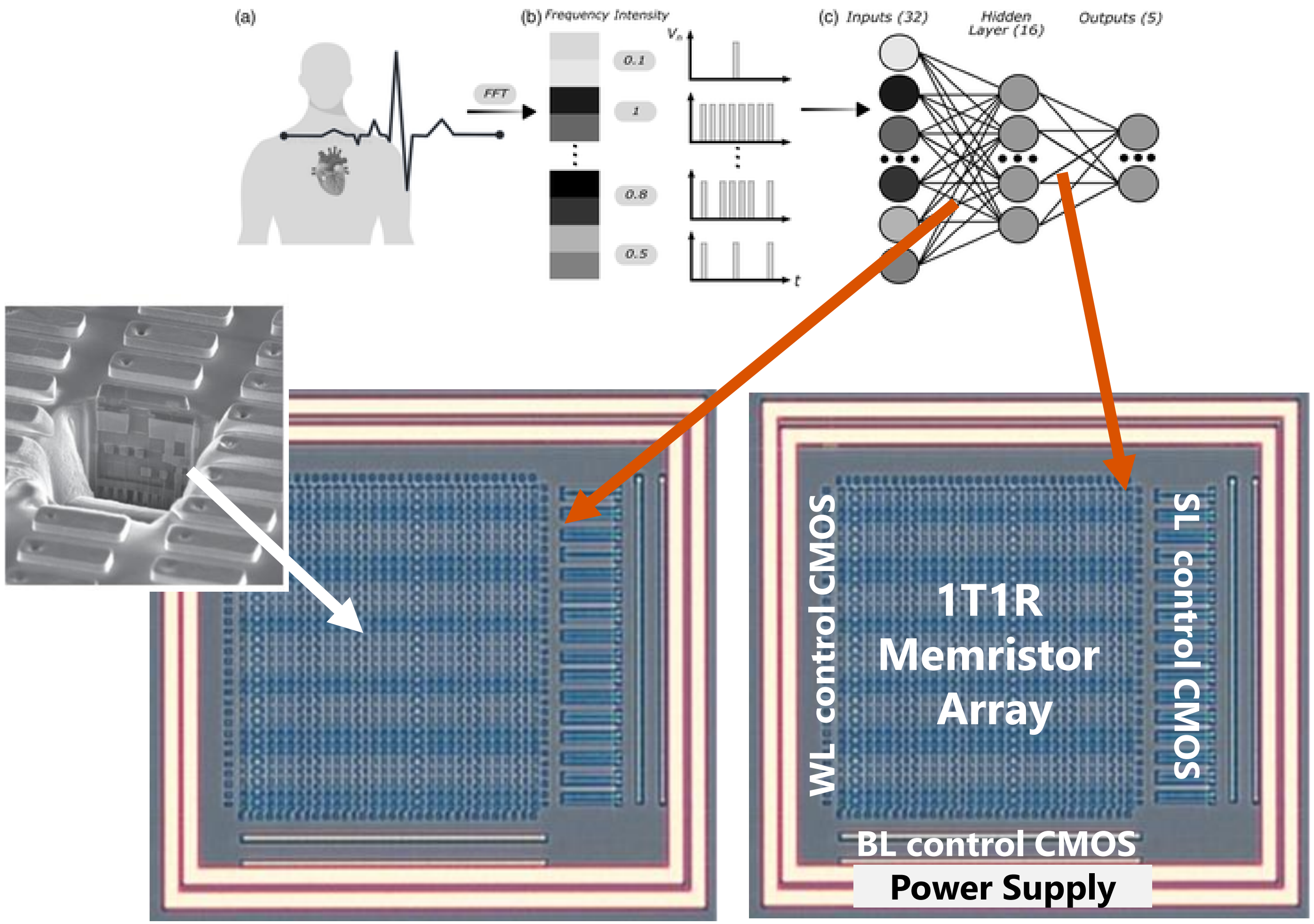
$\equiv$



Neuron value

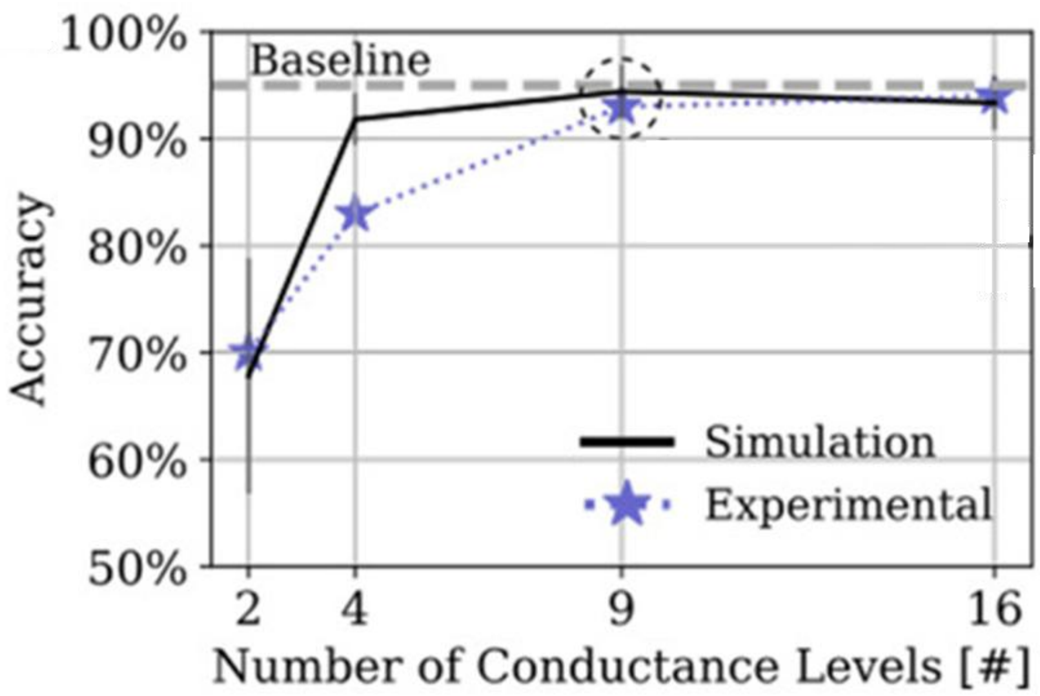
$t_j = f\left(\sum w_{ij} s_i\right)$

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

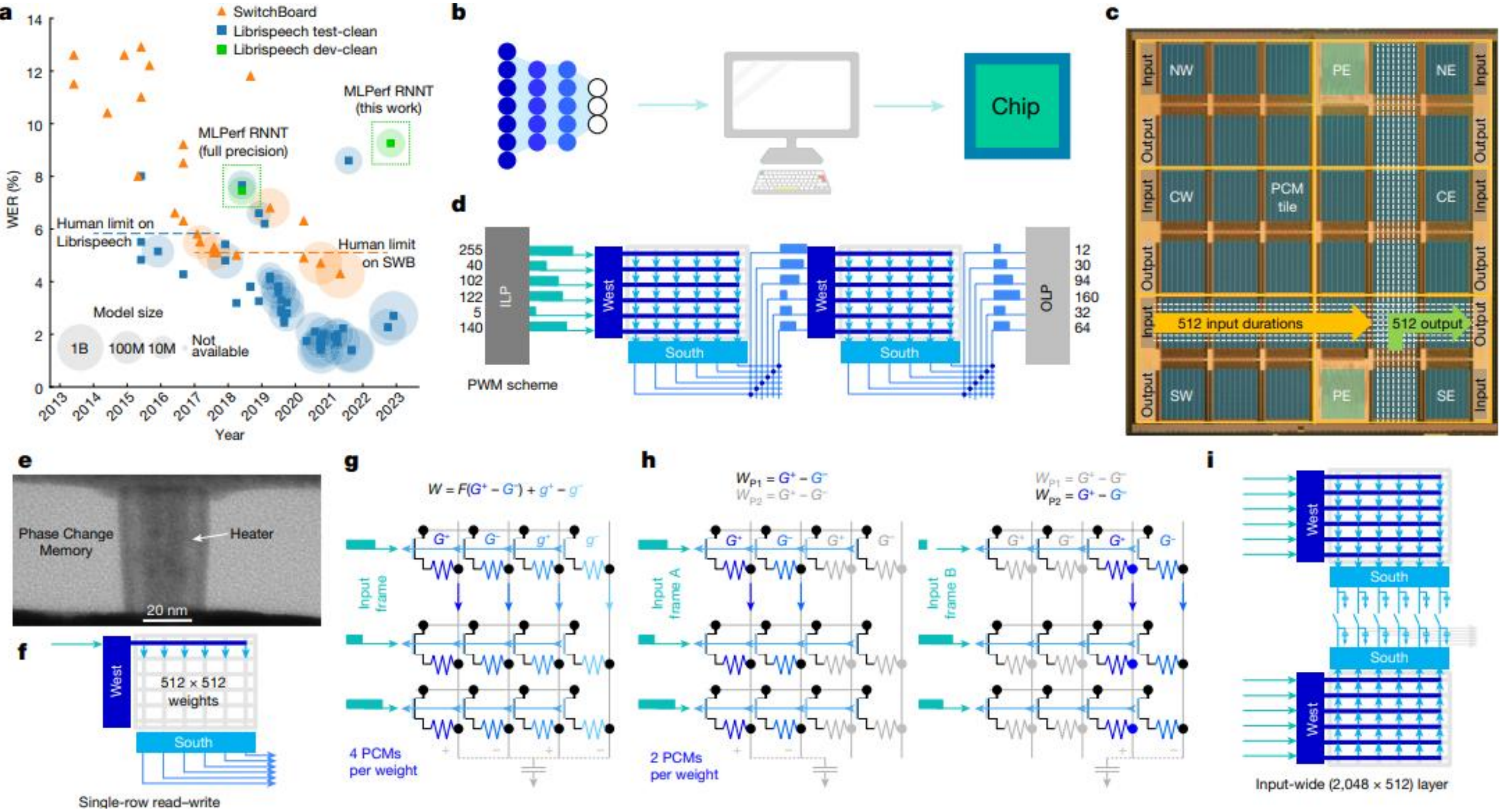
An analog-AI chip for energy-efficient speech recognition and transcription

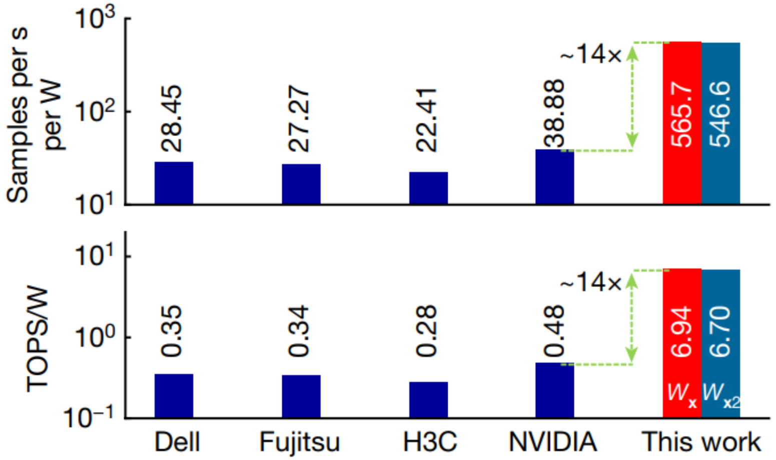
https://doi.org/10.1038/s41586-023-06337-5

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]

Nature 2023





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

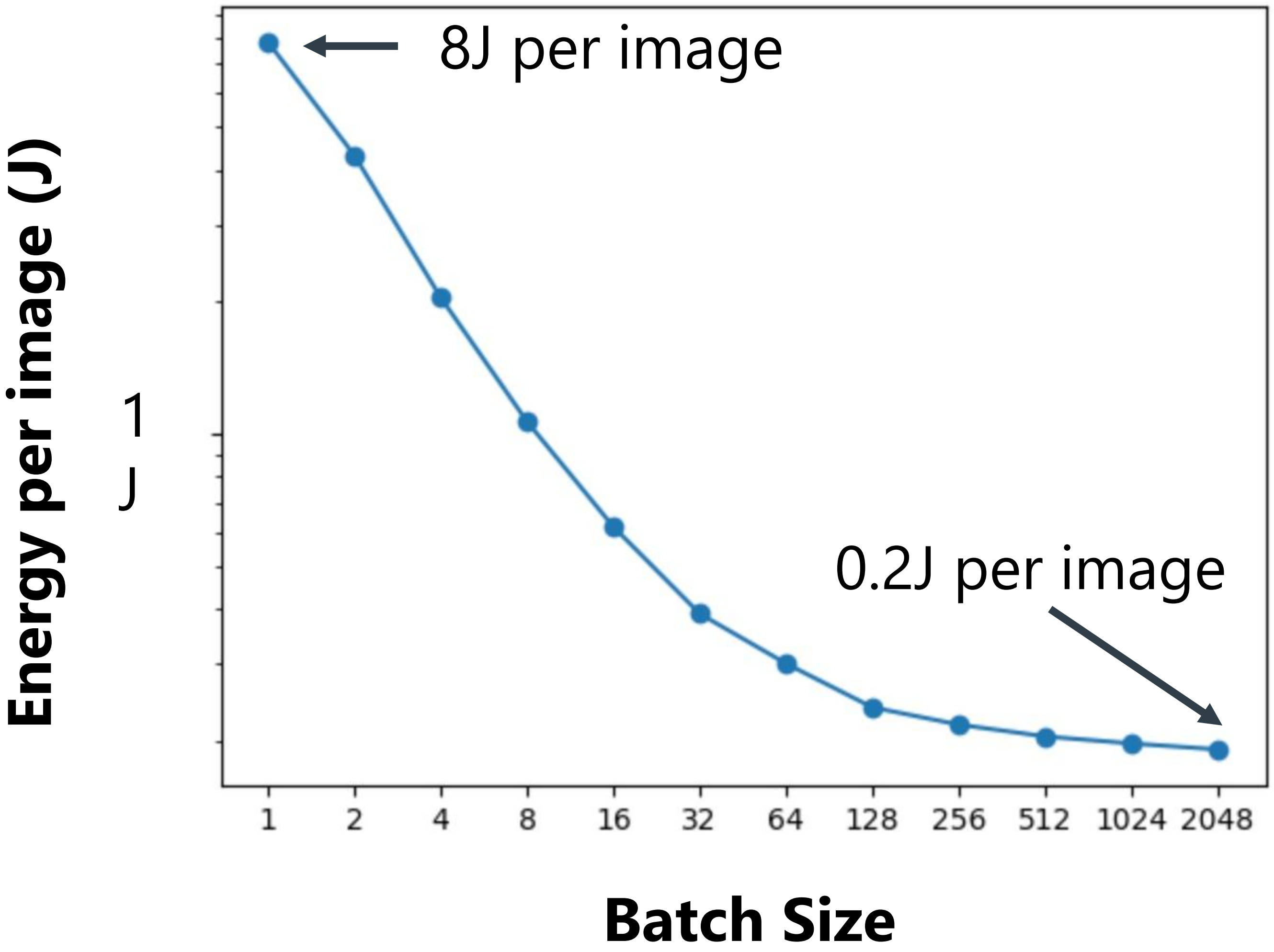
- 35M phase change memories

« 7 TOPS/W Is Not That High »

- Only a prototype

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

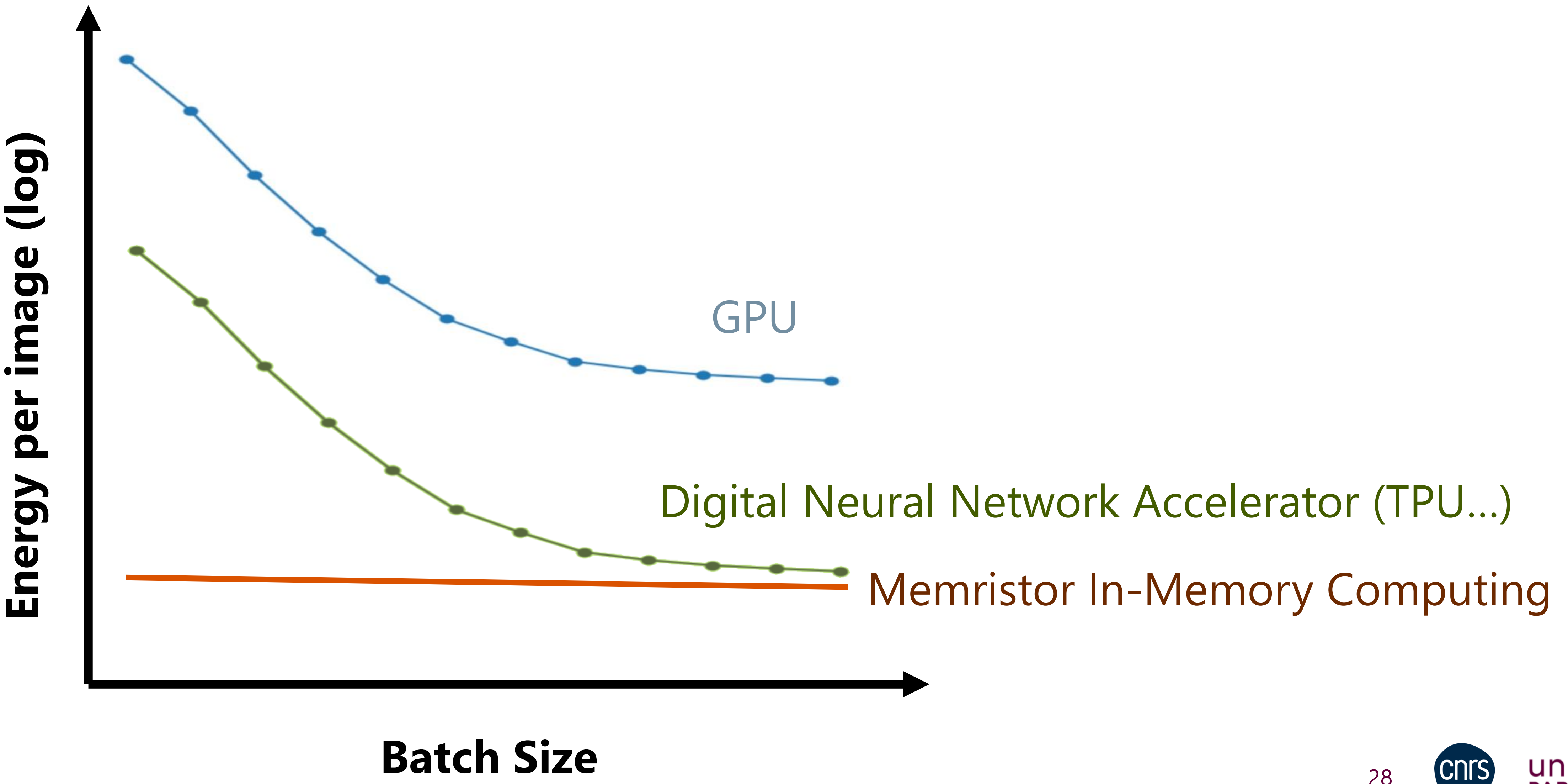
Real Energy Consumption Can Be Dramatically Higher than Peak



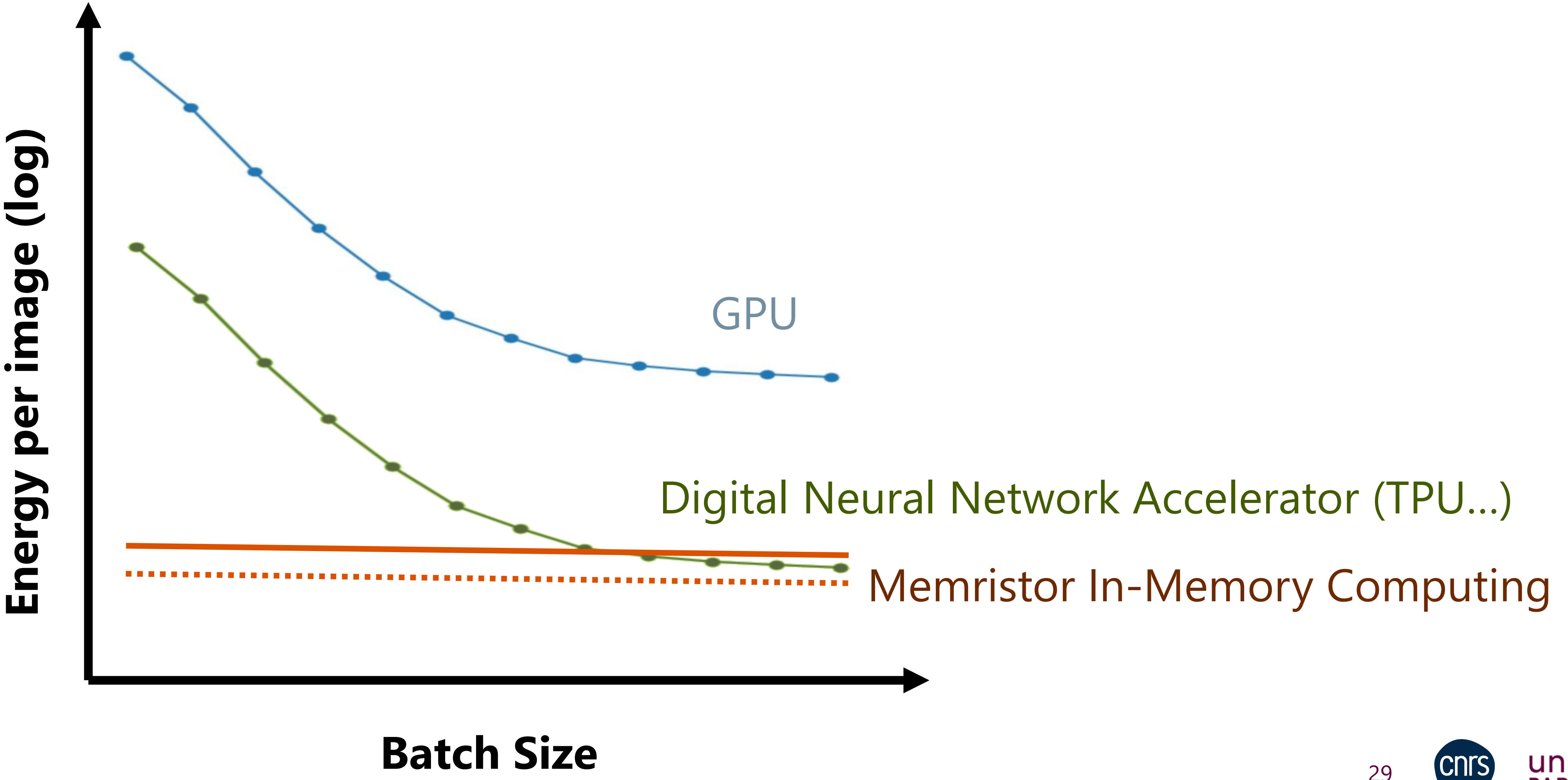
NVIDIA A100 GPU
ResNet50

Courtesy of David Novo (Univ Montpellier/CNRS)

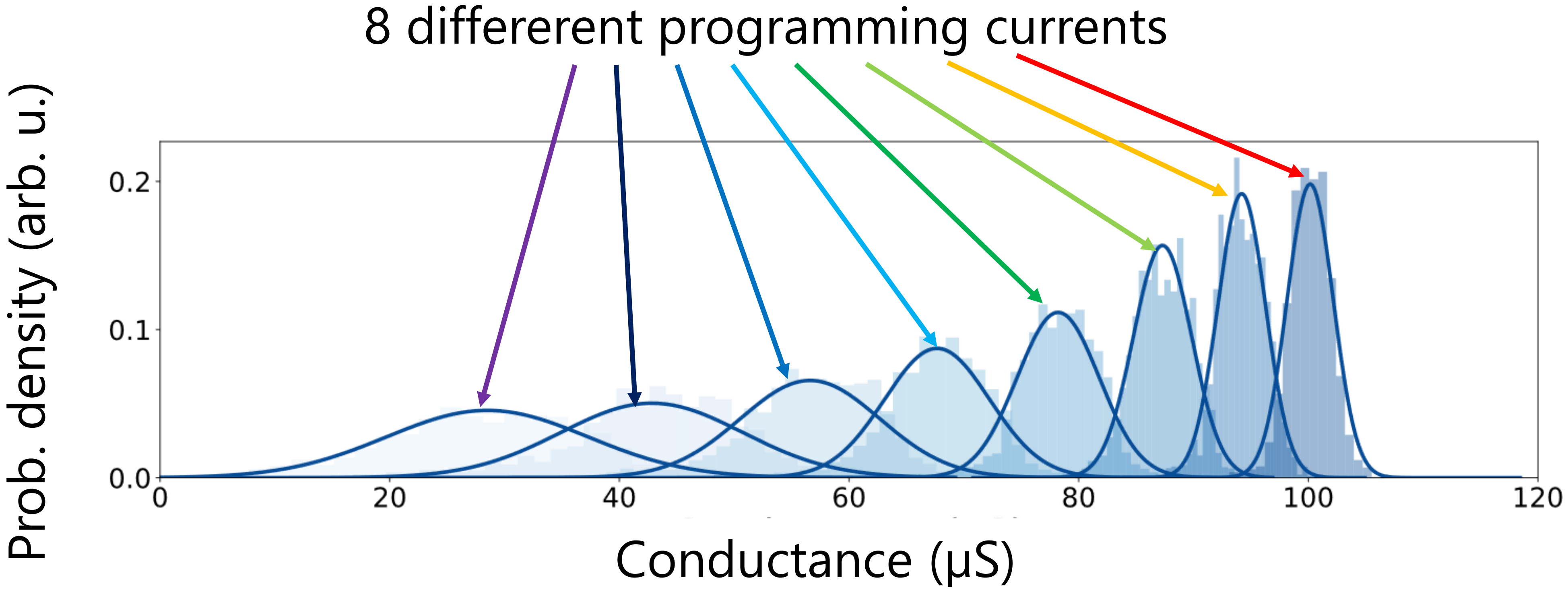
How I View Architecture Research



How I View Architecture Research

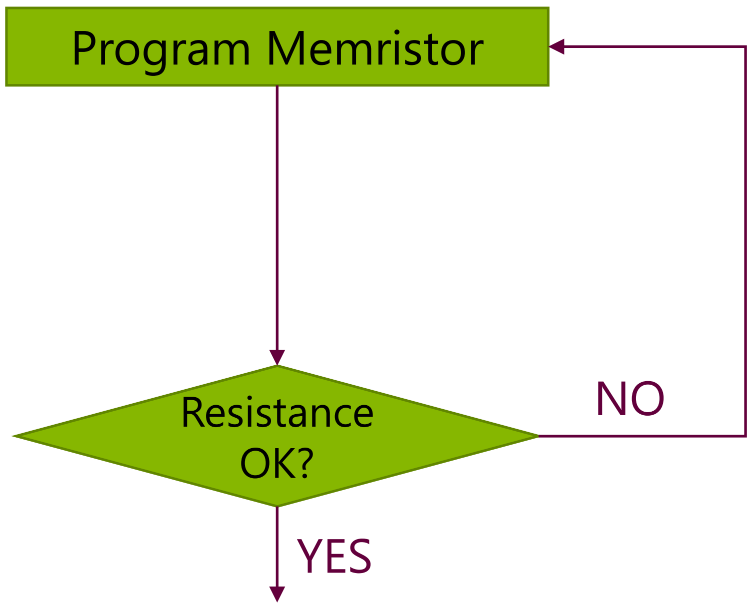


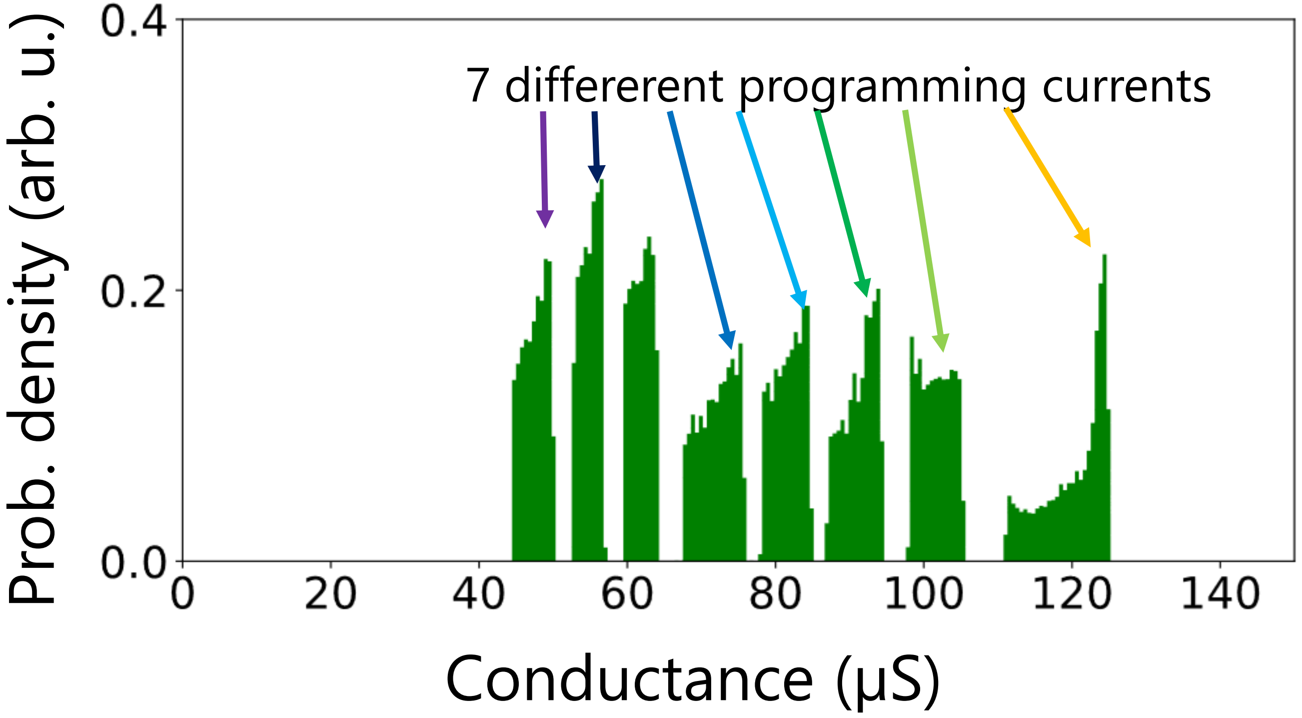
The Challenge of Analog In-Memory Computing: Memristors Act as Random Variables



Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

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



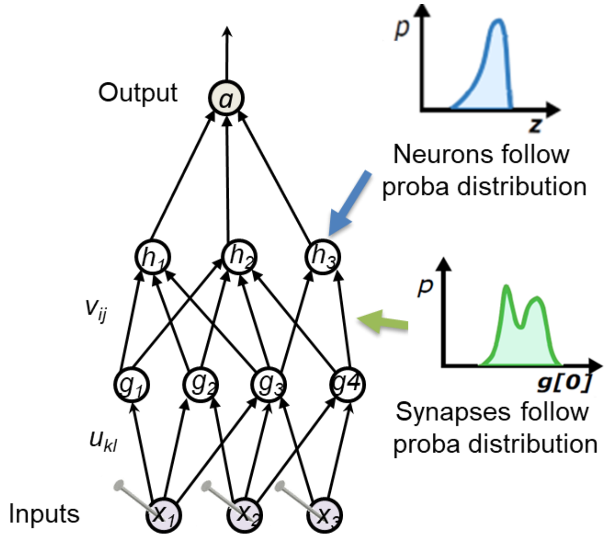


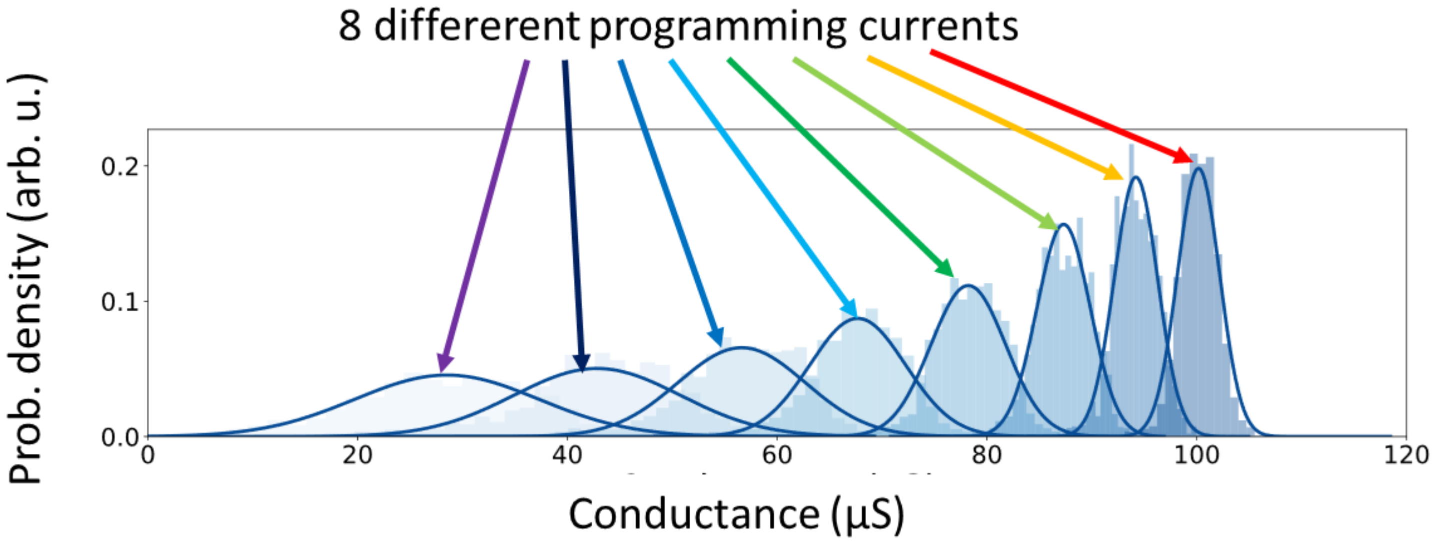
Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

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)

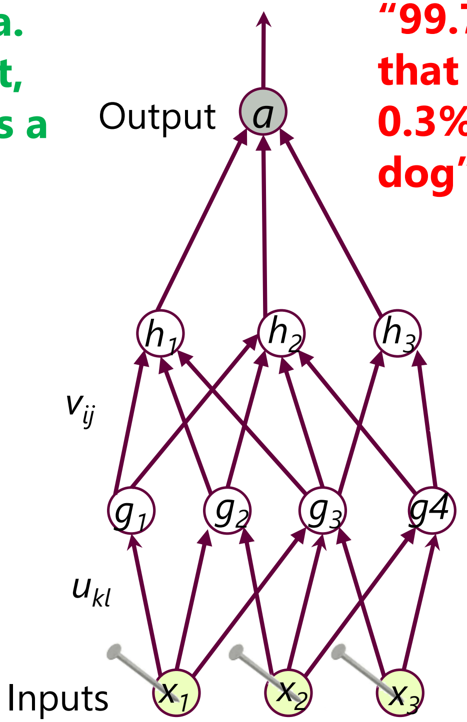
Collaboration

Bayesian Neural Networks

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

Conventional Neural Network

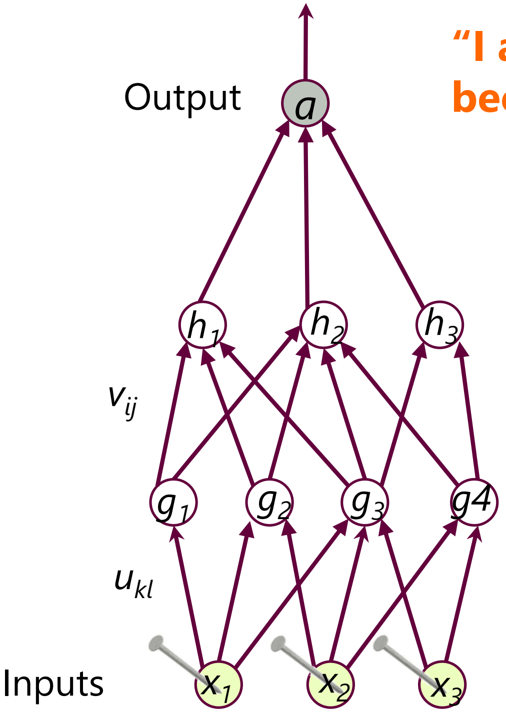
"99.9% proba. that this a cat, 0.1% that it is a dog"



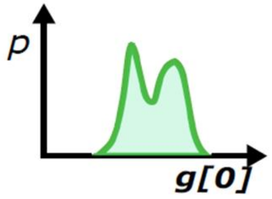
"99.7% proba. that this a cat, 0.3% that it is a dog"

0.7
Each synapse has one weight

Bayesian Neural Network



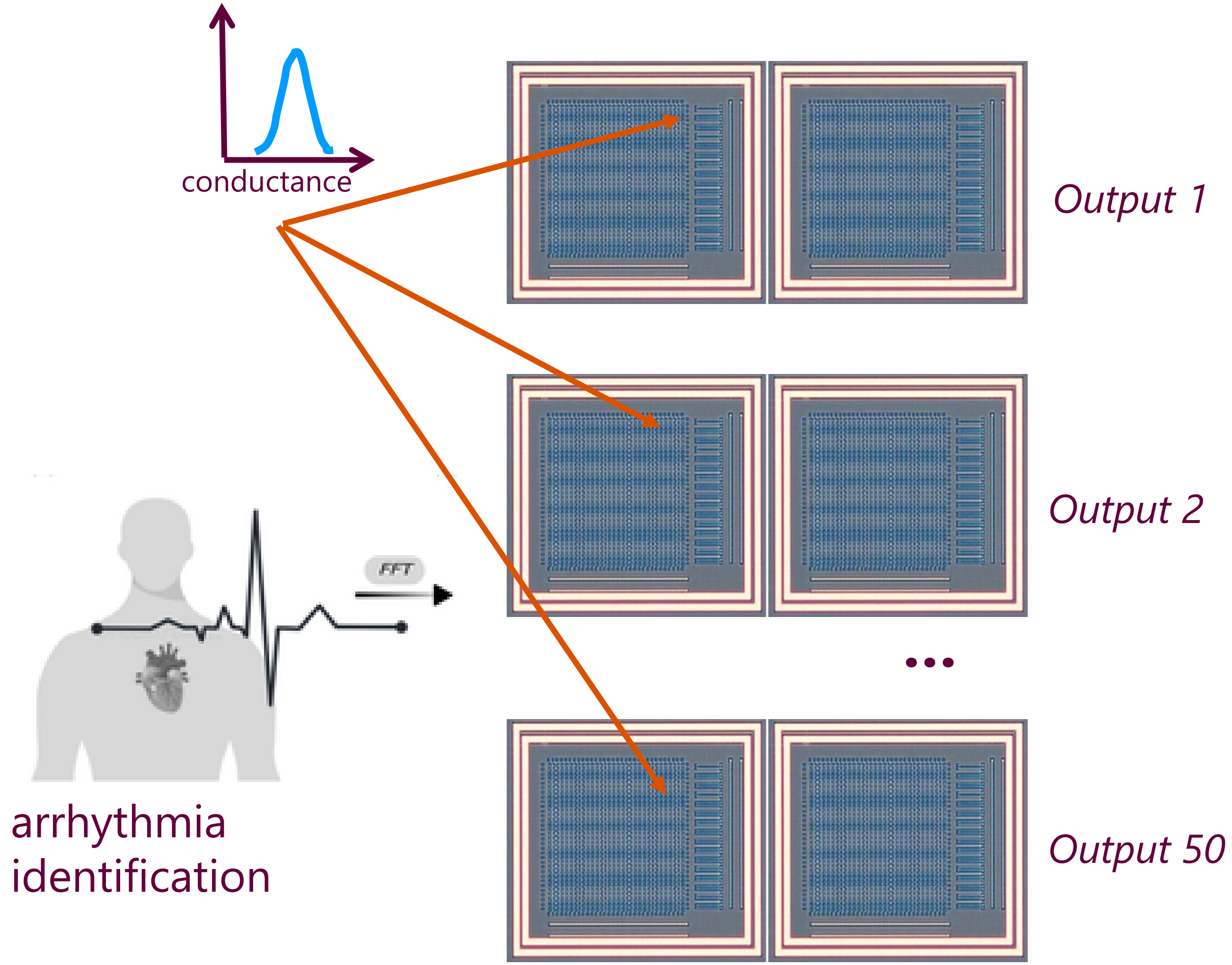
"I am not really sure what this is, because I have never seen it"



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



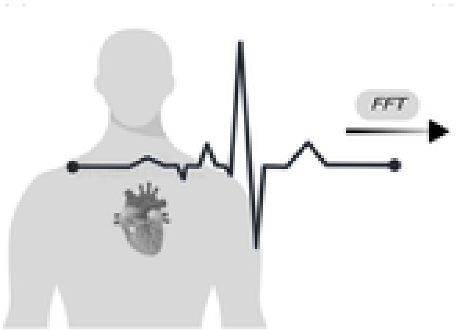
We get 50 outputs:
their dispersion tells about the *certainty* of the neural network

<u>Neural network trained with Variational Inference incorporating a specific "technological loss"</u>

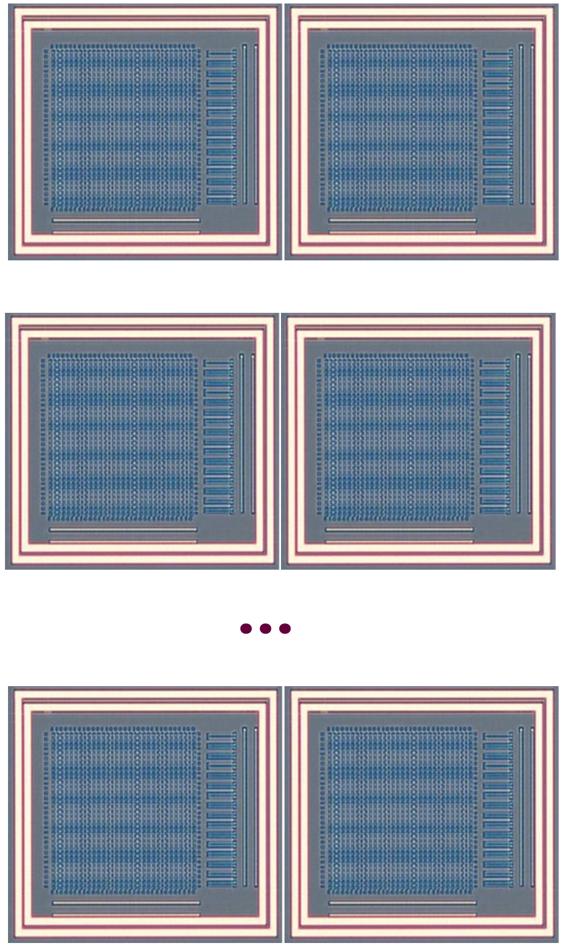
Bonnet et al, Nature Communications 14, 7530 (2023)

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

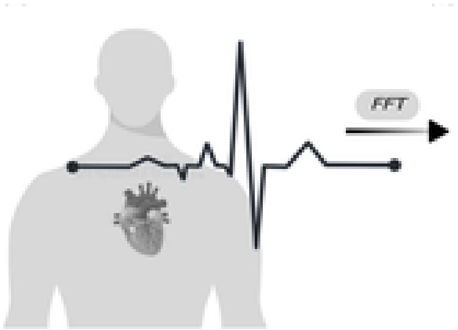
Arrythmia type 1

LOW UNCERTAINTY

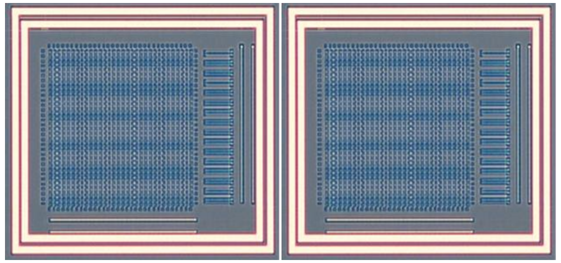
Bonnet et al, Nature Communications 14, 7530 (2023)

Memristor-Based Bayesian Neural Networks

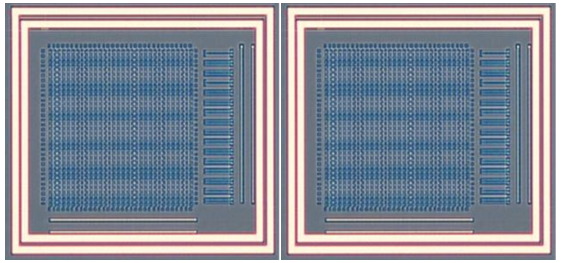
We program 50 memristor arrays, and we apply the same input to them



arrhythmia
identification

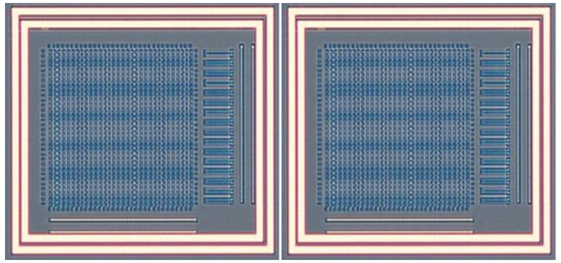


Arrythmia type 1



Hesitates
arrythmia type 1 or 2

...



Hesitates
arrythmia type 3 or 4

Unknown data:

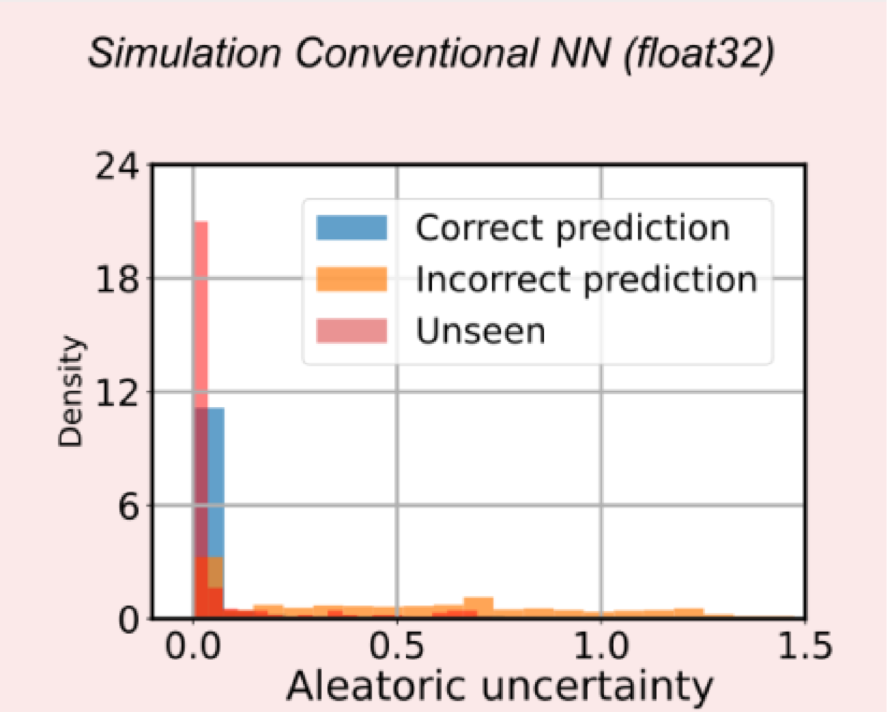
EPISTEMIC UNCERTAINTY

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

Fully Experimental Arrythmia Recogniton

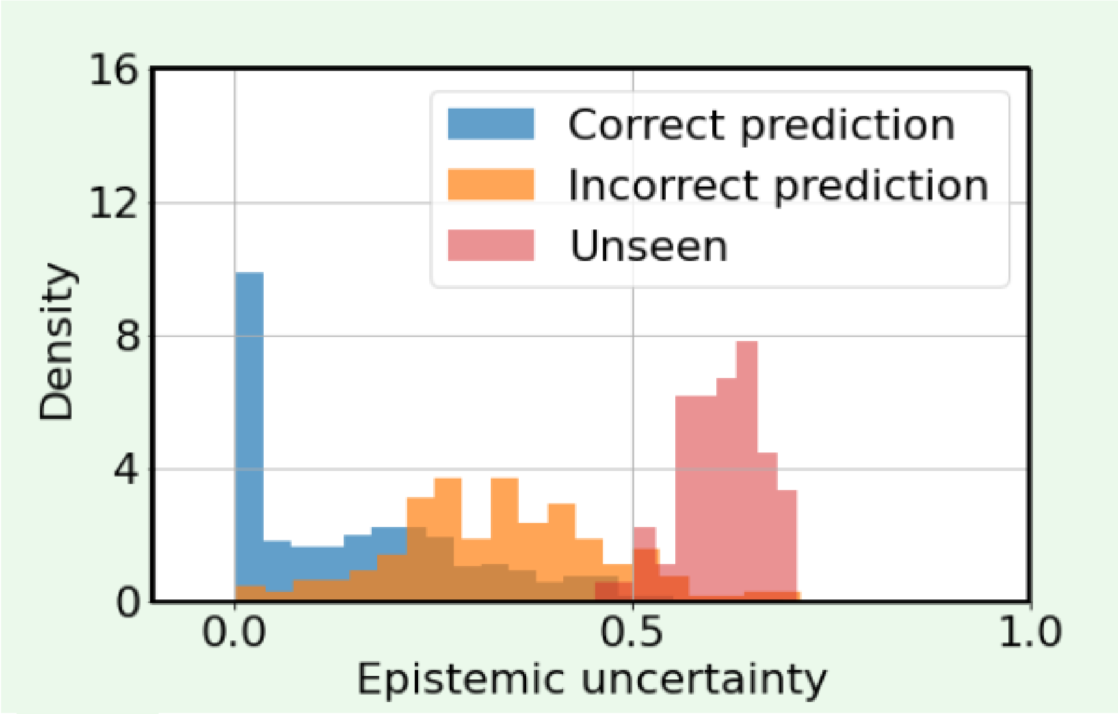
Traditional Neural Network

80% classification accuracy



Bayesian Neural Network

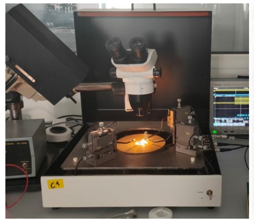
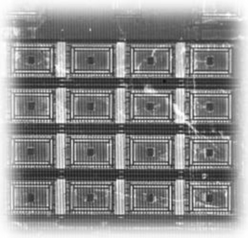
79% classification accuracy



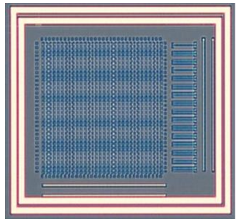
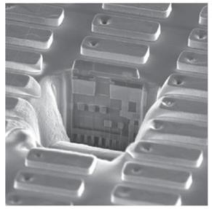
Easily recognizes
unknown types of arrythmia

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

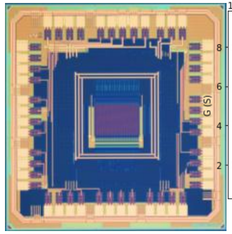
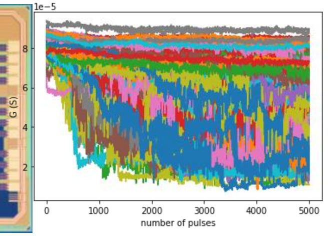
Toward Energy-Efficient and Trustworthy Embedded AI: The In-Memory Computing Approach

- Cointegrating Logic & Memory for Low-Energy AI

- The Analog & Stochastic Paths

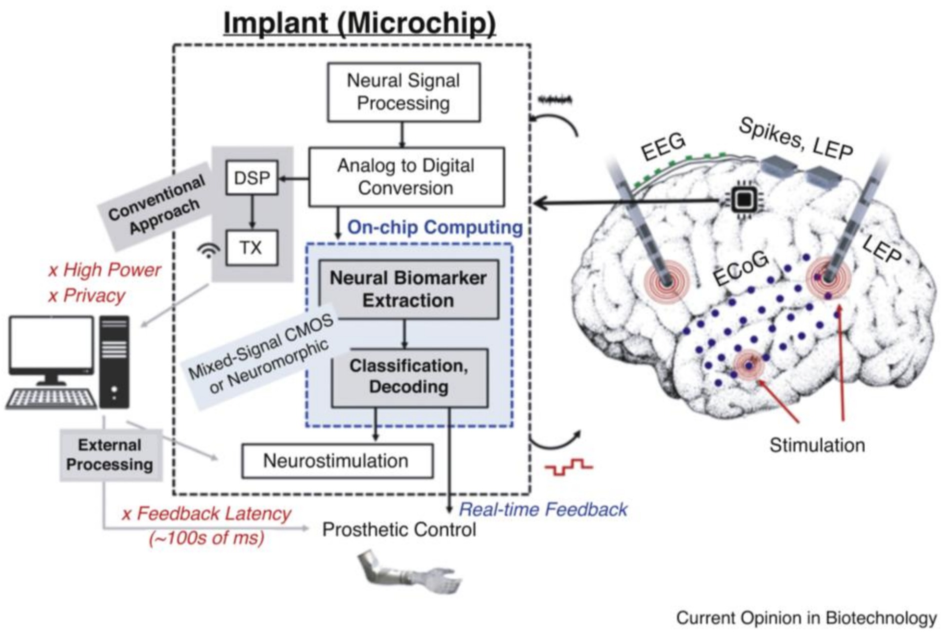
- Learning with Physics

The Grand Challenge of On-Chip Learning

- Current in-memory computing AI accelerators are focused on inference, which makes sense

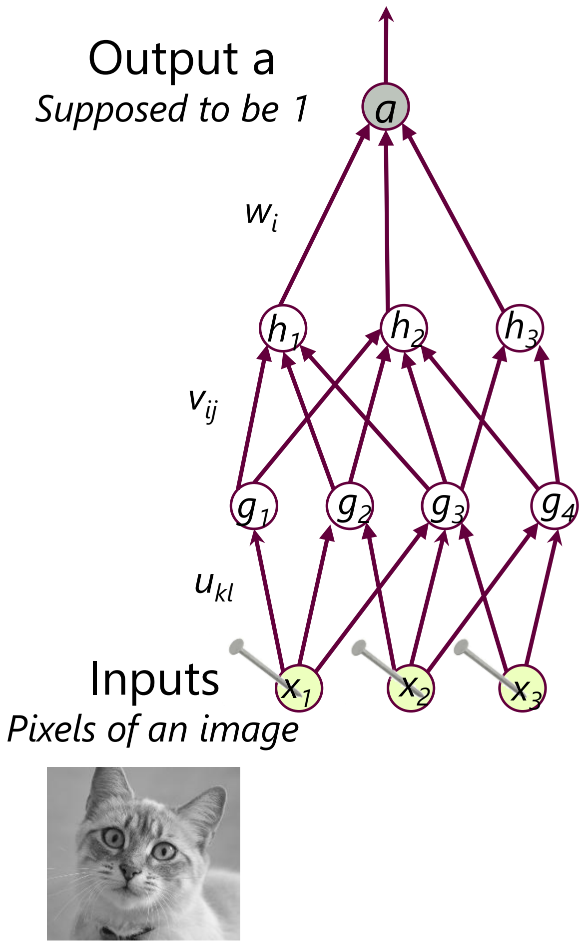
- On-chip learning/adaptation would also have tremendous prospects

Medical applications



Yoo and Shoaran,
Current Opinion Biotechnol., 2021

Backpropagation, the Canonical Method for Training Networks, Is Not Adapted for In-Memory Computing



$$C = -\ln a$$

$$\frac{\partial C}{\partial w_i} = (a-1)h_i$$

$$\frac{\partial C}{\partial v_{ij}} = (a-1)w_i h_i(1-h_i)x_j$$

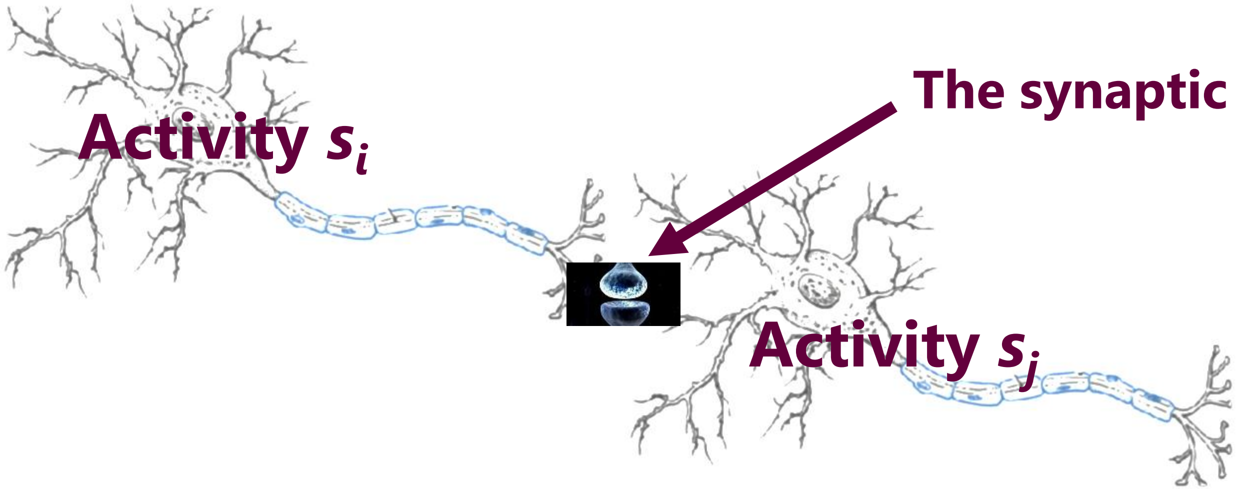
$$\boxed{\frac{\partial C}{\partial u_{kl}} = (a-1)\sum_i w_i h_i(1-h_i)v_{kl}g_k(1-g_k)x_l}$$

To update weight u_{kl}, you need calculation that involves information about the *whole* network!

Learning in the Brain Is Mysterious

- Synapses are physical objects that have only access to local information

- An old idea (Hebb): « cells that fire together wire together »

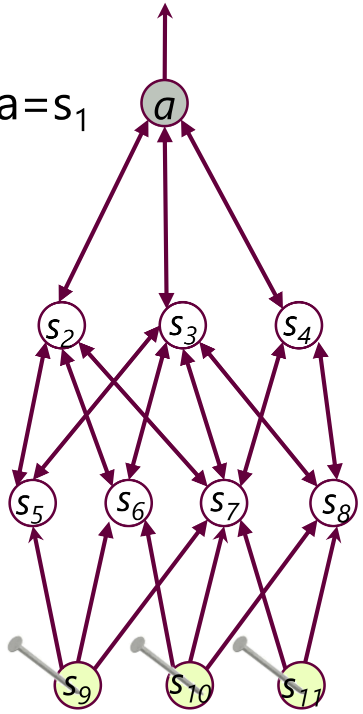


- **Multiple theoretical works show that Hebbian learning rules (e.g., STDP) are not as powerful as backprogation**

EqProp: a Type of Neural Networks Grounded in Physics and with « Brain-like » Learning

<u>Learns with Hebbian-like learning but is equivalent to backpropagation</u>

Output a=s_1



Inputs

Pixels of an image

Energy $E = \frac{1}{2}\sum_i {s_i}^2 - \frac{1}{2}\sum_{i,j} w_{ij}\rho(s_i)\rho(s_j)$

s_i : neuron states, ρ : sigmoid function

The neural network is a dynamical system that goes naturally toward its energy minimum

$$\frac{ds_i}{dt} = -\frac{\partial E}{\partial s_i} \Rightarrow \frac{ds_i}{dt} + \underbrace{s_i = \left[\sum_j w_{ij}\rho(s_j)\right]\rho'(s_i)}$$

Local: depends only on s_i nearest-neighbors

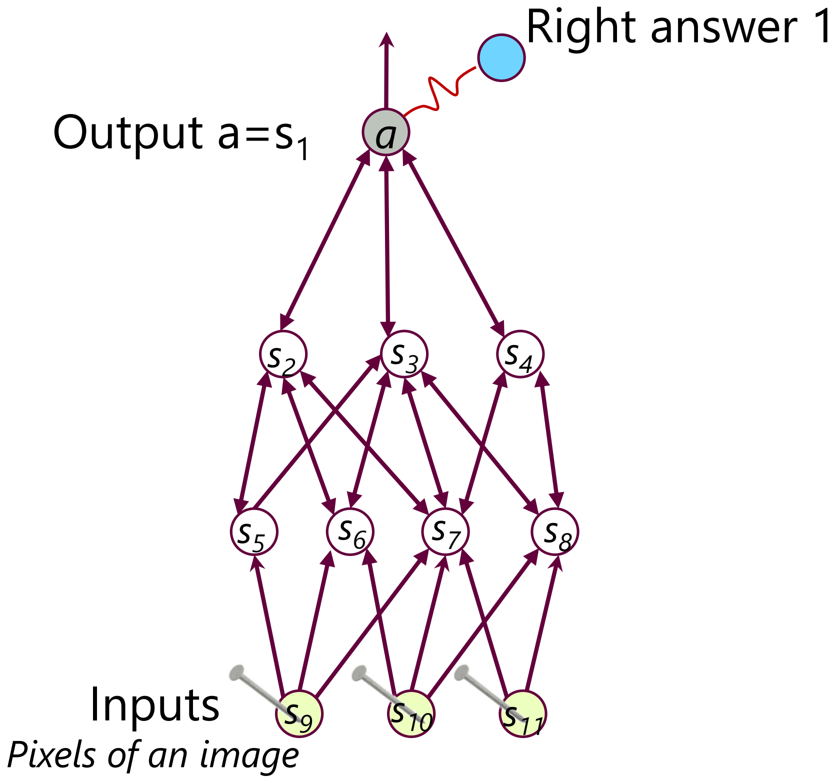
Similar to leaky integrate-and-fire neuron

When the network has converged, we get the output *a*

Scellier & Bengio, fnins 2017

cnrs université PARIS-SACLAY

If the Output a Is Not the Right Value, EqProp then "Nudges" the Network Toward It



Error $C = \frac{1}{2}(a-1)^2$

Nudged Energy $F = E + \beta C$

$$\frac{da}{dt} + a = \left[\sum_j w_{ij}\rho(s_j)\right]\rho'(a) + \underbrace{\beta(a-1)}_{\text{« Nudging »}}$$

The perturbation of the output *a* propagates throughout the network that reaches a new equilibrium

s_{nudged}

Scellier & Bengio, fnins 2017

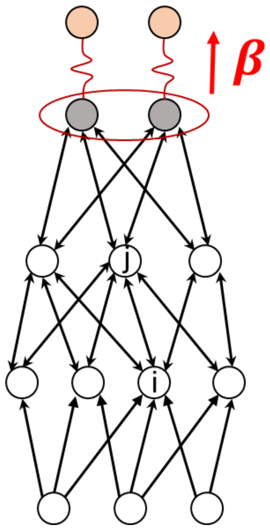
Equilibrium Propagation

- Change the weight w_{ij} by
- $\rho(s_{nudged,j})\rho(s_{nudged,i}) - \rho(s_{free,j})\rho(s_{free,i})$

➢ *If neurons i and j had MORE Hebbian correlation during the nudged phase, then INCREASE their connection w_{ij}*

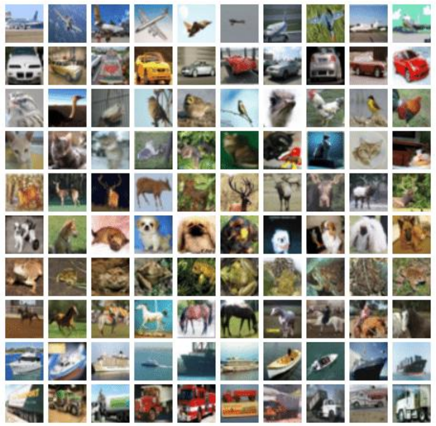
➢ *Otherwise, DECREASE w_{ij}*

The Local Learning Rule of Equilibrium Propagation Leads to High Recognition Rates



Mathematical equivalence of EP gradients with Backpropagation Through Time

M. Ernoult, J. Grollier, D. Querlioz, Y. Bengio, B. Scellier, NeurIPS (2019)



EP scales to CIFAR-10

A. Laborieux, M. Ernoult, B. Scellier, Y. Bengio, J. Grollier, D. Querlioz, fnins 15, 129 (2021)

EP trains Binary Neural Networks

- Binary synapses (CIFAR 10)
- Binary synapses and neurons (MNIST)
- Ternary gradients

J. Laydevant, M. Ernoult, D. Querlioz, J. Grollier, CVPR (2021)

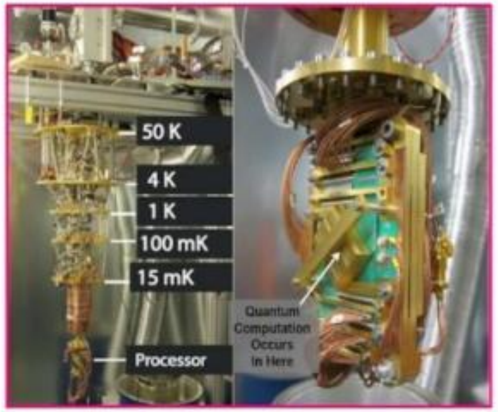
Collaboration

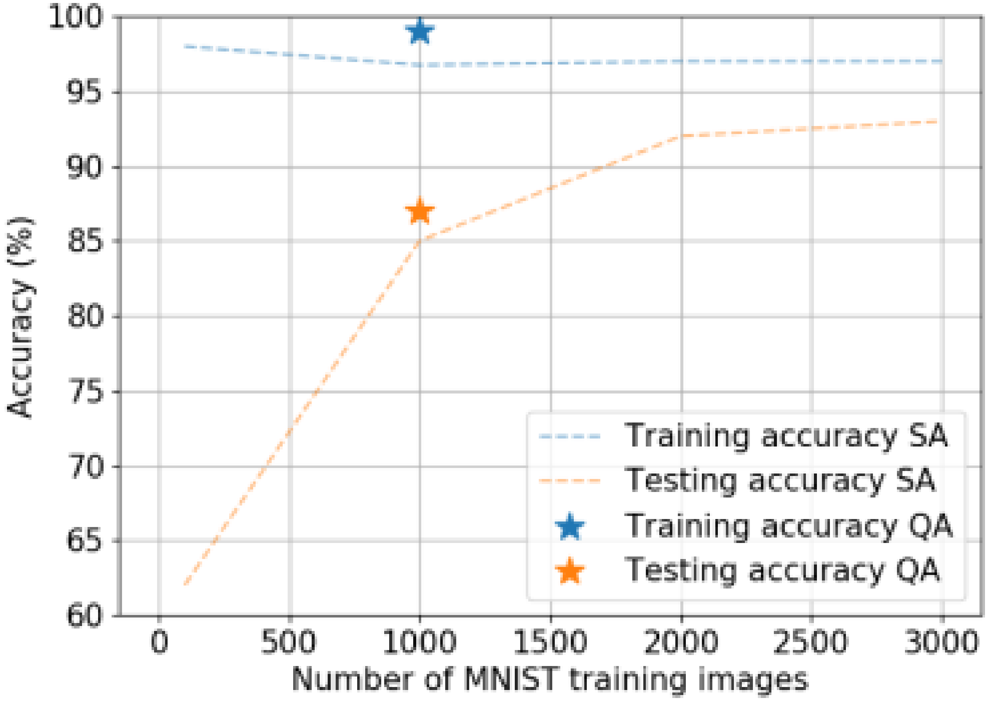
laboratoire Albert Fert C2N Mila

45

cnrs université PARIS-SACLAY

EqProp Can Train a Quantum Ising Machine





- EqProp adapts very naturally to physical systems that compute

Laydevant, Markovic and Grollier, Nature Communications 15, 3671 (2024)

laboratoire Albert Fert
cnrs
université PARIS-SACLAY

Conclusion

- Computing with nanoelectronics devices: huge potential to reinvent electronics for **cognitive-type tasks and AI**

- Memristors are rich devices that can be used in a variety of ways

- Memristors are particulary adapted for Bayesian approaches, paving the way toward trustworthy AI

- **Next grand challenges: scaling and solving learning**

- **Very important & creative time for micro/nano-electronics research. Considerable benefits from algorithm/electronics/technology research**

Acknowledgments

- Kamel-Eddine Harabi
- Clément Turck
- Tifenn Hirtzlin
- Atreya Majumdar
- Marie Drouhin
- Jacques-Olivier Klein
- Maxence Ernoult
- Axel Laborieux
- Adrien Renaudineau
- Théo Ballet

- Elisa Vianello
- Tifenn Hirtzlin
- Eduardo Esmanhotto
- Djohan Bonnet
- Niccolo Castellani
- François Andrieu

- Stéphane Collin
- Oleksandr Bilousov
- Amadéo Michaud

- Julie Grollier
- Jérémie Laydevant
- Maxence Ernoult
- Marie Drouhin

- Jacques Droulez
- Pierre Bessière

- Yoshua Bengio
- Benjamin Scellier

- Jean-Michel Portal
- Marc Bocquet
- Fadi Jebali

- 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/