# Analog Machine Learning Accelerators

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.

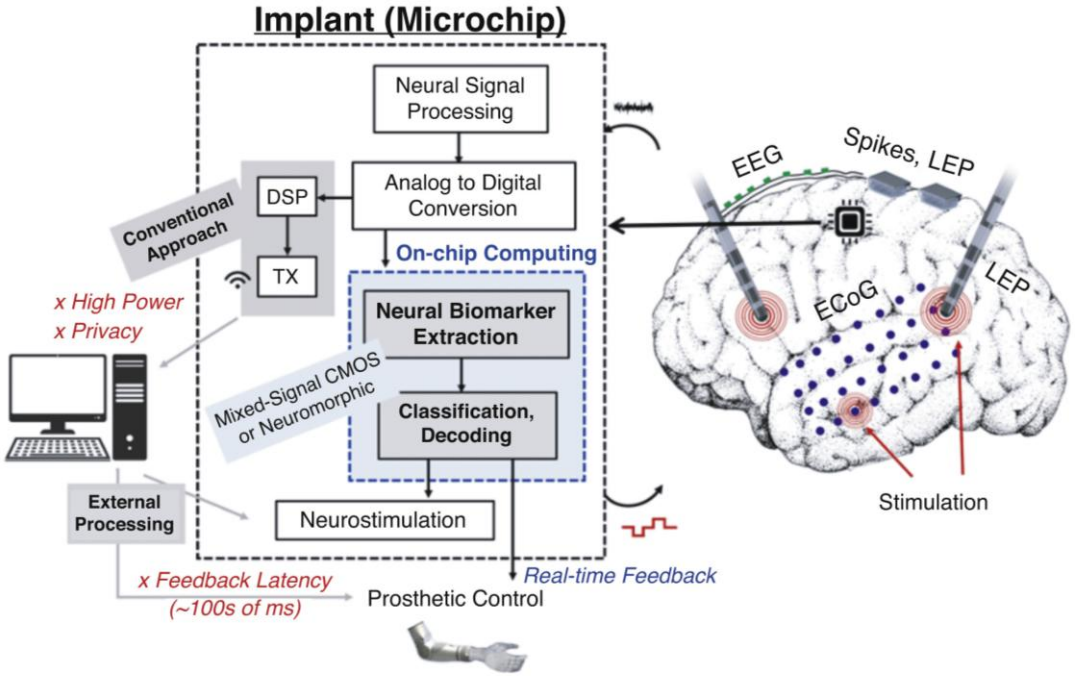
Speaking at a Bloomberg event on the sidelines of the World Economic Forum's annual meeting in Davos, Altman said the silver lining is that more climate-friendly sources of energy, particularly nuclear fusion or cheaper solar power and storage, are the way forward for AI.

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



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**

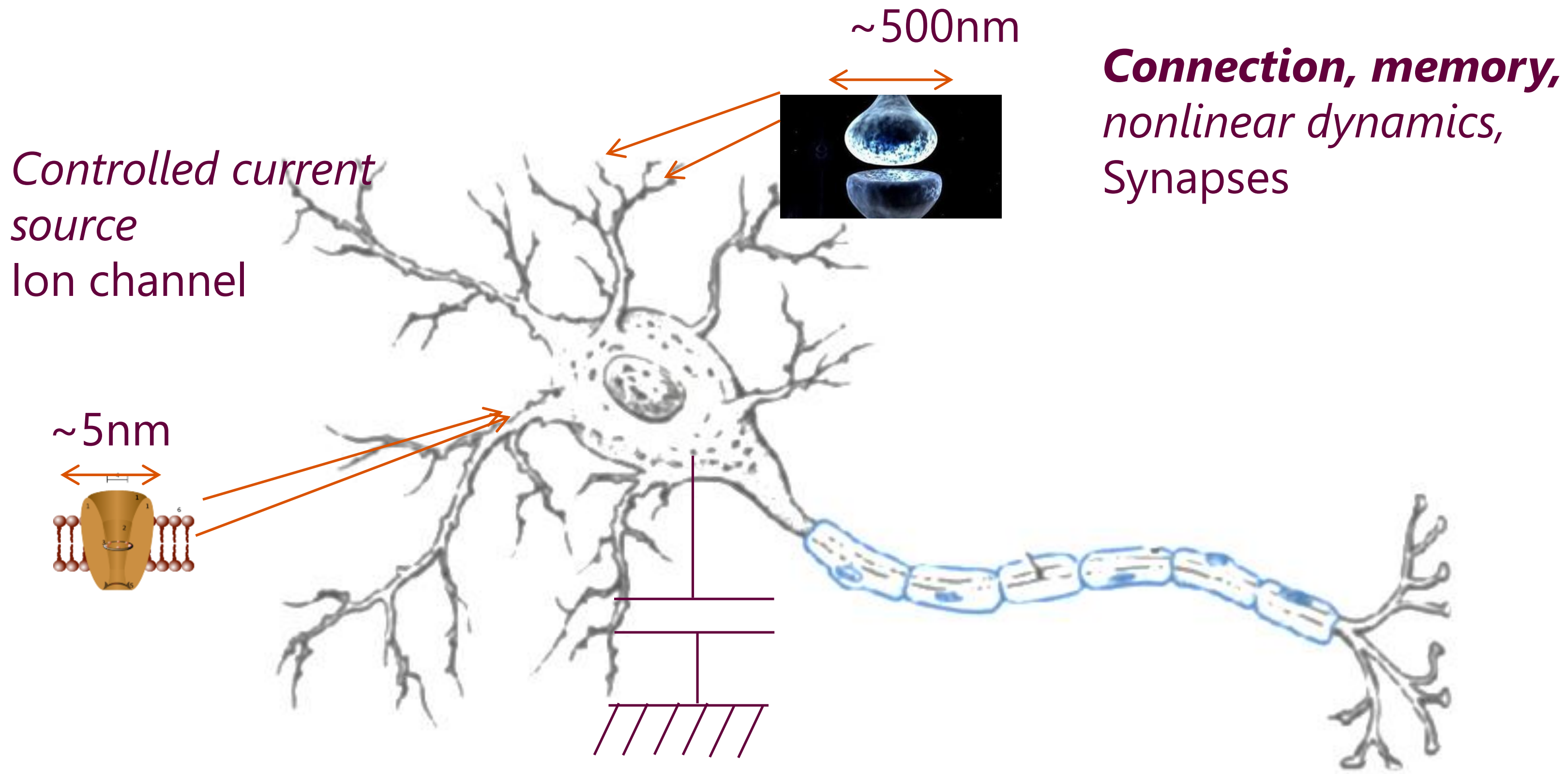
- **In computers, GPUs, and AI accelerators developed by industry (Google TPU, Apple NPU…), memory access is extremely costly**

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

Absent in the brain!

Pedram et al , IEEE D&T 2016

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

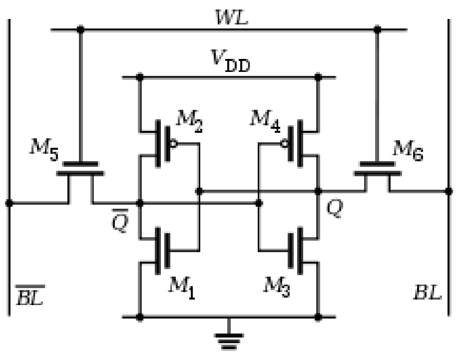



**10,000 times more synapses than neurons!**

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

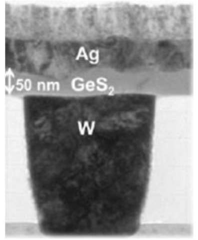
# Integrating Logic & Memory Is a Considerable Challenge

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



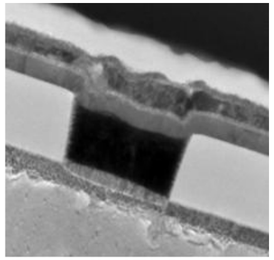

1 bit of SRAM

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




*Phase Change Memory*

*CEA-LETI*



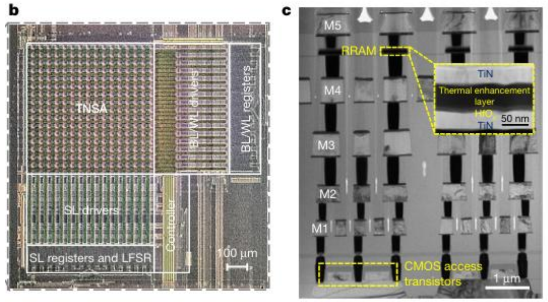
*Memristors, Oxide Resistive Memory*

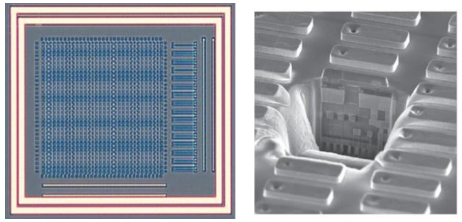
*CEA-LETI*

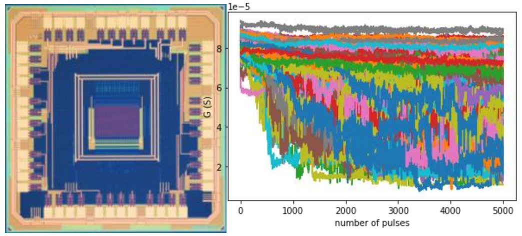
*Spin Torque Memory*

*Samsung*

# Analog Machine Learning Accelerators

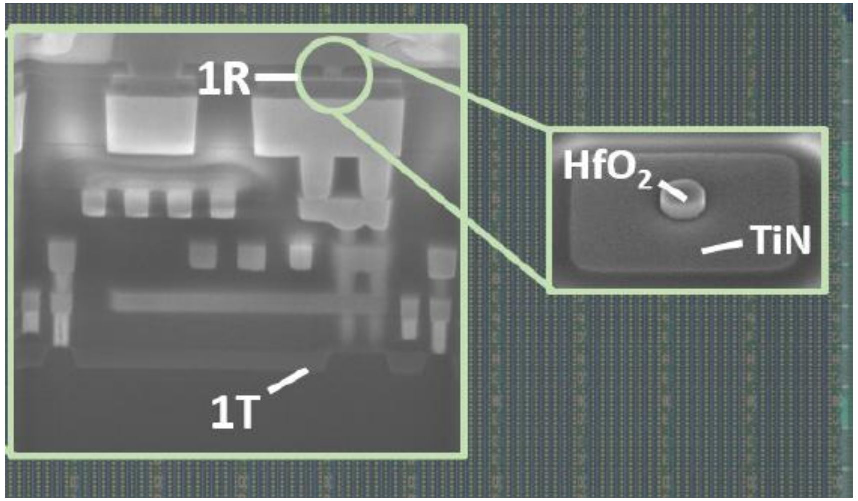


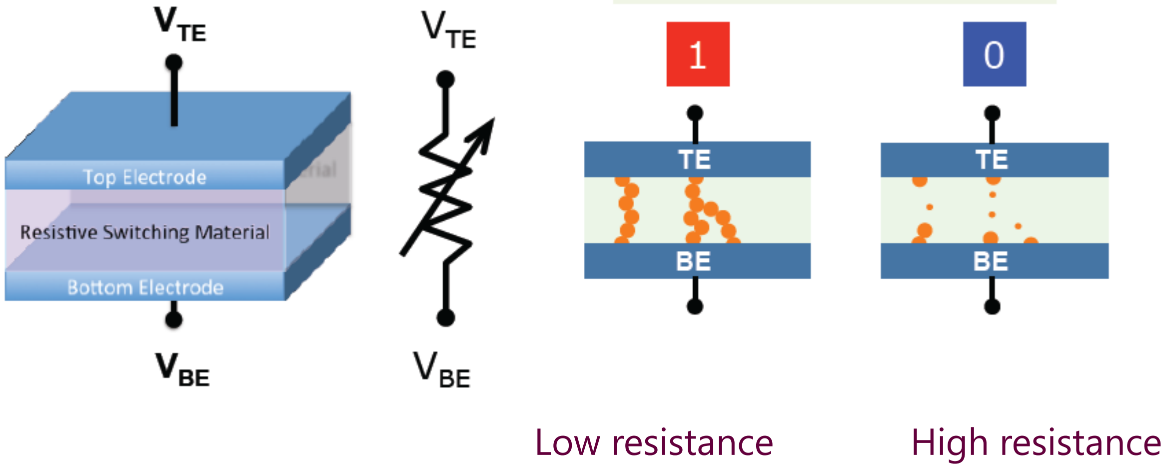






- ➢ The Promise of Analog In-Memory Computing
- ➢ Analog In-Memory Computing in a System
- ➢ Non-Memristor Analog In-Memory Computing
- ➢ Creative In-Memory Computing

# 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: Microcontroler Applications (NOR Flash replacement)

Infineon

All Search Newsletter Contact Wi

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

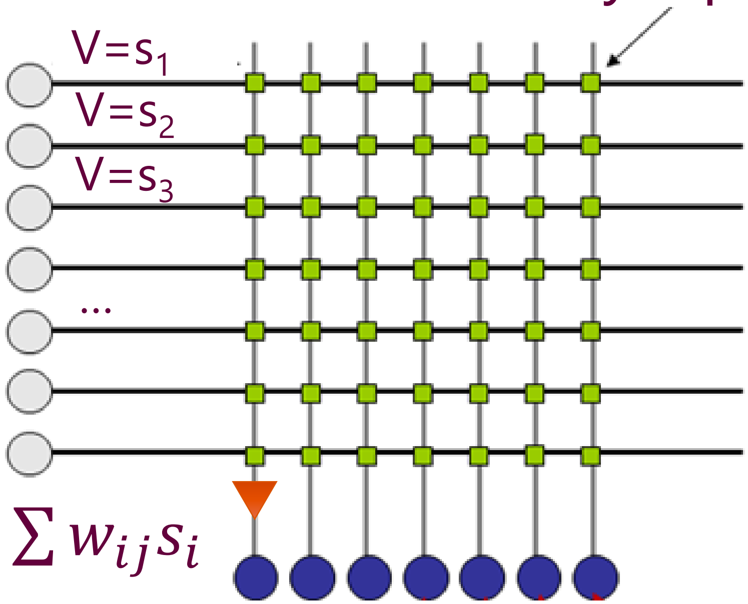
# Memristors Are Very Different From SRAM/DRAM

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

# Analog In-Memory Computing Performs Neural Network Inference Very Naturally

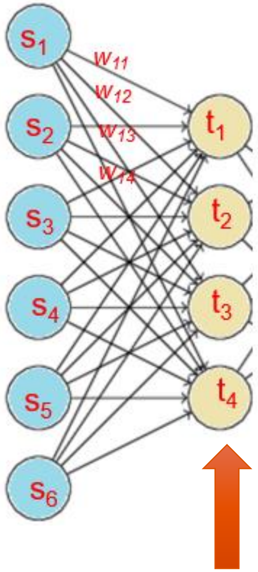
A matrix of analog memristors naturally implements a layer of neural network with **Ohm's and Kirchhoff's laws**!

Memristor conductance G = synaptic weight w




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

$\equiv$




Neuron value

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

# In-Principle Energy Efficiency Is Astonishing

- Multiplication $E = U \times i \times t = U^2 \times t \,/\, R$

0.1 V 1 ns 100 kΩ

0.1 fJ

- Accumulation $E=0$

**0.05 fJ / operation**

# How Do We Measure Efficiency?

Currently, everybody is using

**TOPS/W**

= Tera Operations / Second / Watt

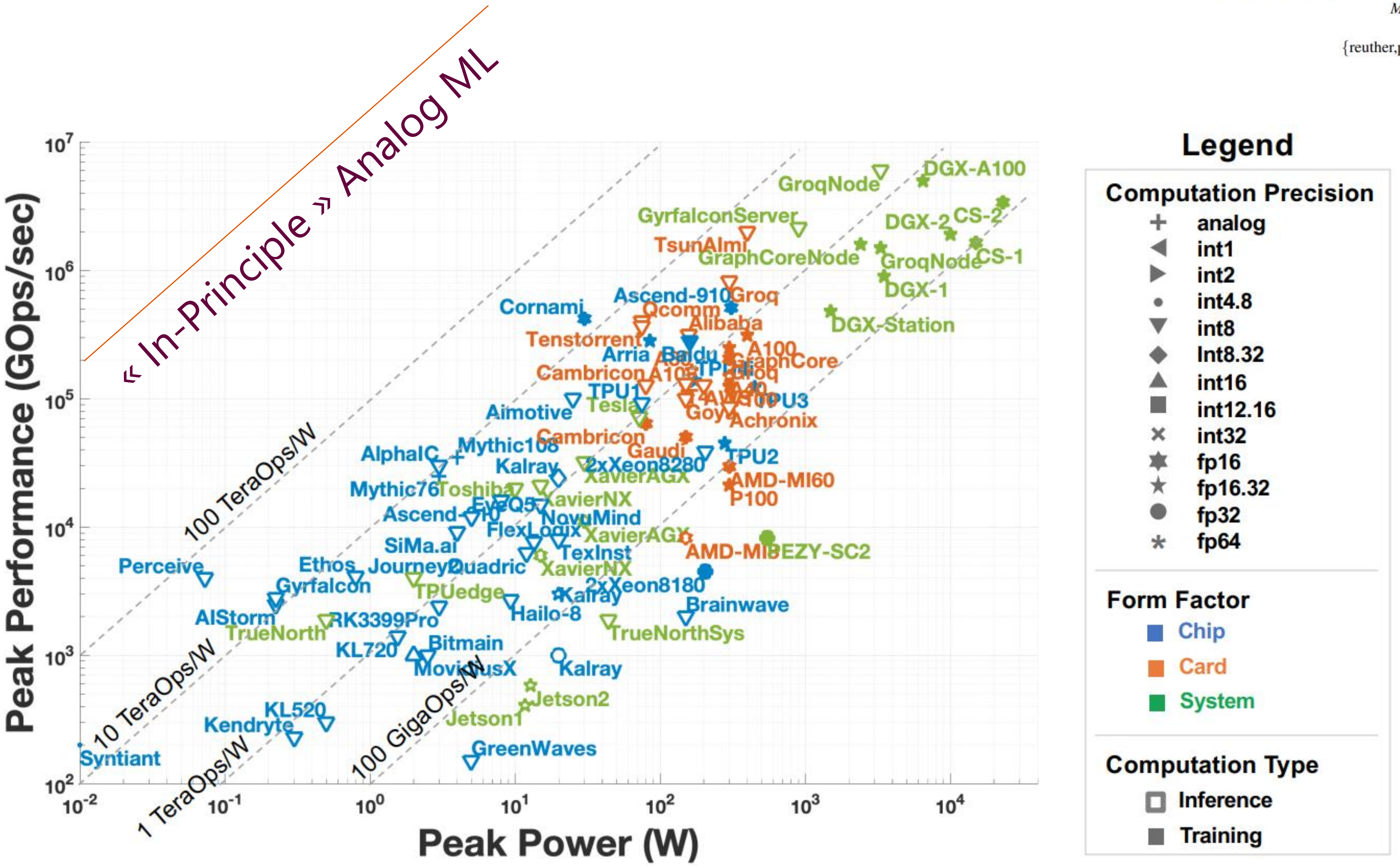
= Tera Operations / Joule

In principle, analog IMC could be 20,000 TOPS/W

# How Do We Measure Efficiency?

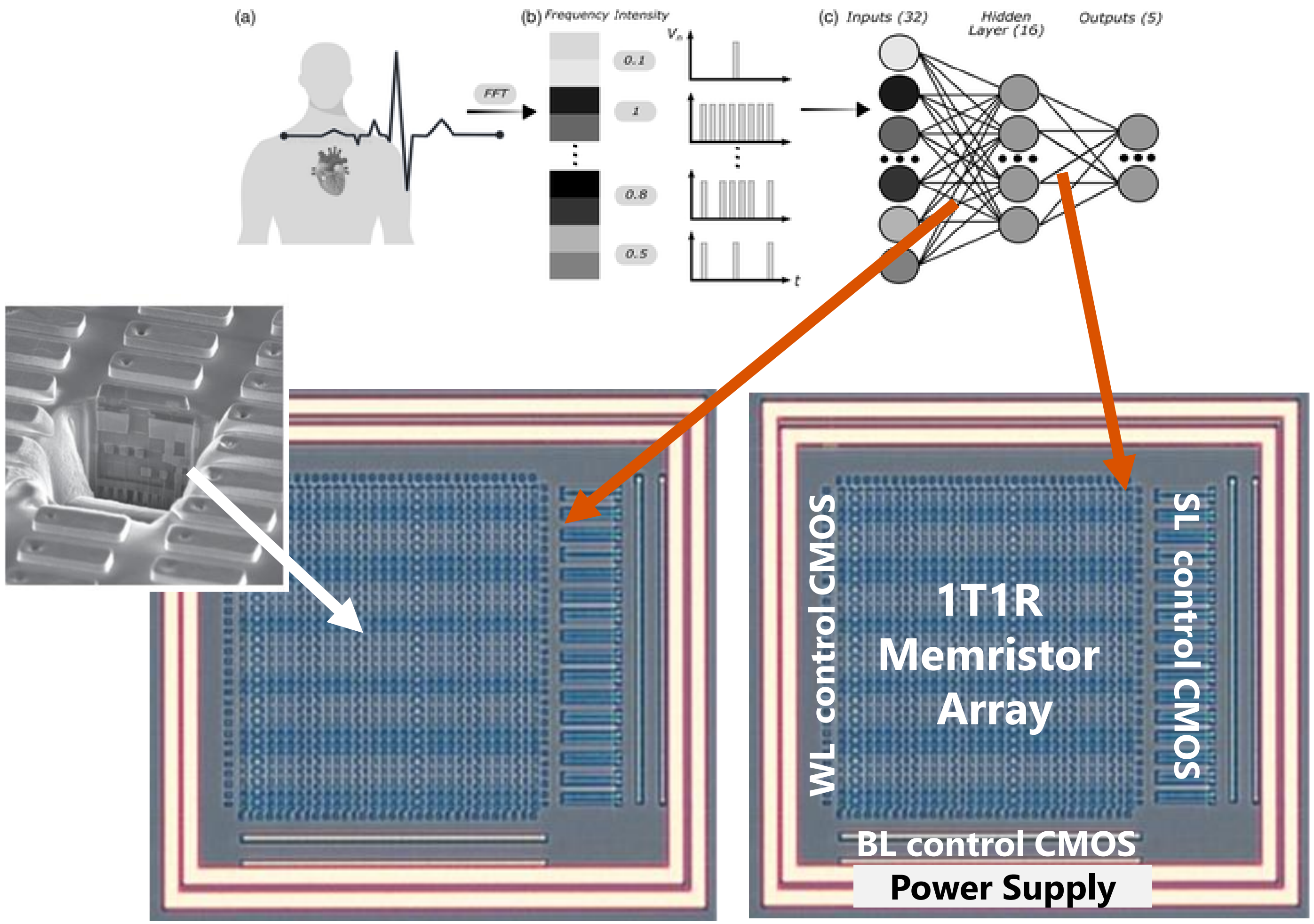
AI Accelerator Survey and Trends

Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner
*MIT Lincoln Laboratory Supercomputing Center*
Lexington, MA, USA
{reuther,pmichaleas,michael.jones,vijayg,sid,kepner}@ll.mit.edu



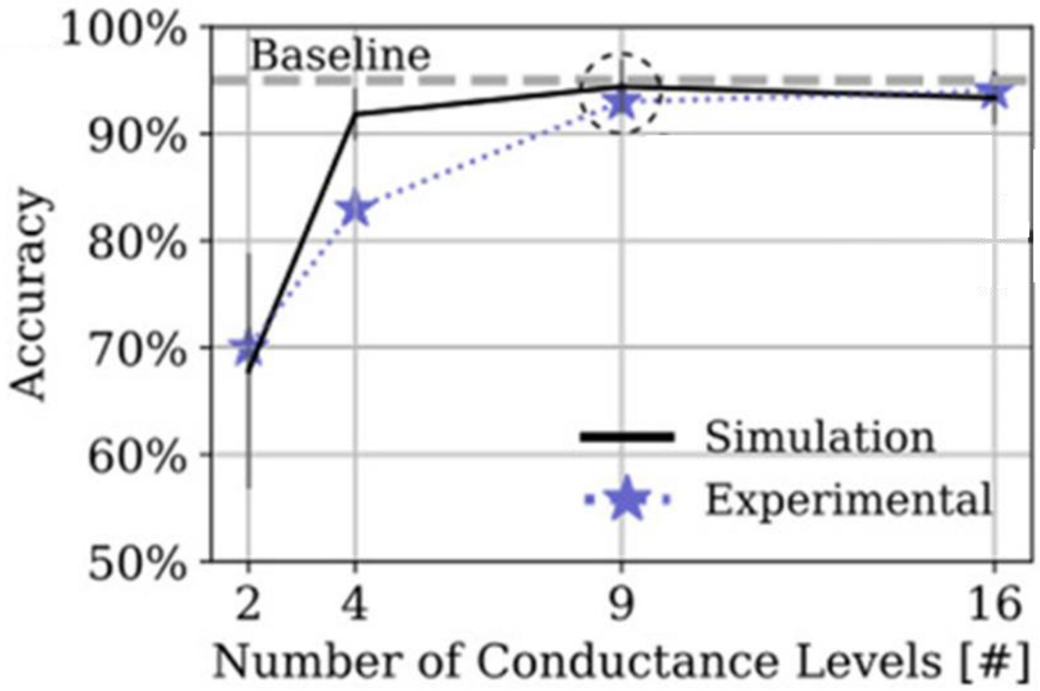

(2023 Data)

# Analog In-Memory Computing with Memristors




HfO$_x$ memristor integrated in BEOL of 130 nm-CMOS

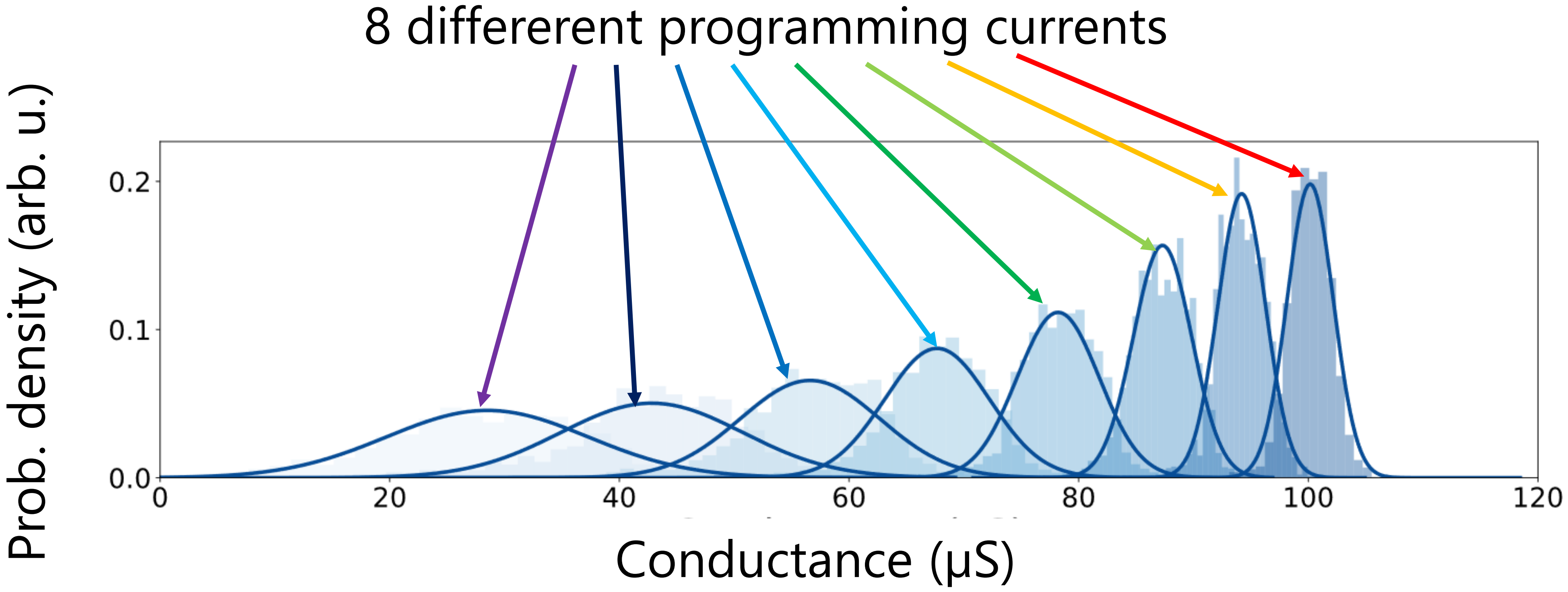
Accuracy on arrhythmia identification




Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

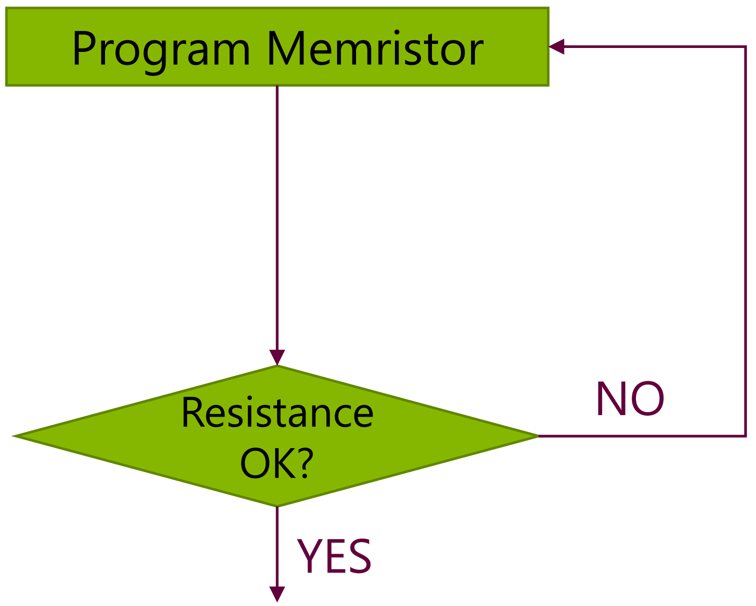
**Collaboration**

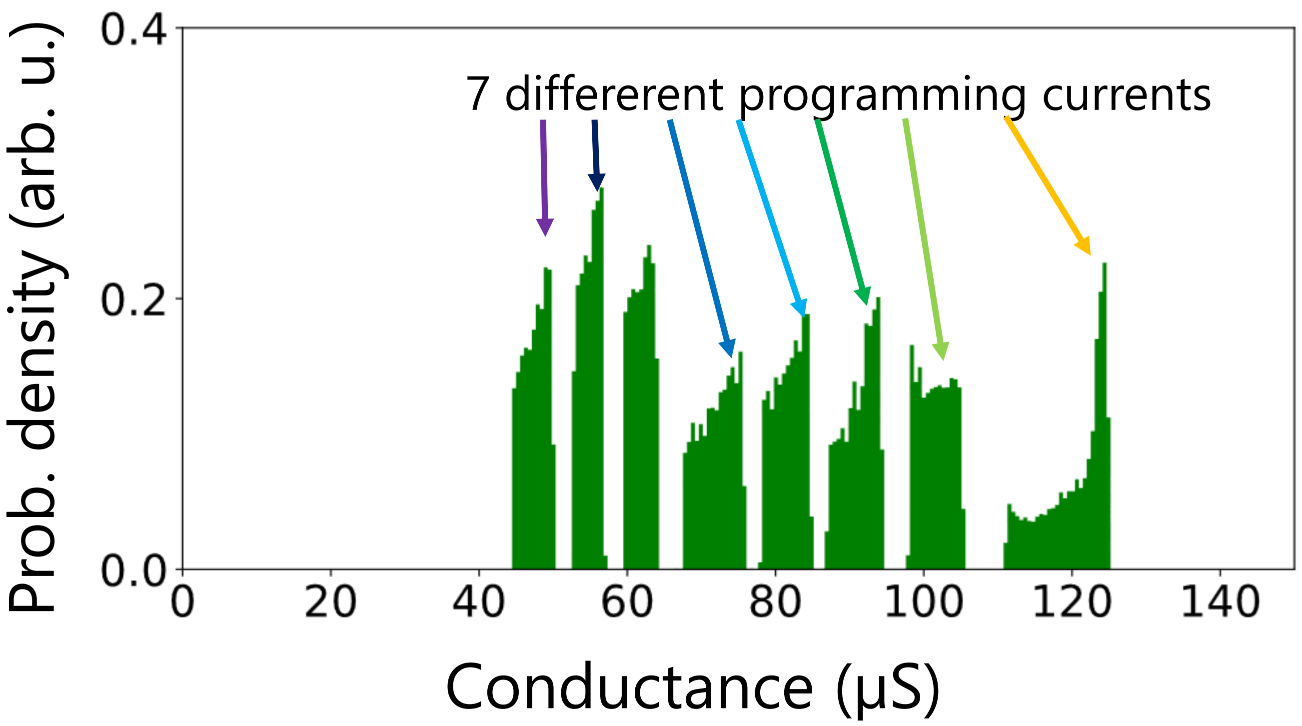
# The Challenge of Analog In-Memory Computing: Memristors Act as Random Variables




Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

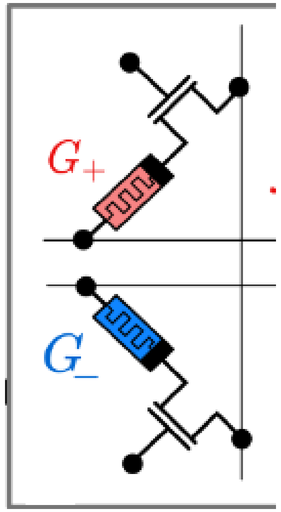
# Fighting the Random Character of Memristors: *Program and Verify*




Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

# Most Analog IMC Uses Differential Schemes




- Weight is difference between two memristor conductance

- ~3b / Memristor ->4b weight

- **Sufficient for almost all neural networks**

# Memristors Are For
# ***Weight Stationary*** Hardware

- Analog programming operations are long

- Write endurance is limited (e.g., one million)

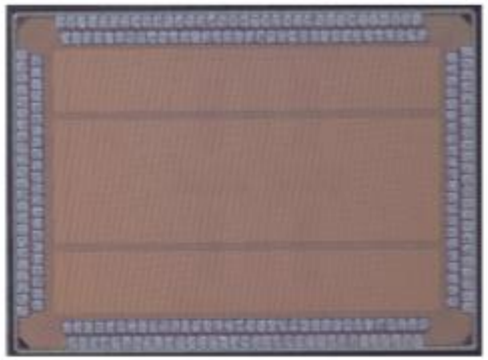
# Scaling Up the Concept

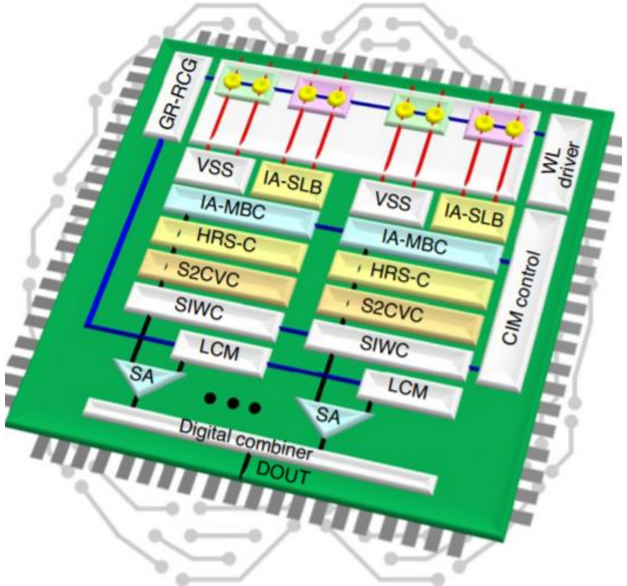

nature electronics
ARTICLES
https://doi.org/10.1038/s41928-020-00505-5
Check for updates

**A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices**

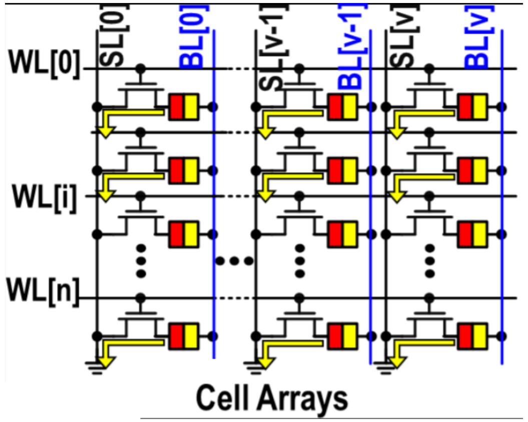
Cheng-Xin Xue[1,4], Yen-Cheng Chiu[1,4], Ta-Wei Liu[1], Tsung-Yuan Huang[1], Je-Syu Liu[1], Ting-Wei Chang[1], Hui-Yao Kao[1], Jing-Hong Wang[1], Shih-Ying Wei[1], Chun-Ying Lee[1], Sheng-Po Huang[1], Je-Min Hung[1], Shih-Hsih Teng[1], Wei-Chen Wei[1], Yi-Ren Chen[1], Tzu-Hsiang Hsu[1], Yen-Kai Chen[1], Yun-Chen Lo[1], Tai-Hsing Wen[1], Chung-Chuan Lo[1], Ren-Shuo Liu[1], Chih-Cheng Hsieh[1], Kea-Tiong Tang[1], Mon-Shu Ho[2], Chin-Yi Su[3], Chung-Cheng Chou[3], Yu-Der Chih[3] and Meng-Fan Chang[1]

2Mb compute-in memory array
(TSMC technology)






(Taiwan)




**2M Resistive Rams, 1T1R**
**<u>Used as binary memory</u>**
**<u>Very complex periphery</u>**
**<u>These periphery circuits require precise calibration</u>**

**146 TOPS/W (22nm CMOS)**

*Xue et al, Nature Electronics,* **4**, 81 (2021)

# A Realization « Truer » to the Original Concept

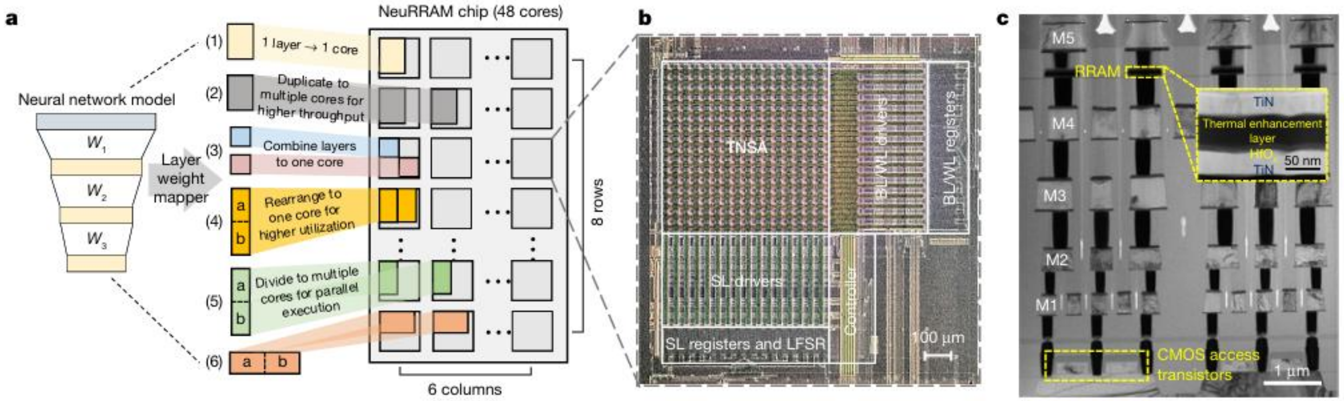
Article

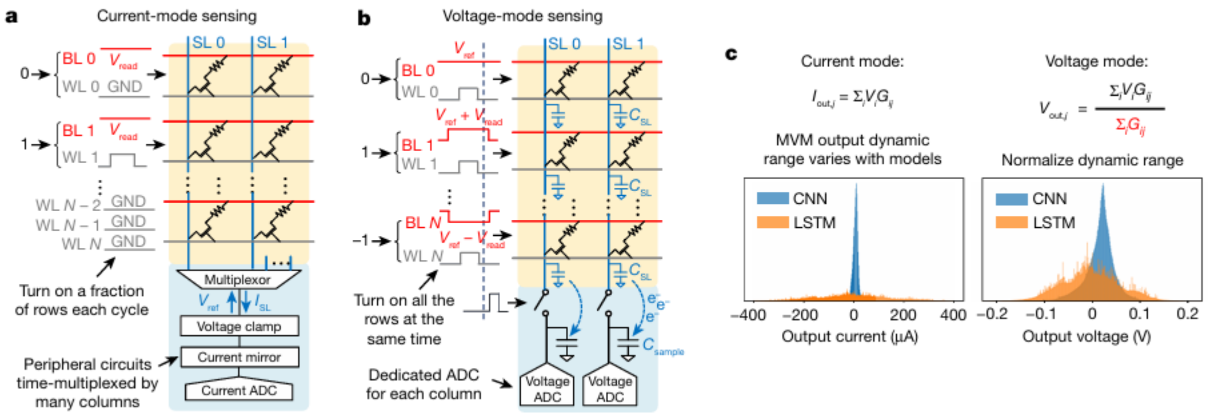
## A compute-in-memory chip based on resistive random-access memory

https://doi.org/10.1038/s41586-022-04992-8
Received: 27 July 2021
Accepted: 17 June 2022

Weier Wan[1,2], Rajkumar Kubendran[2,3], Clemens Schaefer[4], Sukru Burc Eryilmaz[1], Wenqiang Zhang[5], Dabin Wu[5], Stephen Deiss[2], Priyanka Raina[1], He Qian[5], Bin Gao[5], Siddharth Joshi[2,4], Huaqiang Wu[5], H.-S. Philip Wong[1] & Gert Cauwenberghs[2]

*Nature* (2022)




**10-45 TOPS/W in 130nm CMOS**

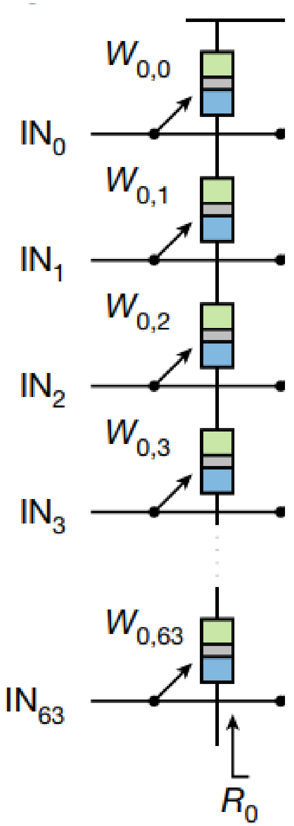
# The Samsung Approach

Article

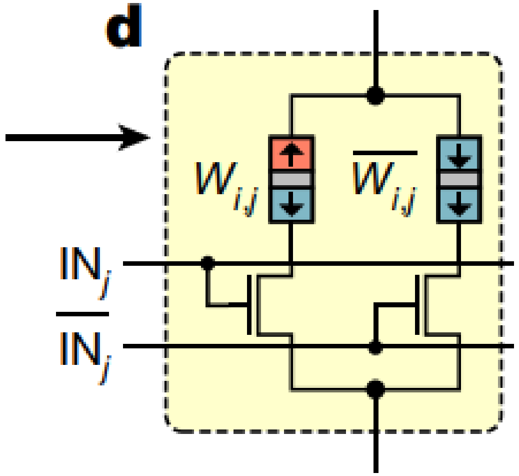
**A crossbar array of magnetoresistive memory devices for in-memory computing**

Nature 2022

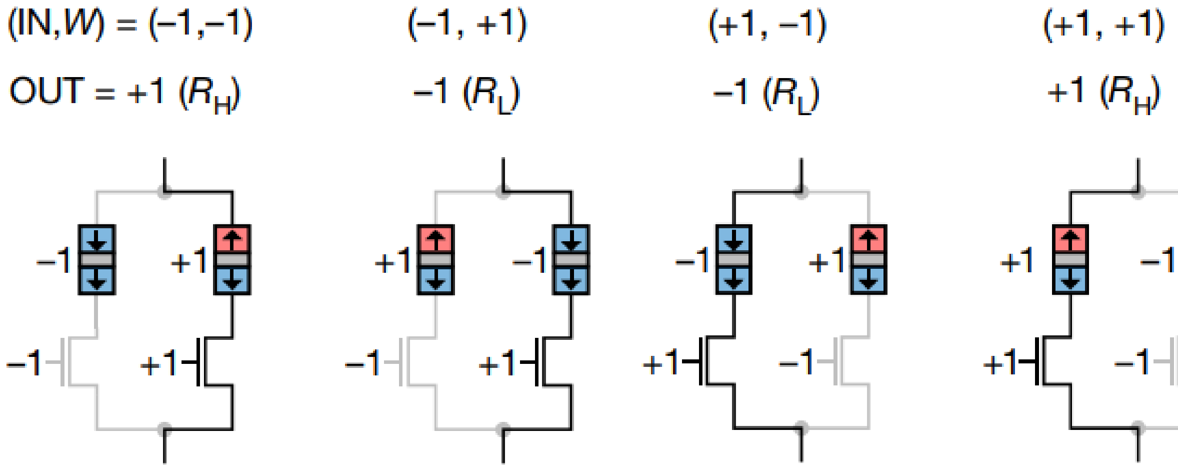
- Implements Binarized Neural Networks
- MRAM
- Uses Resistance (and not Conductance) as a synaptic weight



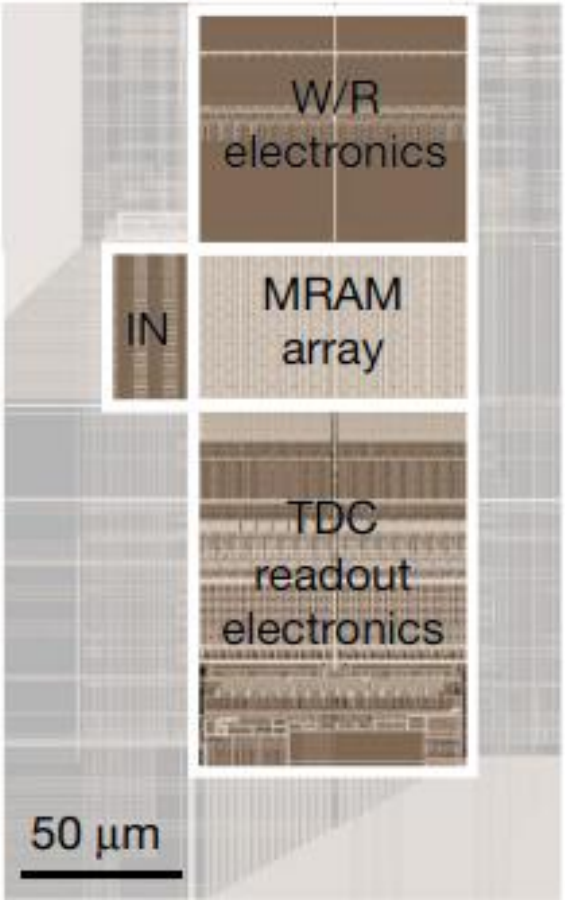

**d**




**e**



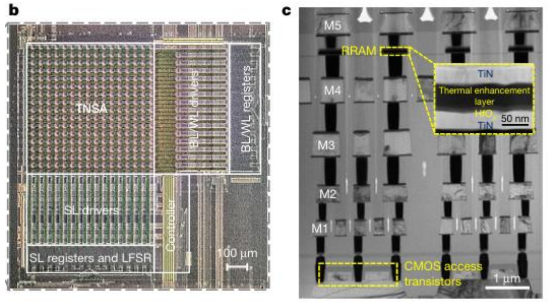
# The Samsung Approach Requires Complex Periphery Circuitry

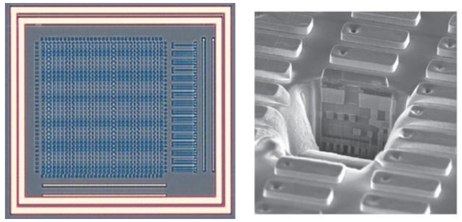


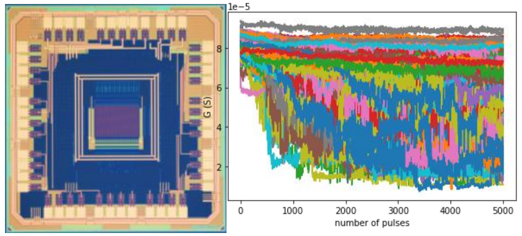

- Need for calibration, compensation, and a quite complex analog to digital scheme using time (TDC)

**405 TOPS/W in 28 nm CMOS**

# Analog Machine Learning Accelerators






- ➢ The Promise of Analog In-Memory Computing
- ➢ Analog In-Memory Computing in a System
- ➢ Non-Memristor Analog In-Memory Computing
- ➢ Creative In-Memory Computing

# How to Go From an Array to a Full Neural Network

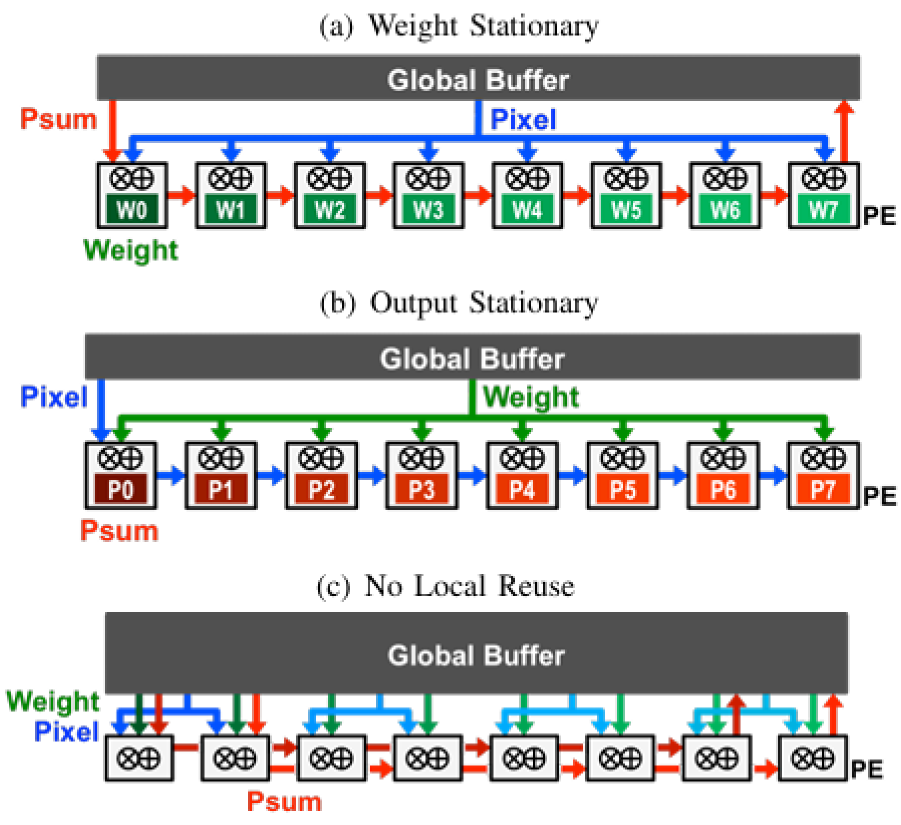


Fig. 8. Dataflows for DNNs.

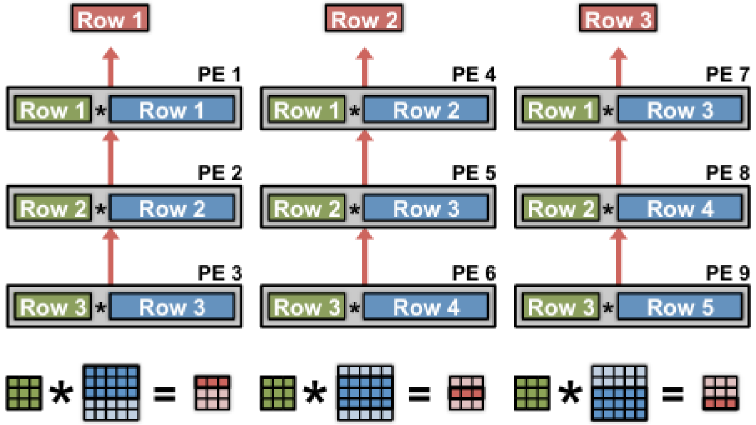
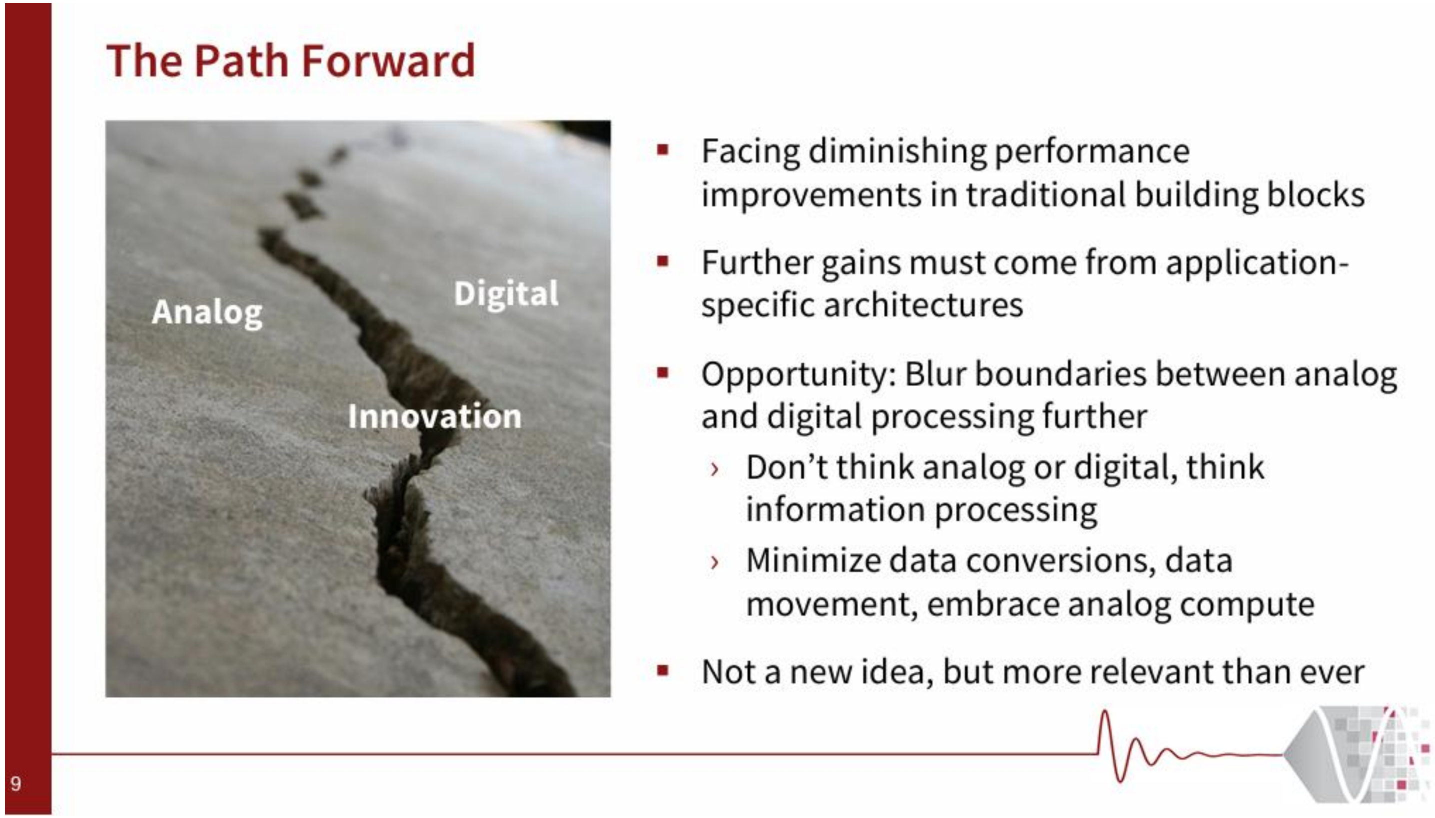


Fig. 9. Row Stationary Dataflow [34].

In digital ML accelerators, weight stationary is not the most researched option.

Sze et al, Proc. IEEE (2017)

# Where to Set the Boundary Between Analog and Digital?




Boris Murmann, BIOCAS, 2019

- **My word of caution**: many computer arch. overestimate the cost of ADC. Use ADCs specifically designed for IMC (e.g., CCO)

# A Full System with Mostly Analog Routing

Article

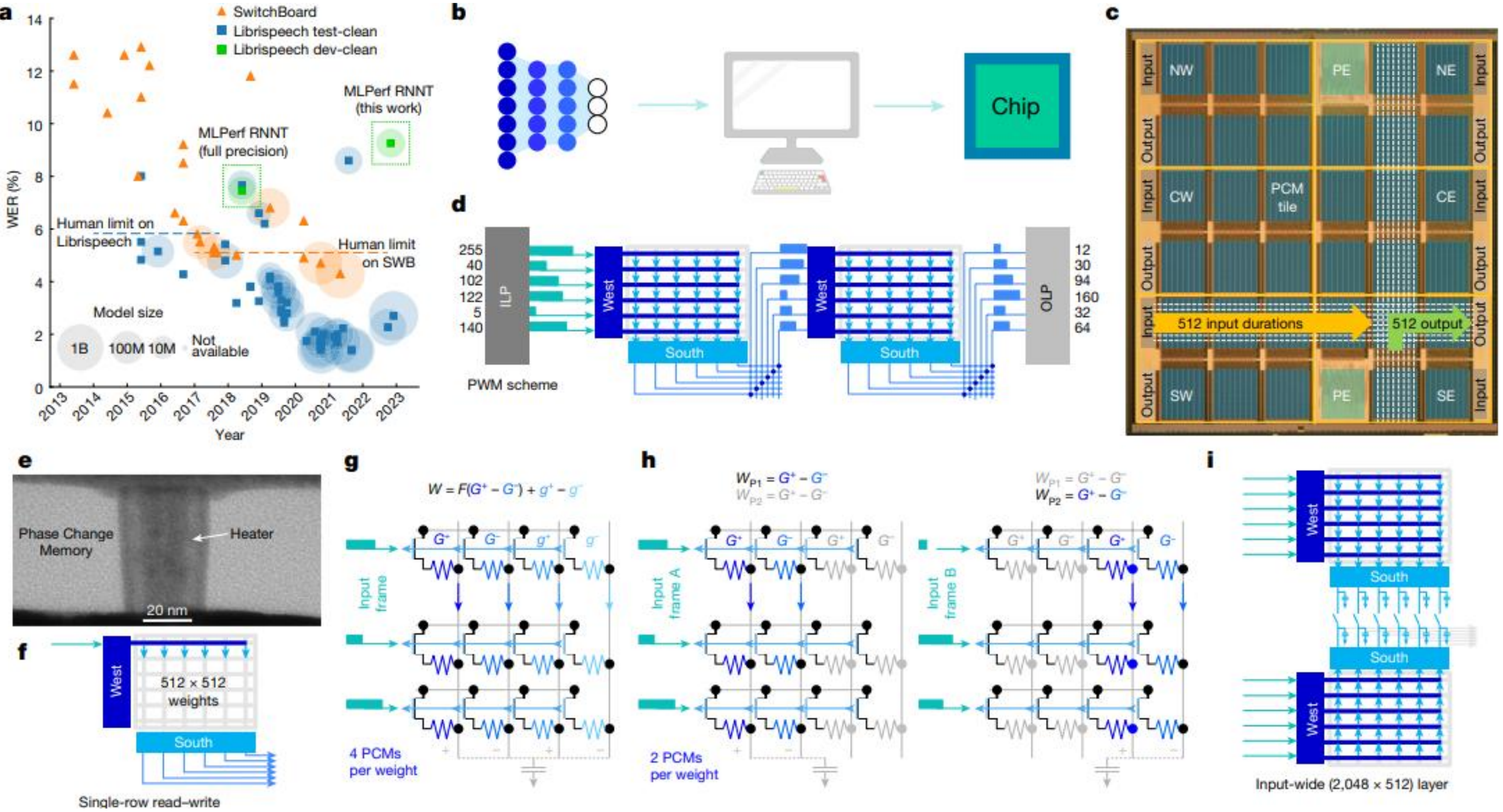
An analog-AI chip for energy-efficient speech recognition and transcription

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

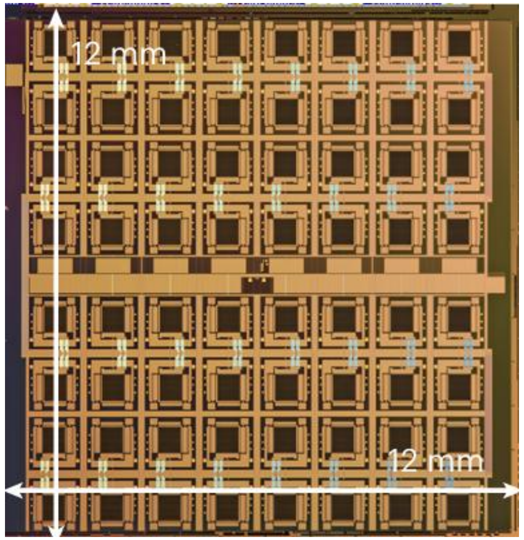


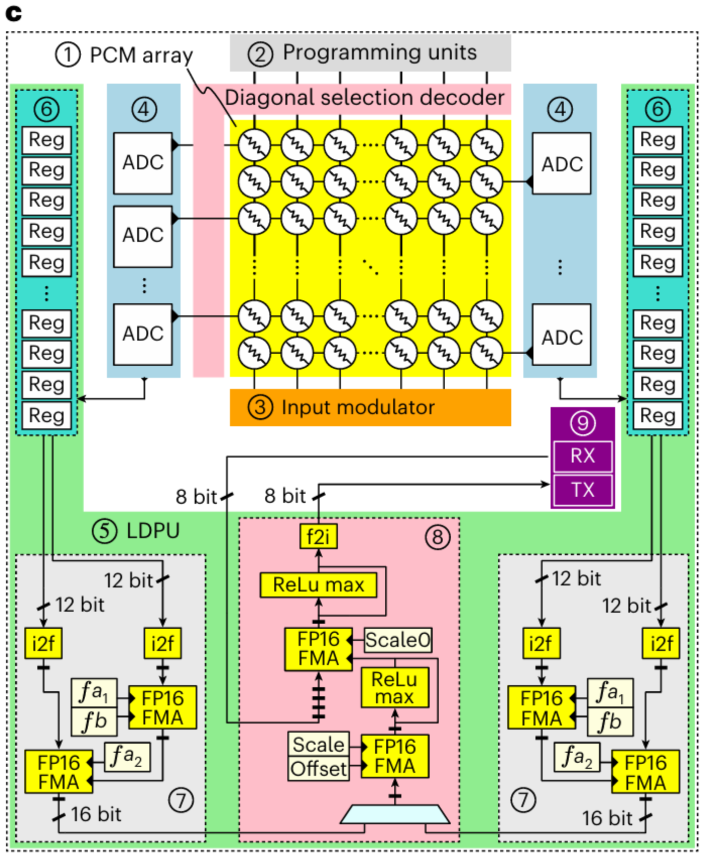

**Full system**
**Mostly analog routing**

**12 TOPS/W**
**(14 nm CMOS)**

- 35M phase change memories

# A Full System with Mostly Digital Routing





nature electronics

Article

https://doi.org/10.1038/s41928-023-01010-1


## A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference


Received: 27 May 2023

Accepted: 10 July 2023

Published online: 10 August 2023

Check for updates



Manuel Le Gallo[1,6], Riduan Khaddam-Aljameh[1,6], Milos Stanisavljevic[1,6], Athanasios Vasilopoulos[1], Benedikt Kersting[1], Martino Dazzi[1], Geethan Karunaratne[1], Matthias Brändli[1], Abhairaj Singh[1], Silvia M. Müller[2], Julian Büchel[1], Xavier Timoneda[1], Vinay Joshi[1], Malte J. Rasch[3], Urs Egger[1], Angelo Garofalo[1], Anastasios Petropoulos[4], Theodore Antonakopoulos[4], Kevin Brew[5], Samuel Choi[5], Injo Ok[5], Timothy Philip[5], Victor Chan[5], Claire Silvestre[5], Ishtiaq Ahsan[5], Nicole Saulnier[5], Vijay Narayanan[3], Pier Andrea Francese[1], Evangelos Eleftheriou[1] & Abu Sebastian[1]


Nature Electronics 2023

**Full system**

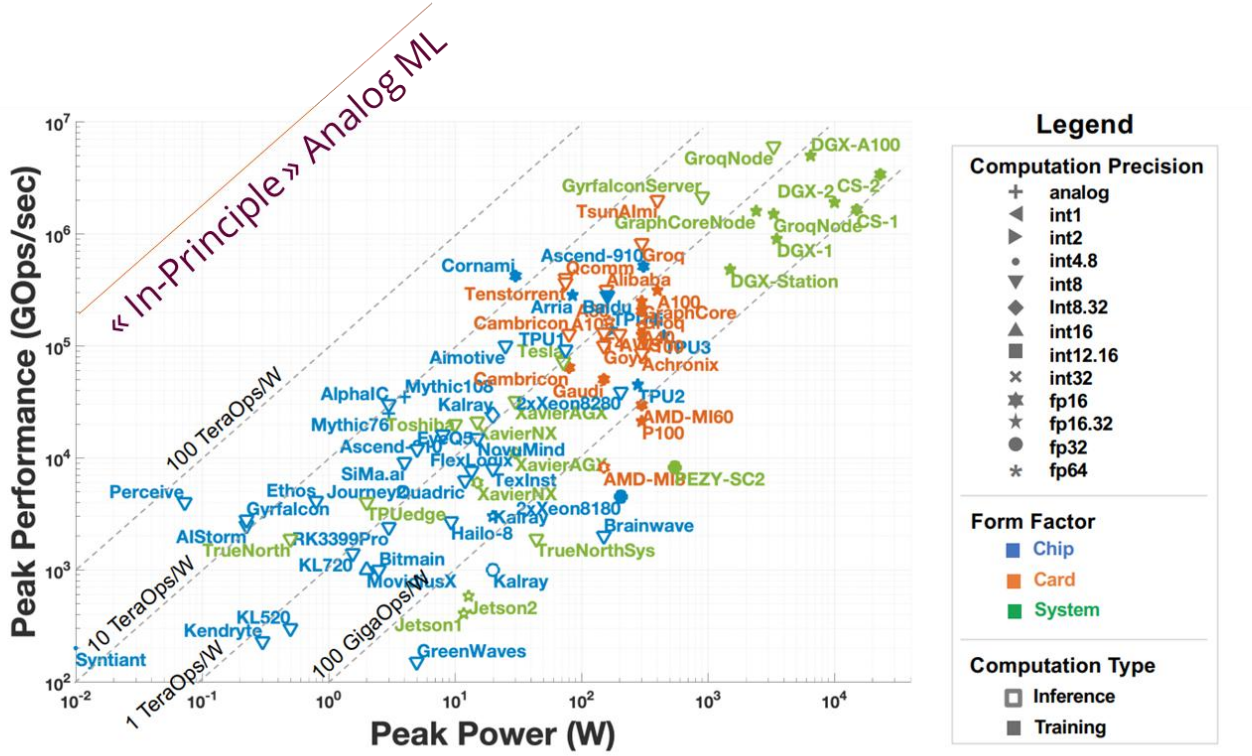
**Mostly digital routing**

**(massive ADC/DAC)**

**10 TOPS/W**

# « But the TOPS/W Are Not That High! »

- **These are early prototypes, optimized for functionality not for perf**



# TOPS/W Can Be a Very Misleading Unit

- « Operation » is ill-defined
- These are for peak conditions, which are rarely reached in practice
- **The difference between real-life and peak conditions varies extensively depending on the different types of hardware**

EETimes

HOME NEWS PERSPECTIVES DESIGNLINES PODCAST EDUCATION STORE

DESIGNLINES | AI & BIG DATA DESIGNLINE

TOPS: The Truth Behind a Deep Learning Lie

By Ludovic Larzul, Mipsology 06.25.2021 0

Jan Werth

Apr 26, 2021 · 11 min read · Listen

**When "TOPS" are Misleading**

Neural accelerators are often characterized with the performance feature "TOPS" — Trillion operations per second. But that alone is not enough. It is important to know how these accelerators work and what else should be considered when making a comparison.

AUTO, SECURITY & PERVASIVE COMPUTING OPINION

**Lies, Damn Lies, And TOPS/Watt**

249 Shares 20 7 20

*Questions you need to ask to make sure you understand AI hardware performance.*

JANUARY 7TH, 2019 - BY: GEOFF TATE

There are almost a dozen vendors promoting inferencing IP, but none of them

Aron Kirschen

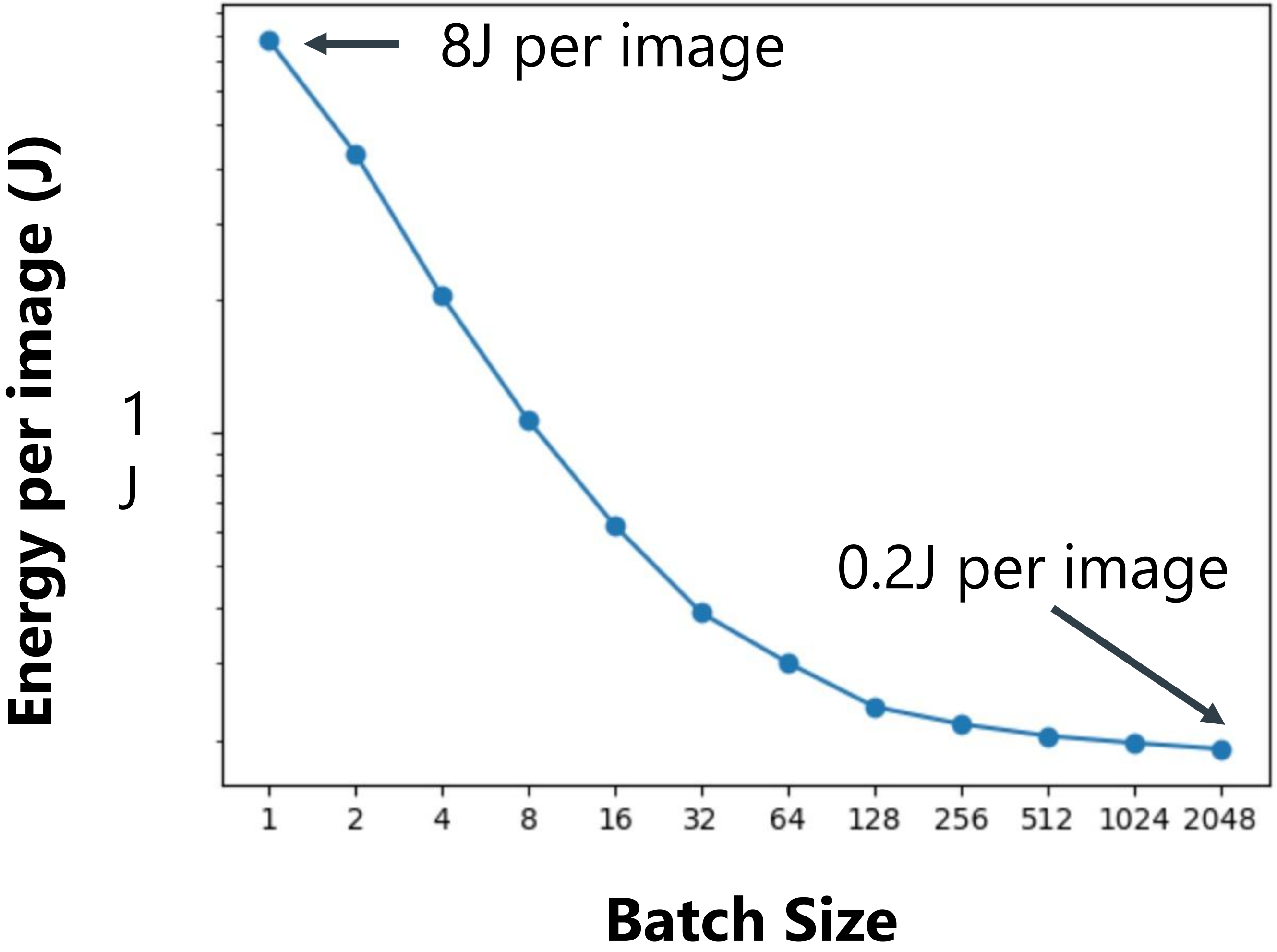
Aug 12, 2020 · 10 min read · Listen

**Why TOPS/W is a bad unit to benchmark next-gen AI chips**



cnrs université PARIS-SACLAY

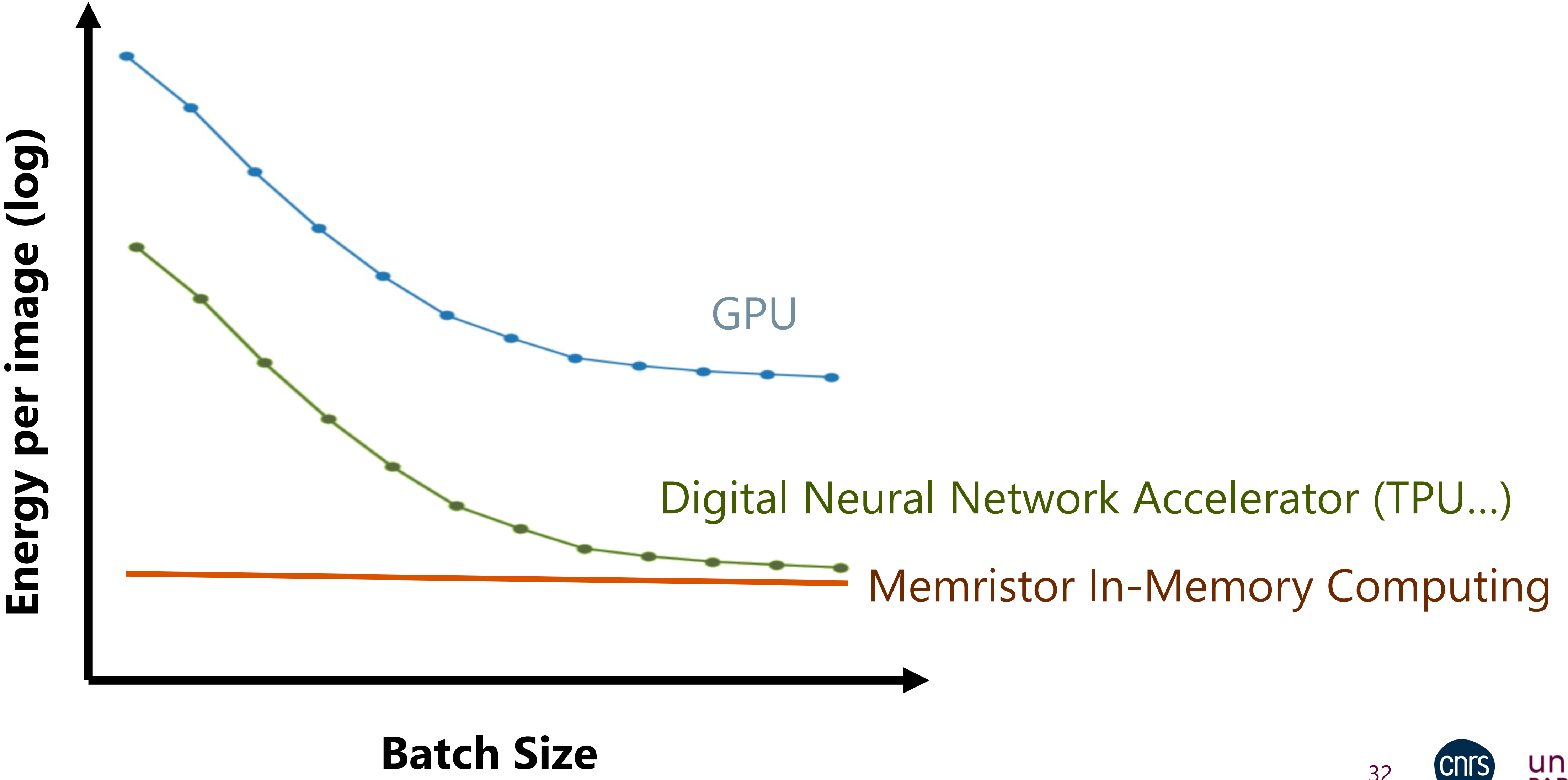
# Real Energy Consumption Can Be Dramatically Higher than Peak



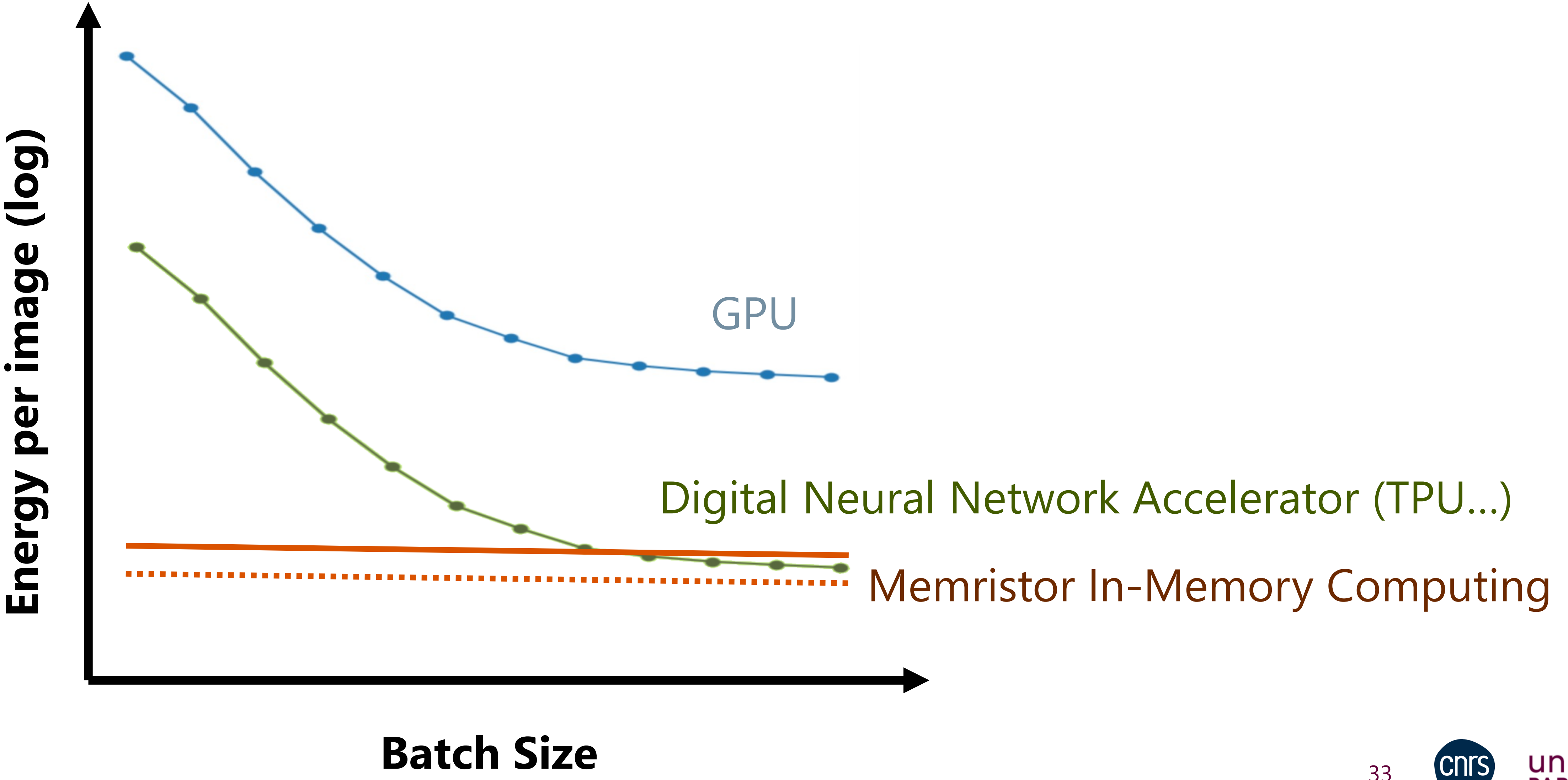

NVIDIA A100 GPU
ResNet50

*Courtesy of David Novo (Univ Montpellier/CNRS)*

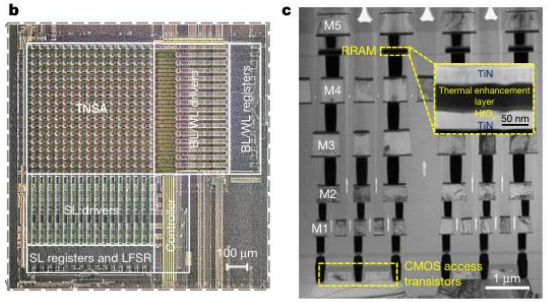
# How I View Architecture Research

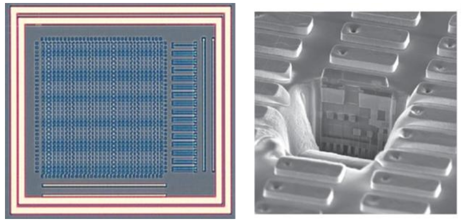


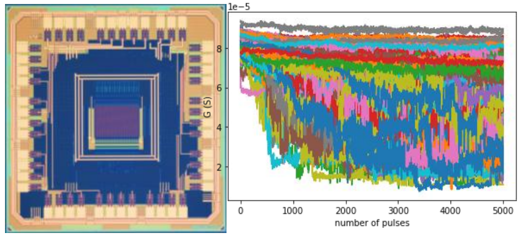
# How I View Architecture Research



# Analog Machine Learning Accelerators






- ➢ The Promise of Analog In-Memory Computing
- ➢ Analog In-Memory Computing in a System
- ➢ Non-Memristor Analog In-Memory Computing
- ➢ Creative In-Memory Computing

# From Unconventional to More Conventional?

- For cultural reason, many analog IMC ideas are developed for emerging technology

- There's a movement to adapt these ideas to more established technology

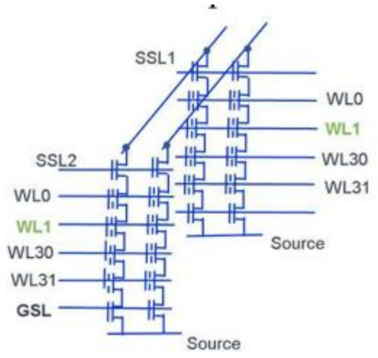
# IMC with Flash Memory

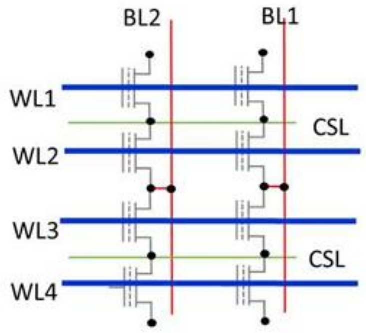
**Prospects of Computing In or Near Flash Memories**

Hang-Ting Lue, Chun-Hsiung Hung, Keh-Chung Wang and Chih-Yuan Lu

Macronix International Co., Ltd
16 Li-Hsin Road, Hsinchu Science Park, Hsinchu, Taiwan.
E-mail: htlue@mxic.com.tw

IEDM 2024




**NAND: Mainstream data storage (SSD)**
- Page (16KB*N) read/write (slower, 10-100us)
- Big block (> 4MB) erase (GC, write amplification)
- Rely on controller managements. Intensive ECC
- 3D charge-trapping memories (>300 layers), with MLC/TLC/QLC

**NOR: classic legacy for code Flash**
- Random access read (faster, ~100ns)
- Byte addressable; Small-unit P/E
- No controller needed. Almost no ECC
- 2D FG stops at 45nm node → *Going to 3D charge-trapping NOR*

- NAND flash constrains are challenging for IMC
- NOR Flash memory can be used for IMC,
  but does not scale to advanced CMOS

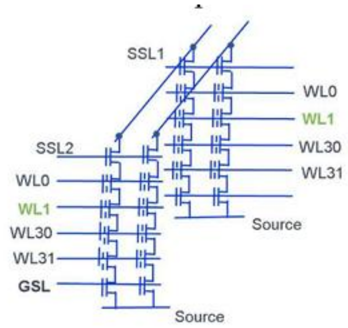
# Toward 3D IMC?

## Prospects of Computing In or Near Flash Memories

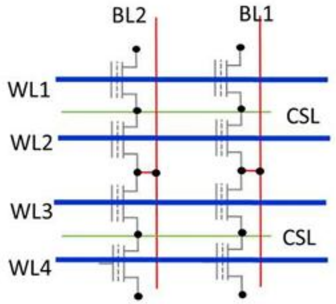
Hang-Ting Lue, Chun-Hsiung Hung, Keh-Chung Wang and Chih-Yuan Lu

Macronix International Co., Ltd
16 Li-Hsin Road, Hsinchu Science Park, Hsinchu, Taiwan.
E-mail: htlue@mxic.com.tw

IEDM 2024




**NAND: Mainstream data storage (SSD)**
- Page (16KB*N) read/write (slower, 10-100us)
- Big block (> 4MB) erase (GC, write amplification)
- Rely on controller managements. Intensive ECC
- 3D charge-trapping memories (>300 layers), with MLC/TLC/QLC




**NOR: classic legacy for code Flash**
- Random access read (faster, ~100ns)
- Byte addressable; Small-unit P/E
- No controller needed. Almost no ECC
- 2D FG stops at 45nm node → *Going to 3D charge-trapping NOR*

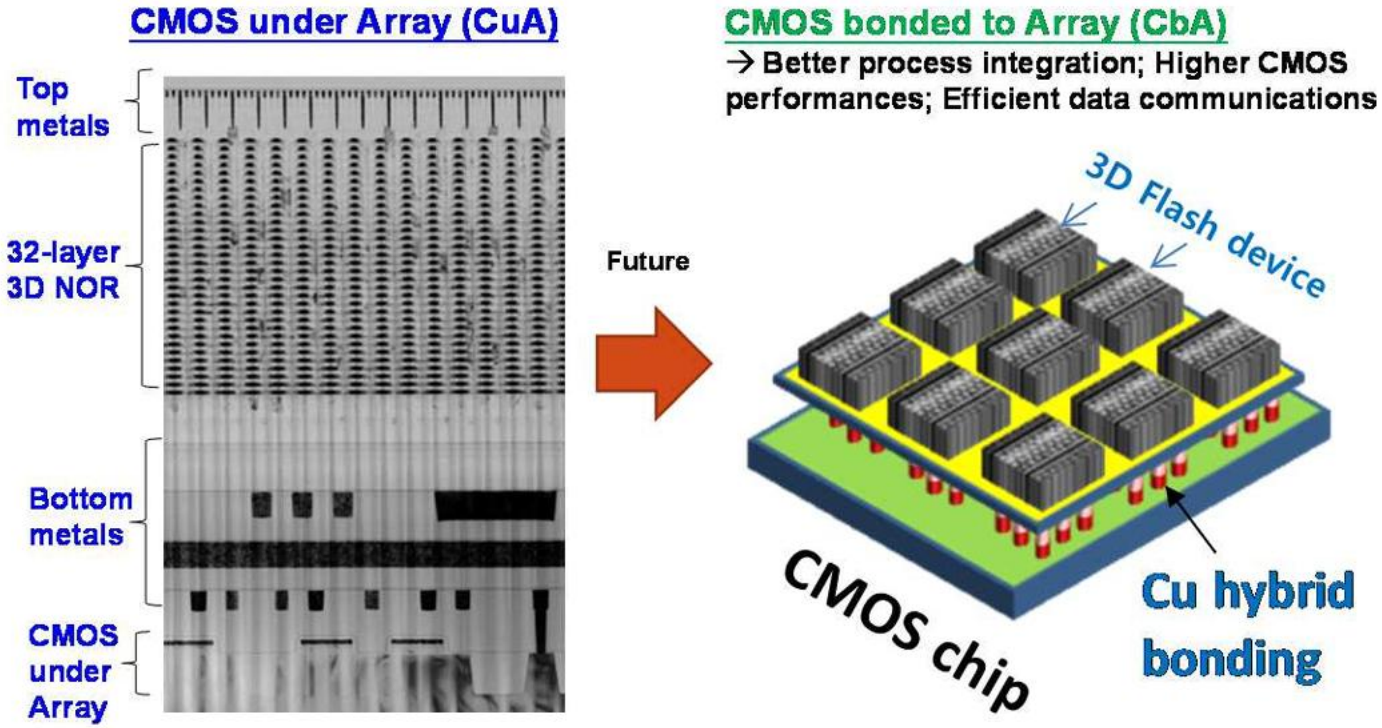
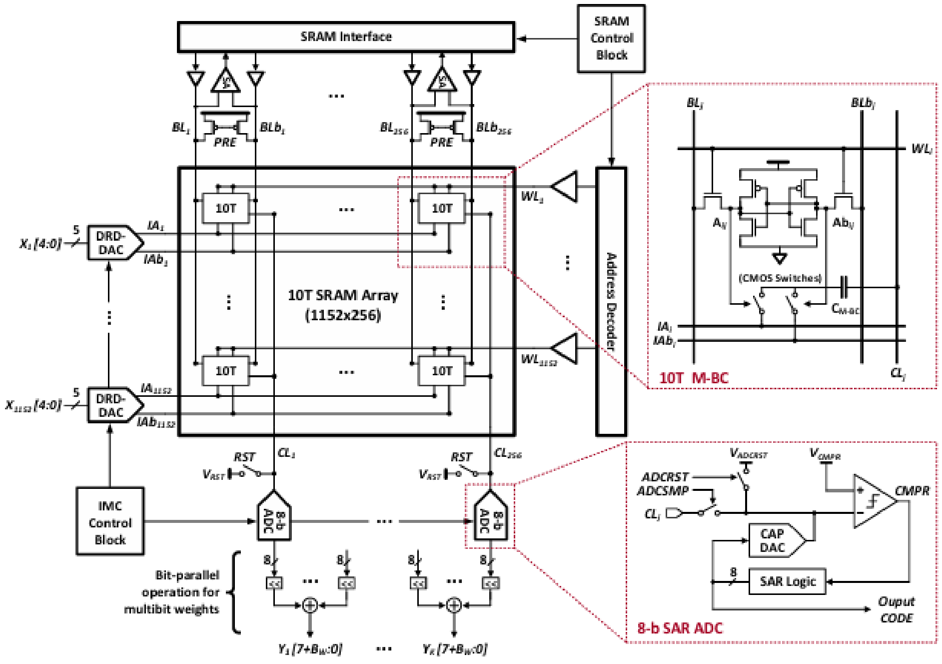



**Figure 4** Current 3D NOR adopts CMOS under Array (CuA) process integration. In the future, we will go for Cu hybrid bonding to connect separated CMOS chip with 3D array (like 3D NAND).

# SRAM-Based IMC

- SRAM can be used for both digital and analog IMC
- Most exciting analog designs rely on switched cap principles




**Fully Row/Column-Parallel In-memory Computing SRAM Macro employing Capacitor-based Mixed-signal Computation with 5-b Inputs**


Jinseok Lee, Hossein Valavi, Yinqi Tang and Naveen Verma
Princeton University, Princeton, NJ, USA (jinseokl@princeton.edu)


VLSI 2021

- Positioning: edge or HPC?

# DRAM-Based IMC

Memory-Centric Computing:
Recent Advances in Processing-in-DRAM (Invited)

Onur Mutlu, Ataberk Olgun, Geraldo F. Oliveira, Ismail E. Yuksel
ETH Zürich

IEDM 2024

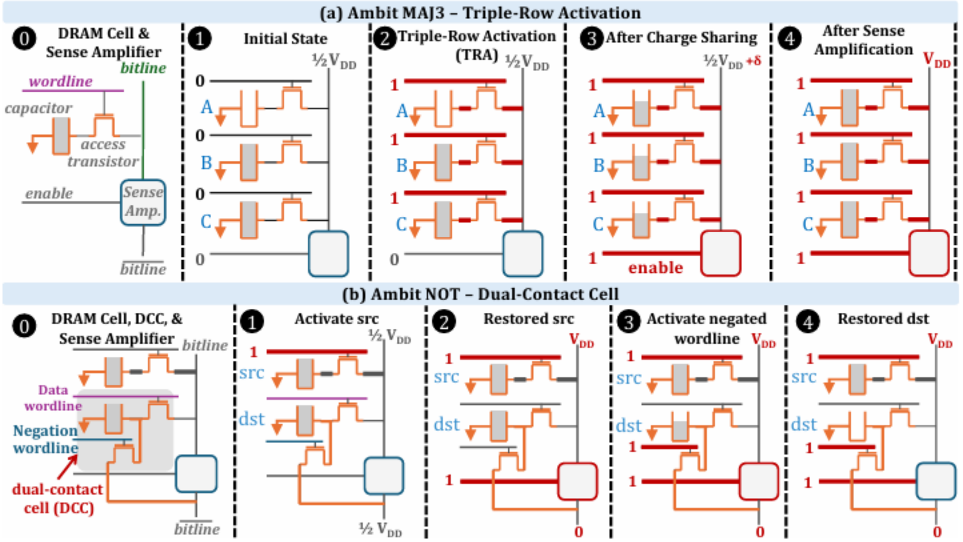
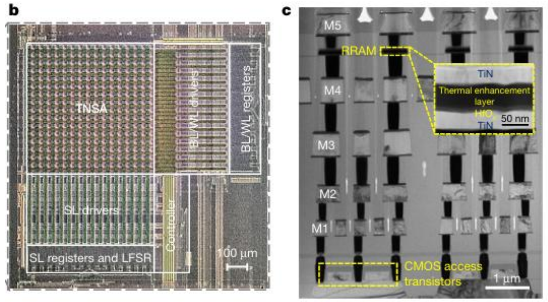
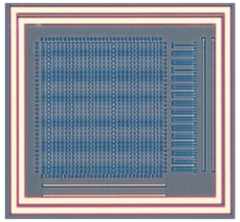


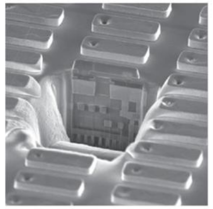

Fig. 1: **An example of performing the MAJority-of-three operation (i.e., MAJ3(A,B,C)) (a) and the NOT operation (i.e., dst=NOT(src)) in Ambit [111].** In (a), we focus on DRAM cell and sense amplifier operations (⓿). Initially, cells A, B, C, and bitline have voltage levels of GND, VDD, VDD, and VDD/2, respectively (❶). We first perform a triple-row activation (TRA) to simultaneously activate cells A, B, and C (❷). When the wordlines of all three cells are raised simultaneously, charge sharing

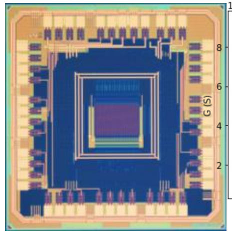
- DRAM based-IMC is possible but destructive read, and small signals -> most adapted for digital

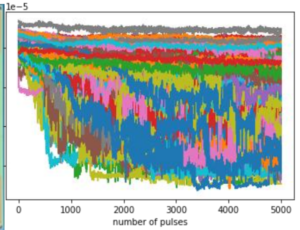
# Analog Machine Learning Accelerators







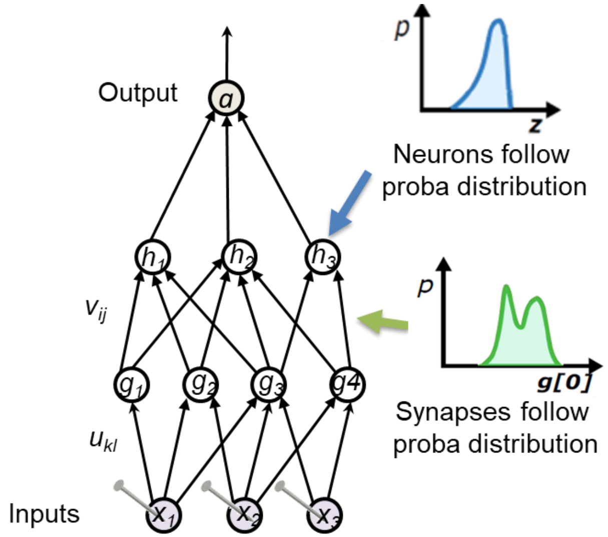


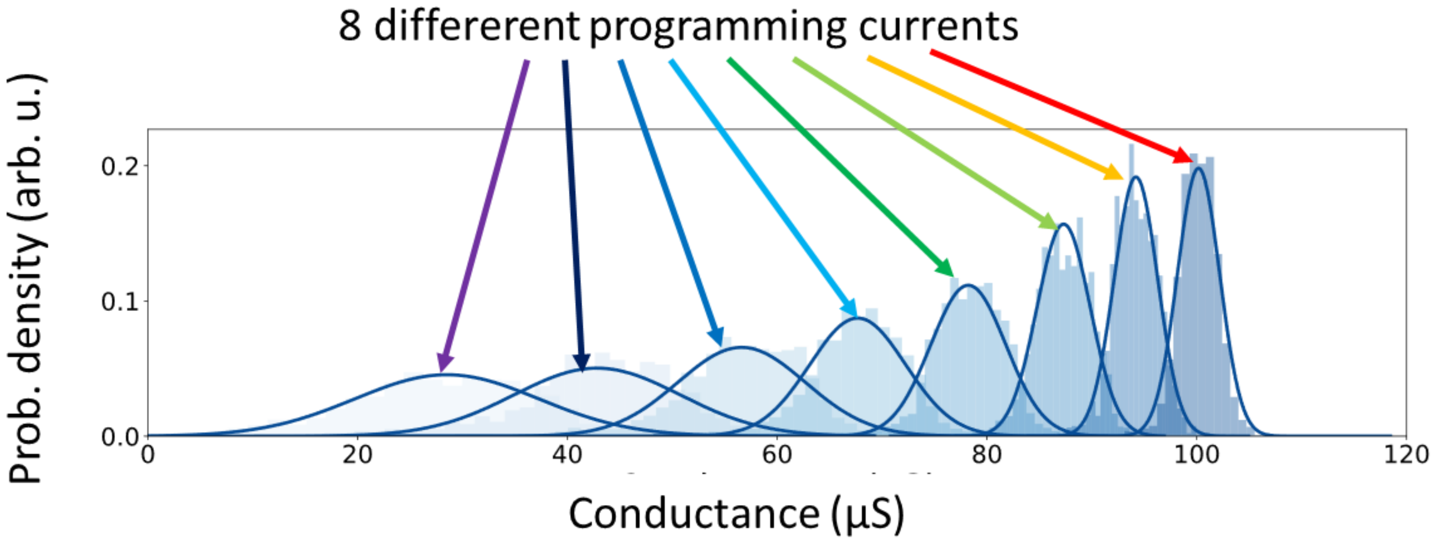

- The Promise of Analog In-Memory Computing
- Analog In-Memory Computing in a System
- Non-Memristor Analog In-Memory Computing
- Creative In-Memory Computing

# 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)

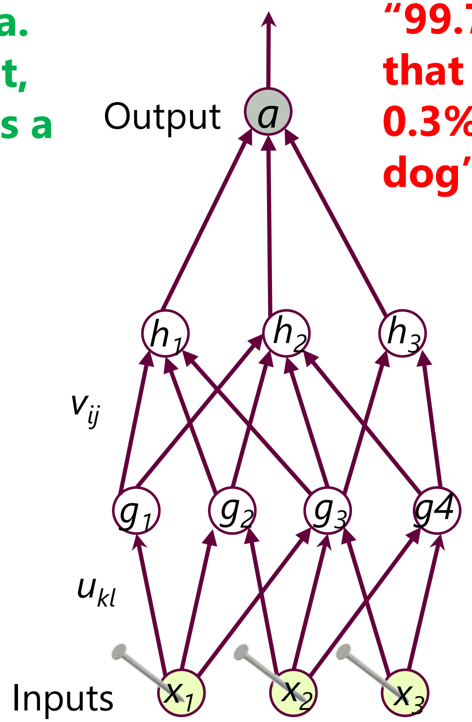
Collaboration

# Bayesian Neural Networks

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

### *Conventional Neural Network*

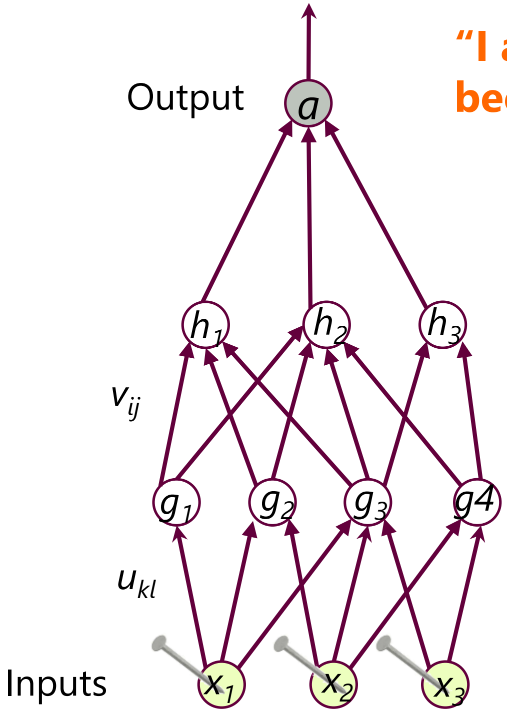
**"99.9% proba. that this a cat, 0.1% that it is a dog"**



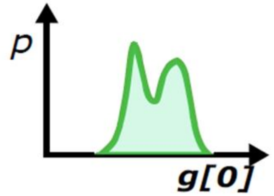

**"99.7% proba. that this a cat, 0.3% that it is a dog"**

0.7
Each synapse has one weight

### *Bayesian Neural Network*



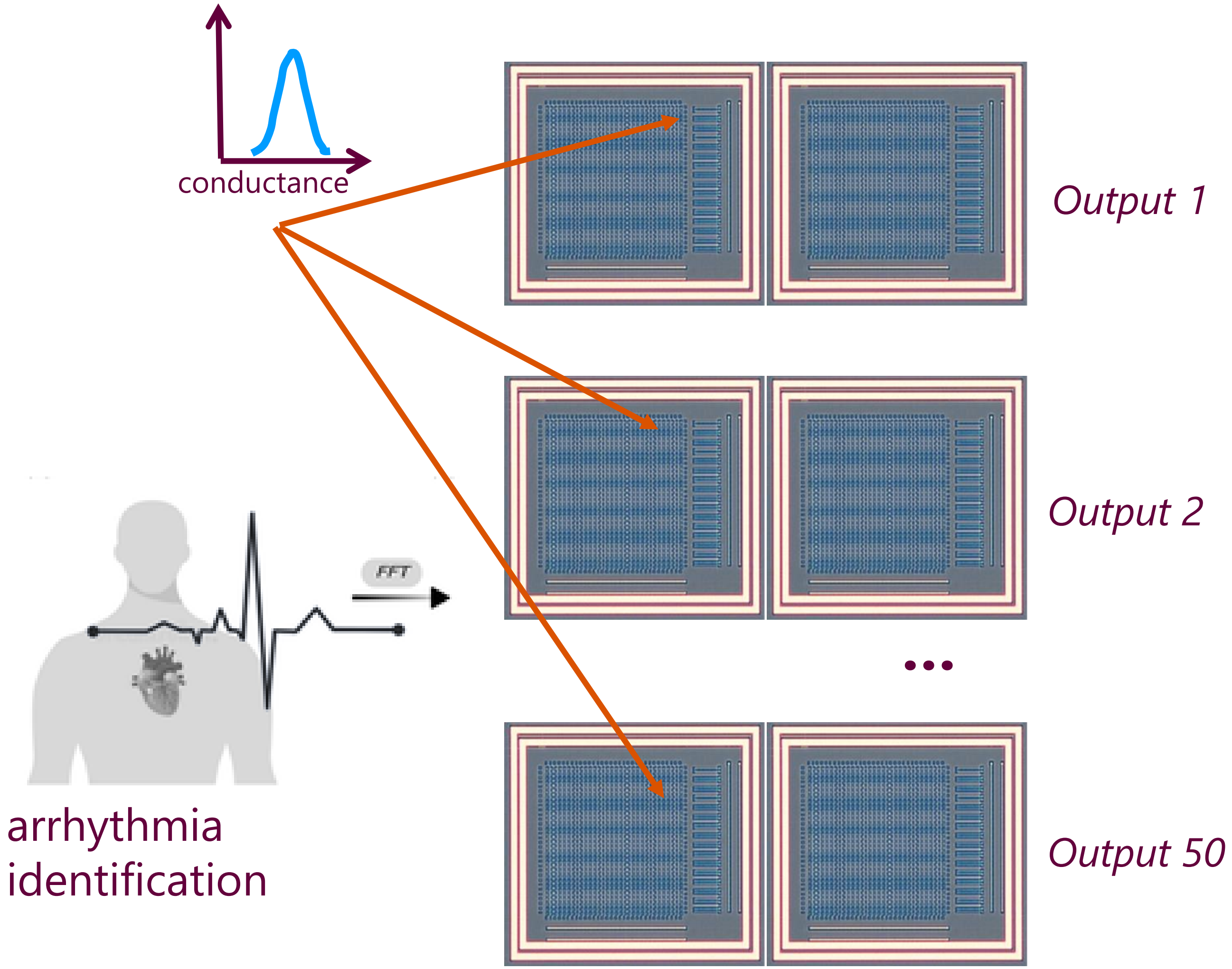

**"I am not really sure what this is, because I have never seen it"**




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



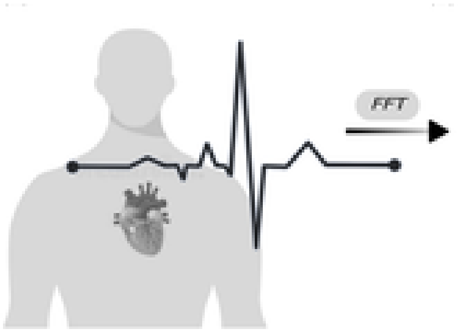

**We get 50 outputs:**
their dispersion tells about the *certainty* of the neural network

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

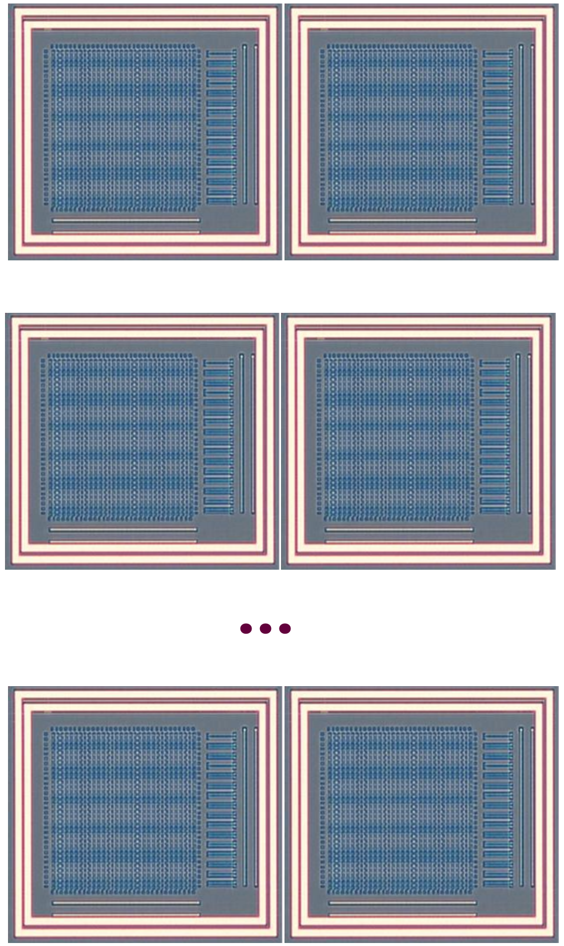
Bonnet et al, Nature Communications 14, 7530 (2023)

# 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*

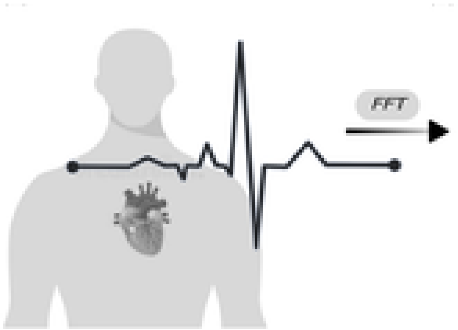
*Arrythmia type 1*

**LOW UNCERTAINTY**

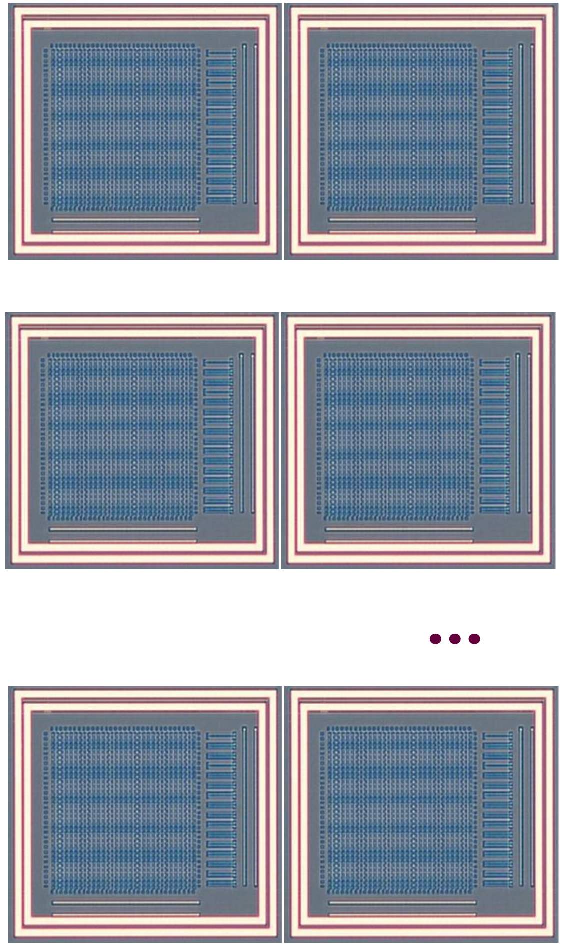
Bonnet et al, Nature Communications 14, 7530 (2023)

# Memristor-Based Bayesian Neural Networks

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




arrhythmia
identification



*Hesitates*
*arrythmia type 1 or 3*

*Hesitates*
*arrythmia type 1 or 2*

*Hesitates*
*arrythmia type 3 or 4*

**Unknown data:**

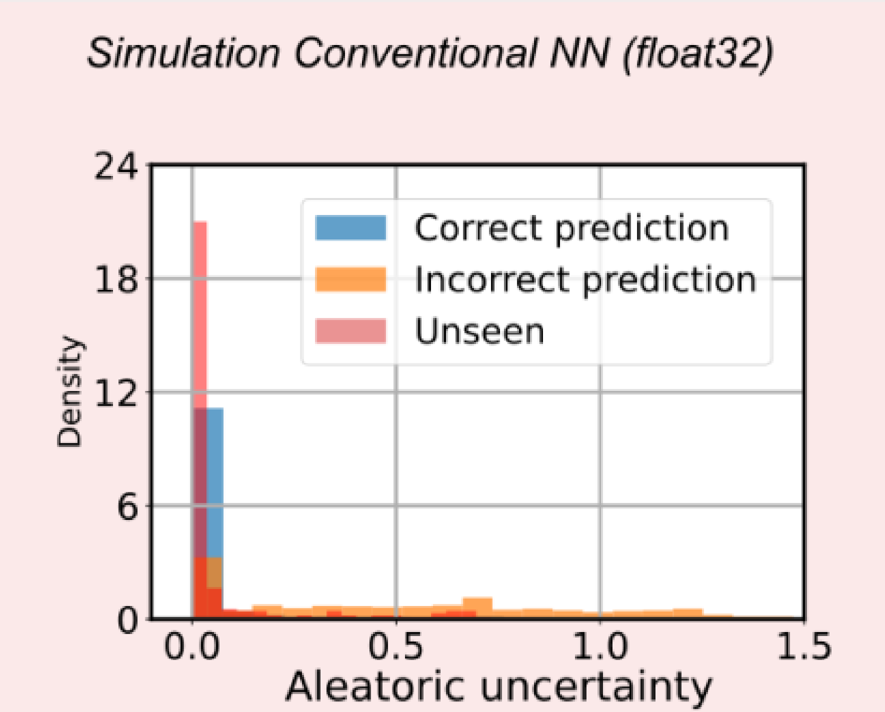
**EPISTEMIC UNCERTAINTY**

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

# Fully Experimental Arrythmia Recogniton

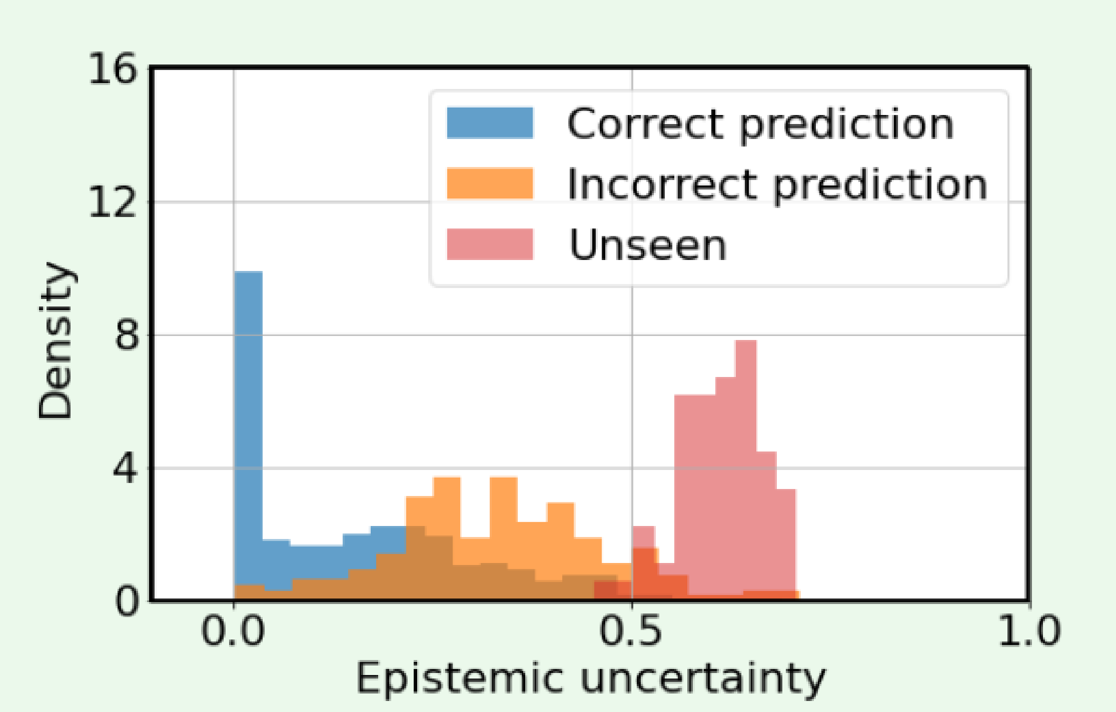
***Traditional Neural Network***

80% classification accuracy




***Bayesian Neural Network***

79% classification accuracy




Easily recognizes
unknown types of arrythmia

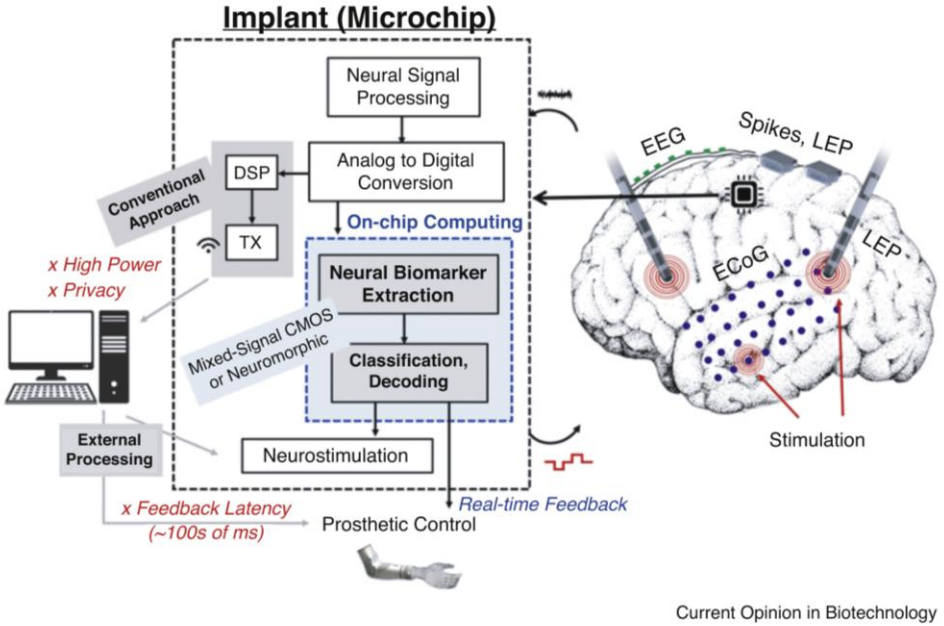
Bonnet et al, Nature Communications 14, 7530 (2023)

# The Grand Challenge of On-Chip Learning

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

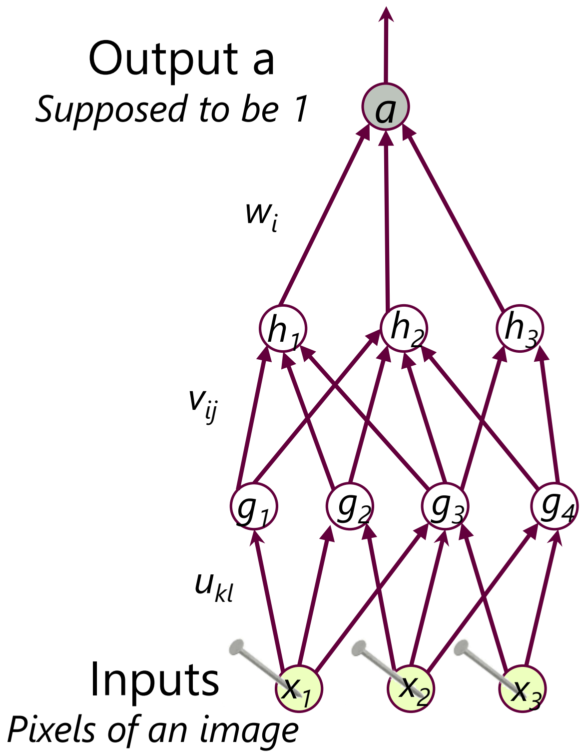
- On-chip learning/adaptation would also have tremendous prospects

*Medical applications*




Yoo and Shoaran,
Current Opinion Biotechnol., 2021

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




$$C = -\ln a$$

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

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

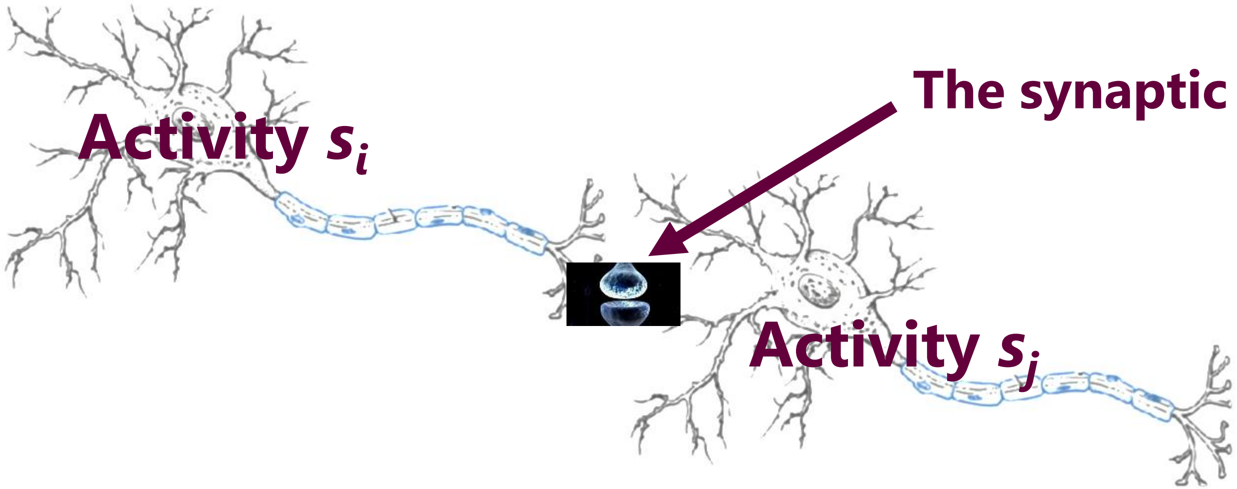
$$\boxed{\frac{\partial C}{\partial u_{kl}} = (a-1)\sum_i w_i h_i(1-h_i)v_{kl}g_k(1-g_k)x_l}$$

**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

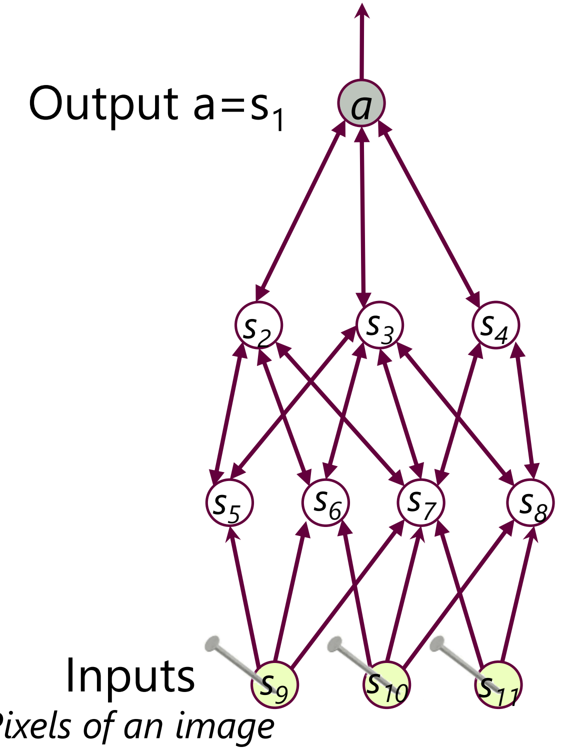
- An old idea (Hebb): « cells that fire together wire together »




- **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

*__Learns with Hebbian-like learning but is equivalent to backpropagation__*




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, $\rho$ : 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*

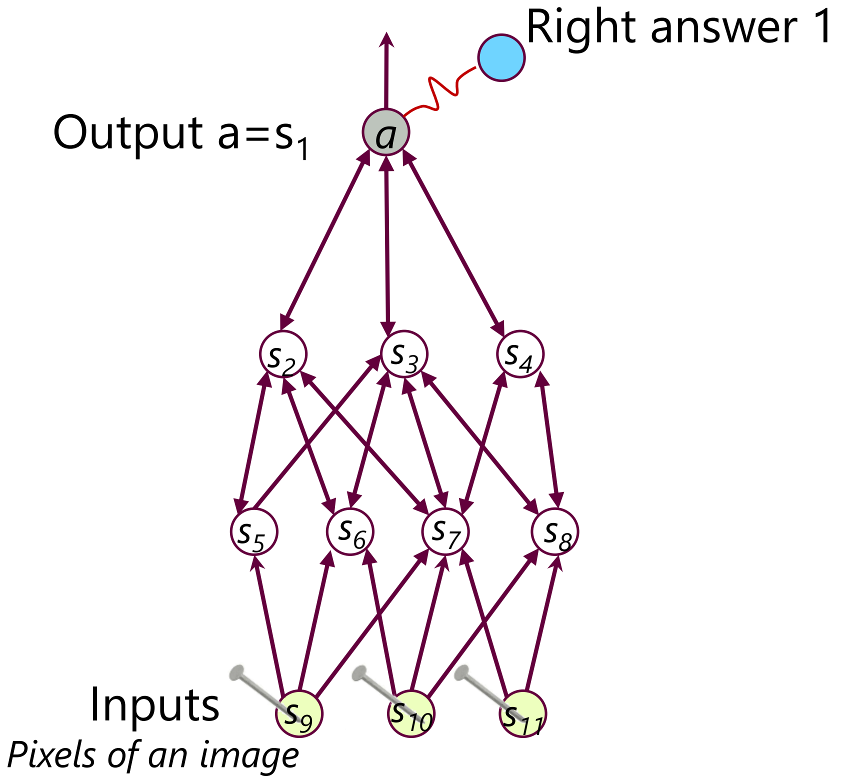
**When the network has converged, we get the output *a***

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

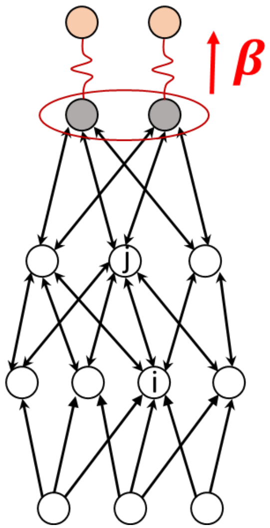
# Equilibrium Propagation

- 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 $w_{ij}$*

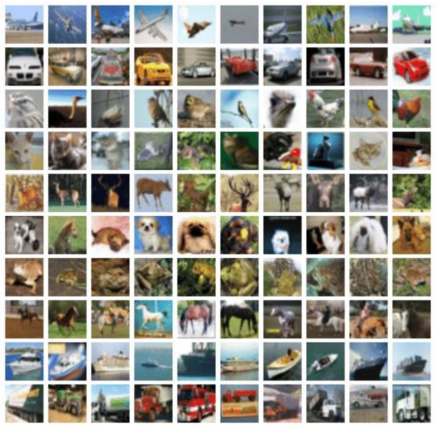
➢ *Otherwise, DECREASE $w_{ij}$*

# 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**

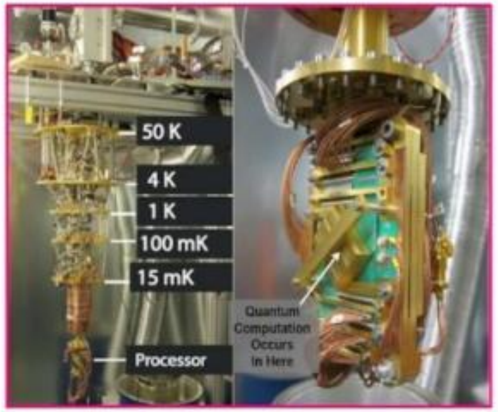
laboratoire Albert Fert C2N Mila

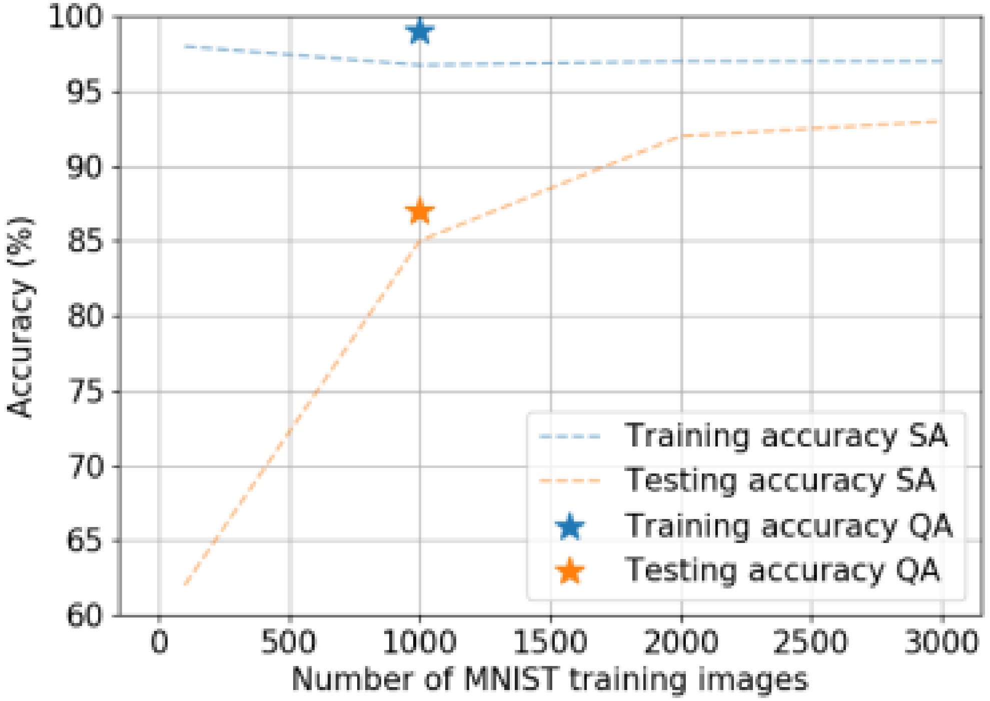


53

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)

# Conclusion

- Analog in-memory computing: huge potential to reinvent electronics for **cognitive-type tasks and AI**

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

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

- **Next ARCHI 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

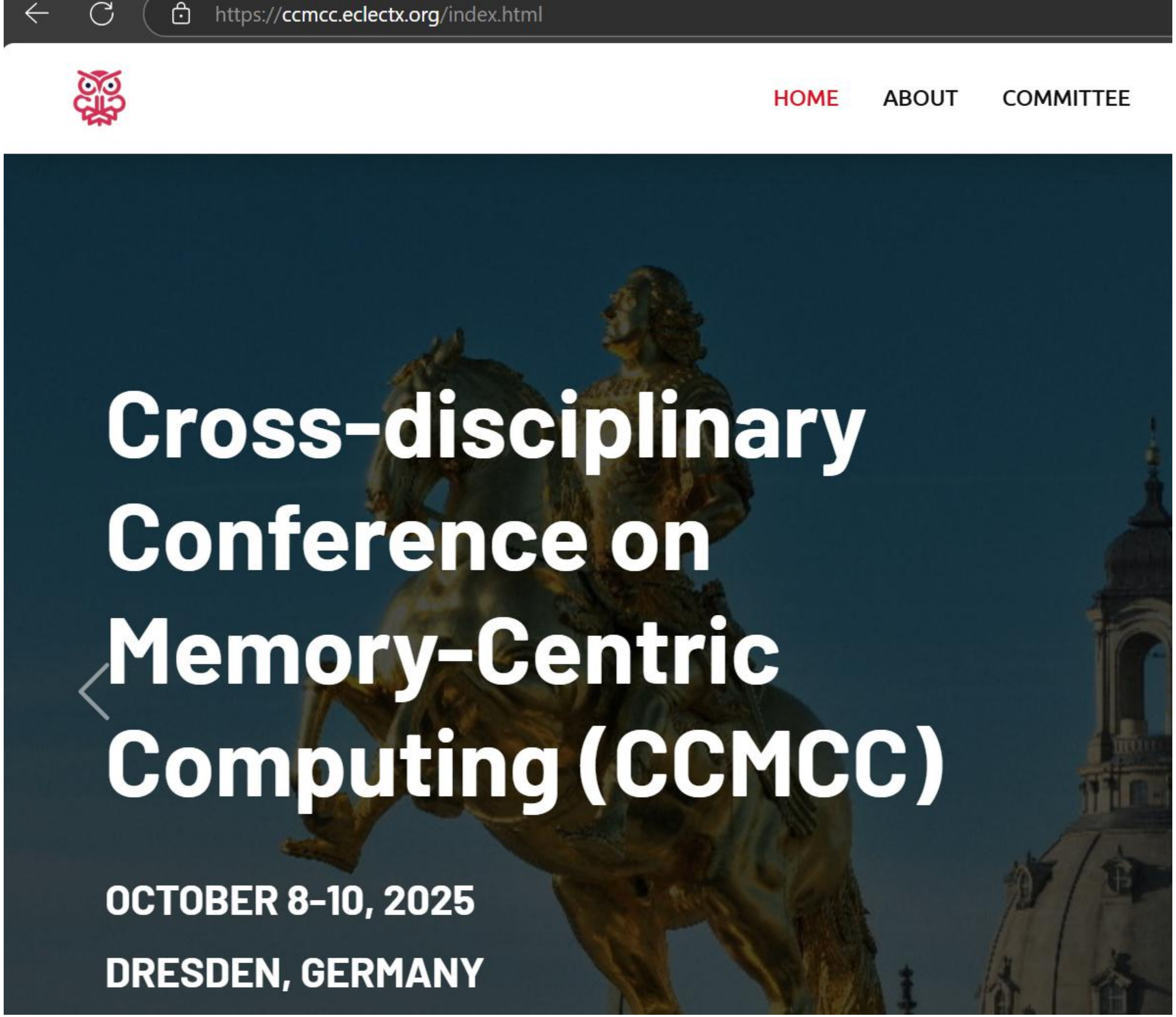
- David Novo
- Pascal Benoit
- Paul Delestrac
- Aymen Romdhane
- Bruno Lovison-Franco

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

- Yoshua Bengio
- Benjamin Scellier

- Jean-Michel Portal
- Marc Bocquet
- Fadi Jebali

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


https://ccmcc.eclectx.org/index.html
HOME
ABOUT
COMMITTEE
Cross-disciplinary Conference on Memory-Centric Computing (CCMCC)
OCTOBER 8-10, 2025
DRESDEN, GERMANY

# Thank you for your attention!

**Postdoc positions available**

**damien.querlioz@universite-paris-saclay.fr**

**https://sites.google.com/site/damienquerlioz/**