In-Memory Computing with Imperfect or Unreliable Memory Devices

Damien Querlioz, Tifenn Hirtzlin, Jacques-Olivier Klein, Clément Turck, Kamel-Eddine Harabi *Université Paris-Saclay, CNRS, C2N, Palaiseau*

Marc Bocquet, Jean-Michel Portal *IM2NP, Aix-Marseille Univ, CNRS*

Elisa Vianello, Thomas Dalgaty, Niccolo Castellani, Etienne Nowak, *CEA, LETI, Grenoble*

Context: the AI Energy Problem

Waymo

- AI of self-driving car prototypes
  ~ 2,000 W
- Human brain ~20W

- Training the GPT-3 language model required 190,000 kWh
- *1,000 years of brain operation!*

Due to this energy problem, most AI is done on the cloud

This Energy Cost Largely Originates from the Separation of Computation and Memory

- In a modern computer

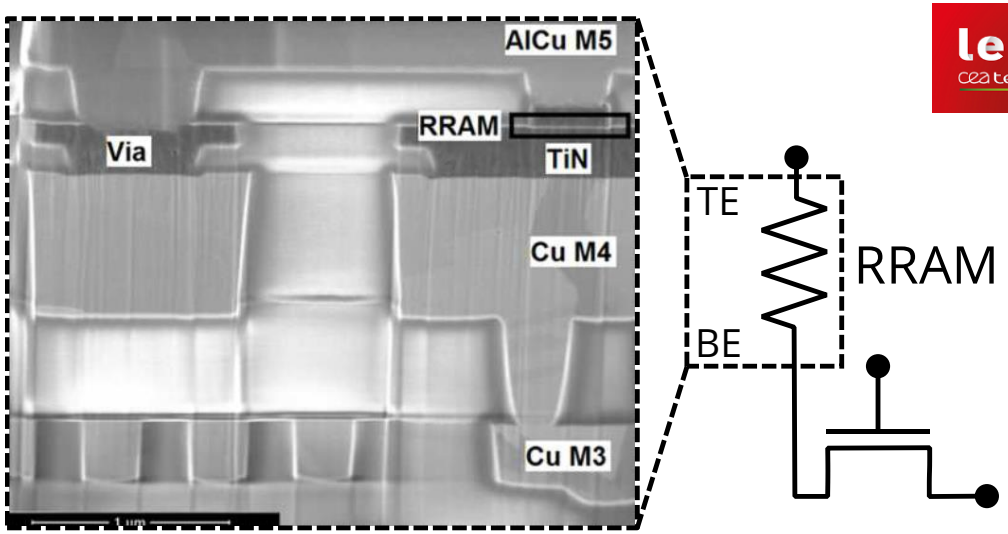
Operation	Energy
Addition of data (fixed point)	1x
Access data (onchip cache)	60x
Access data (offchip RAM)	3500x

Pedram et al , IEEE D&T 2016

Absent in the brain!

In-memory computing approach of the brain could dramatically reduce the energy consumption of AI

The Solution: Resistive Memory (Memristor) and Other Emerging Memories



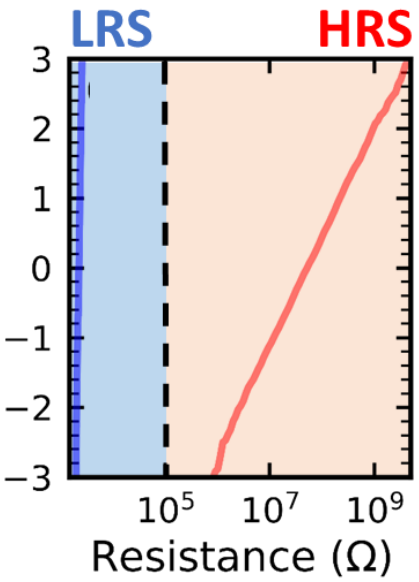
- **Fast, non-volatile memory that can be embedded at the core of CMOS**
- Memory state is the electrical resistance of the device
- Many variations (memristive, phase change / PCM, magnetoresistive / MRAM)
- In industry test production (Samsung, TSMC, Intel, ST Microelectronics...)

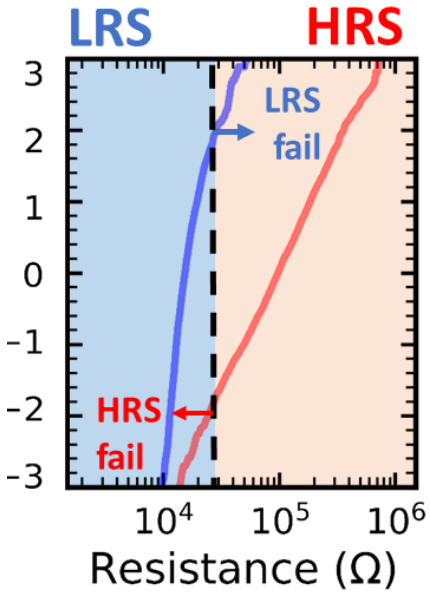
Can be ideal technology for merging logic and memory

The Challenge: Device Imperfection!

- Nanometer-scale physics is very noisy and prone to variations
- Emerging memories well described by the tools of **statistics**

Cumulative Distribution Function (Number of Sigmas)





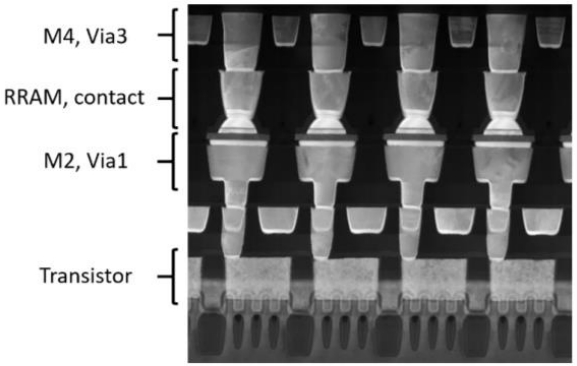
HfO_2 RRAM

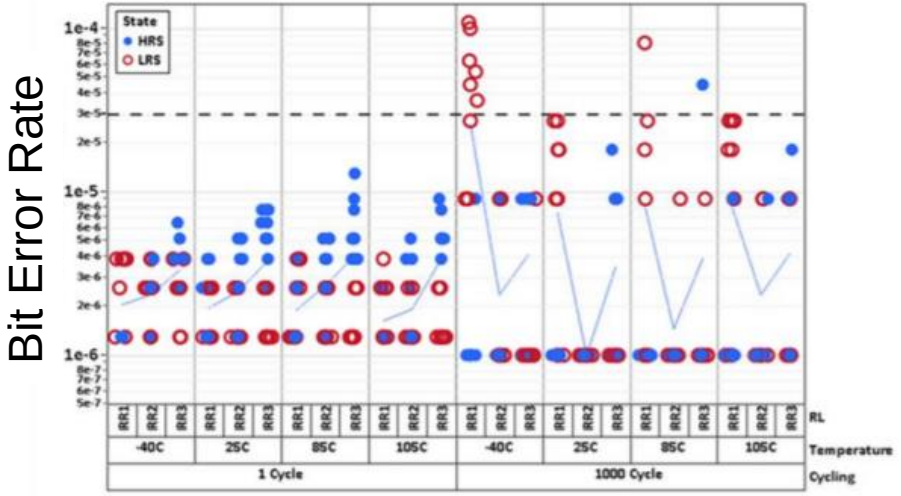
programmed at high voltage

programmed at low voltage

Dealing with Device Imperfections

- Intel RRAM (memristor) 22 nm process





- **Current products with emerging memories: multiple error correcting codes (ECC)**
  - **Intel RRAM and MRAM: ECC corrects up to 3 errors (TEC) per 128b**

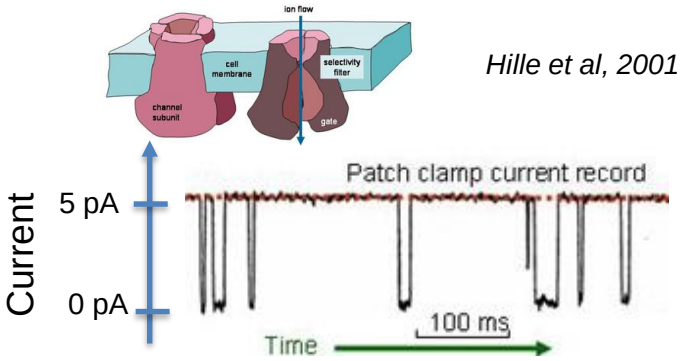
Chang et al, IRPS 2020

Error detection and correction is costly in terms of delay, energy, area

Dealing with Device Imperfections

- The brain also functions with noisy nanodevices and does **not** use formal ECC

Ion channels

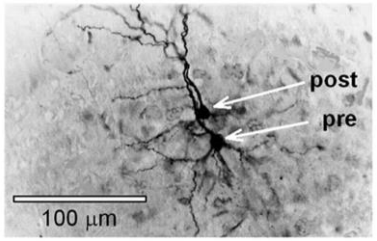


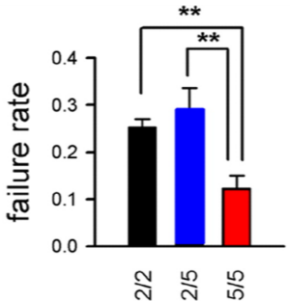
Hille et al, 2001

Ion channel's current is pure telegraph noise

Synapses

Hardingham et al, J. Neurosci, 2010

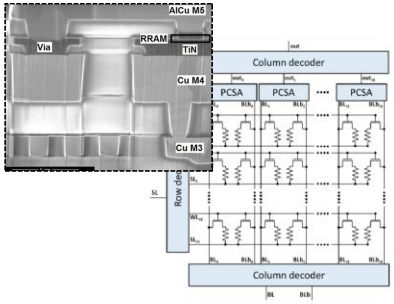




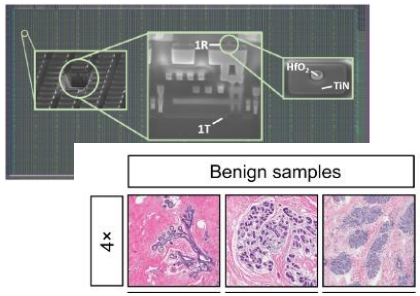
Synapses are massively unreliable!

This talk: embrace the statistical nature of emerging memories for neuromorphic computing

Memory-Centric Artificial Intelligence



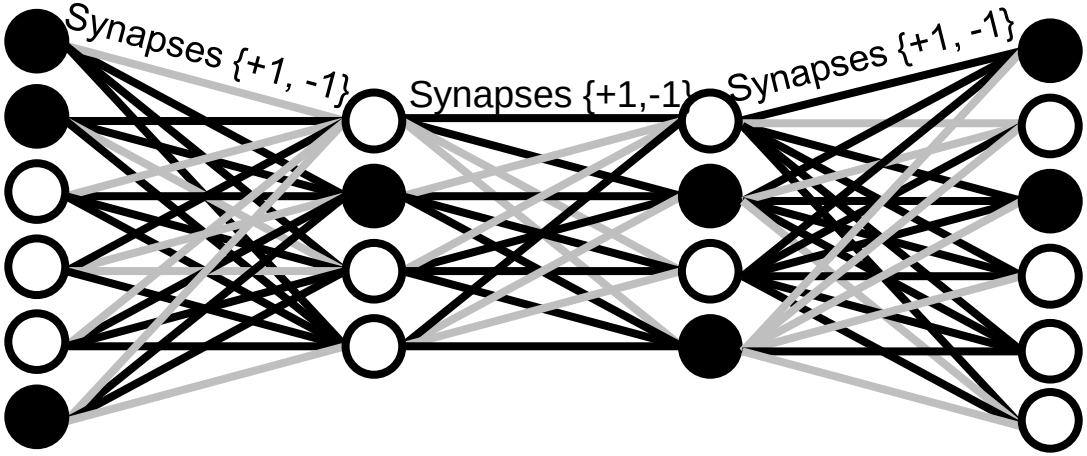
- Circuits that work even if the devices make many errors – *Binarized Neural Networks*

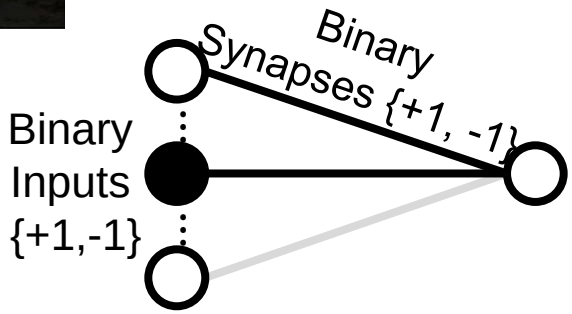


- Exploiting the statistical nature of memristors – *Markov Chain Monte Carlo learning*

Binarized Neural network: A “Super Low Precison” Neural Networks

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





Our work :
RRAM as binary synapse

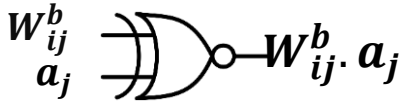
Can approach state of the art performance on vision tasks!

Binarized Neural network: Very Simple Logical Operations

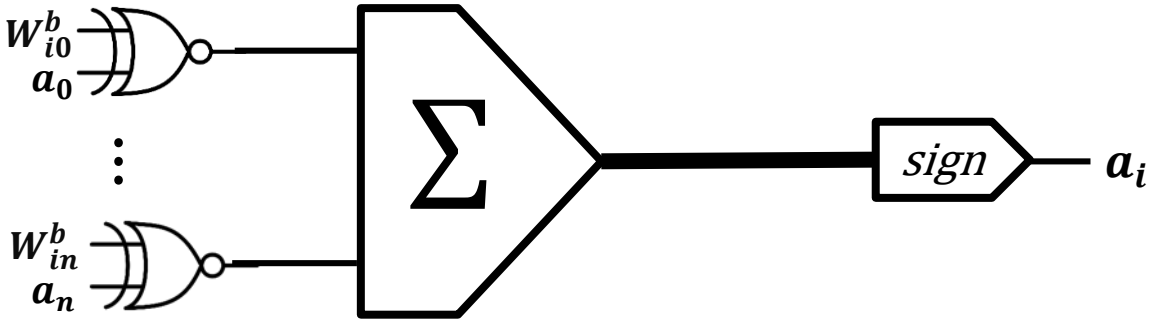
Multiplication → Accumulation → Non-linear function

$$W_{ij} \cdot a_j \qquad \sum_j W_{ij} \cdot a_j \qquad a_i = f\left(\sum_j W_{ij} \cdot a_j\right)$$

XNOR → Bitcount → Sign



a_j	W^b_{ij}	$W^b_{ij}.a_j$
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1



Binarized Neural Networks inference is particularly adapted for in-Memory / near-Memory Computing

Training Is *Not* Binarized

- During training, synapses are associated with a *hidden* real weight
- The binary weight is the sign of the real weight

Binarized Neural Networks are truly attractive for **inference** hardware

How to Deal With Bit Errors?

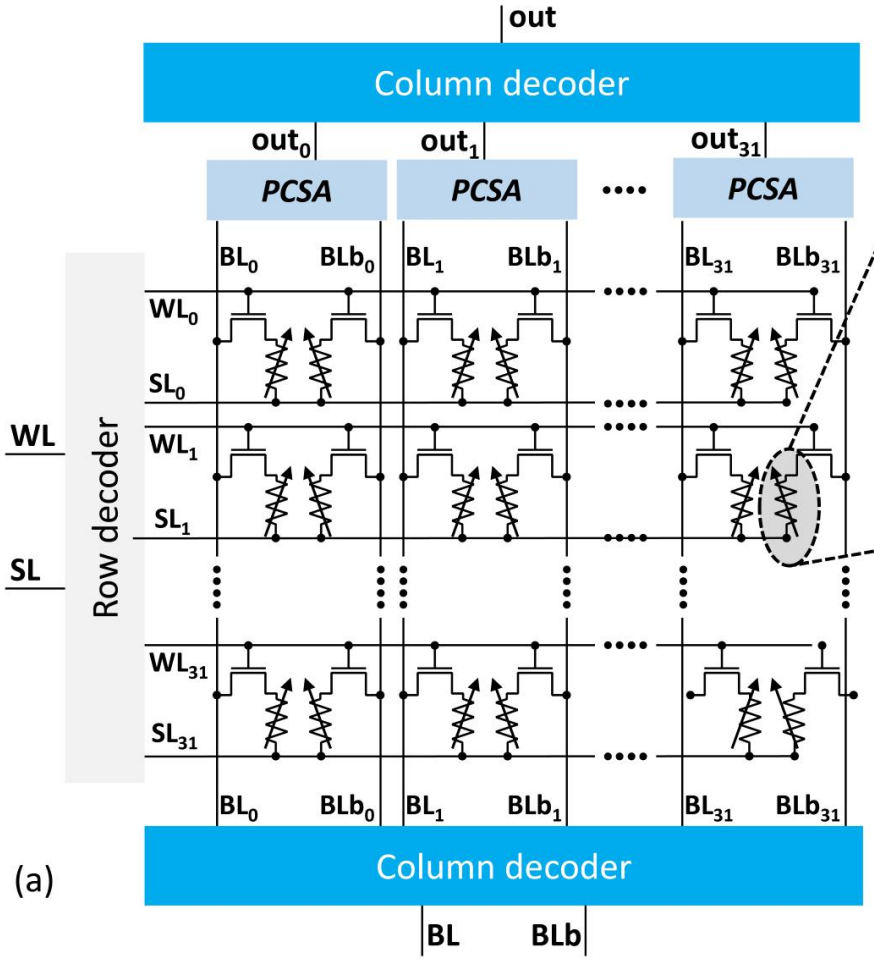
Classical Approach : Error Correction Code

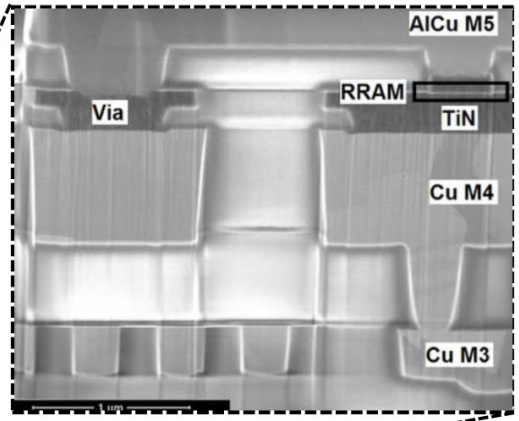
→ **A problem with in-memory computing**

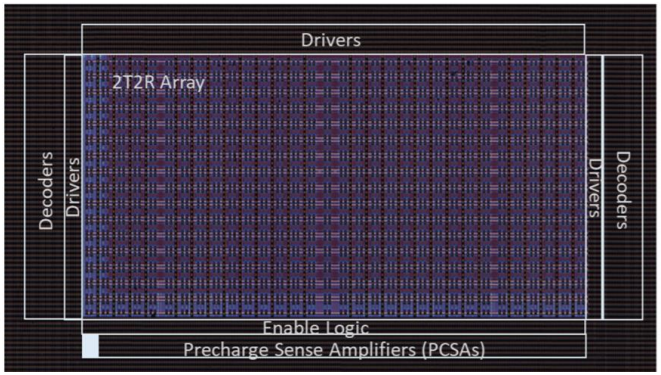
ECC decoding would use most of the area and energy consumption!

Memory Array for ECC-less In-Memory Binarized Neural Networks

130nm CMOS + HfO_2 RRAM

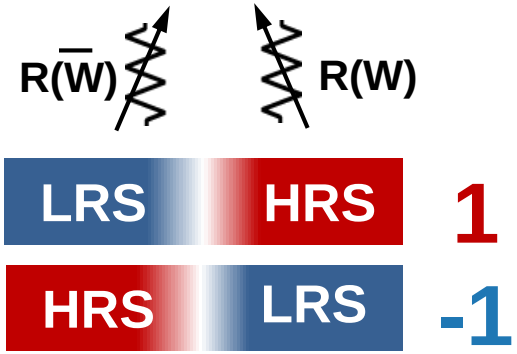






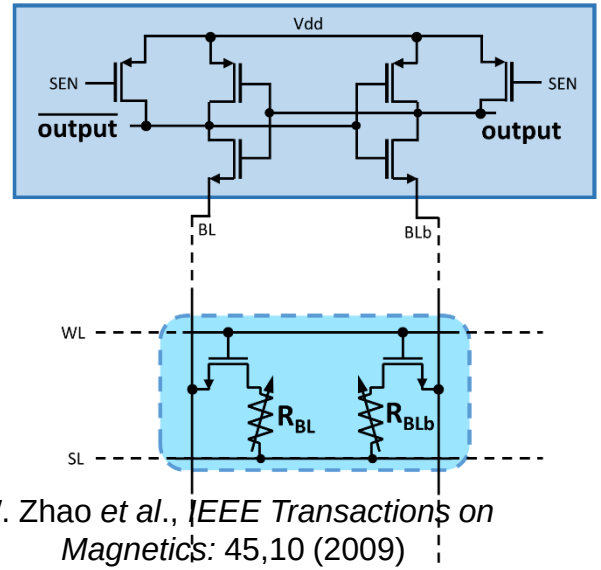
Reducing the Error Rate without ECC: Two RRAM Devices as One Binary Synapse

Devices programmed in a complementary fashion



Circuit to differentiate resistance state

Pre-Charge Sense Amplifier (PCSA)



W. Zhao *et al.*, *IEEE Transactions on Magnetics:* 45,10 (2009)

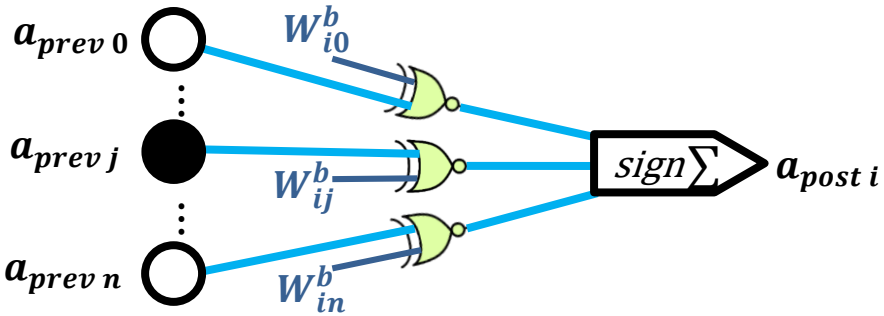
Doubles the number of devices

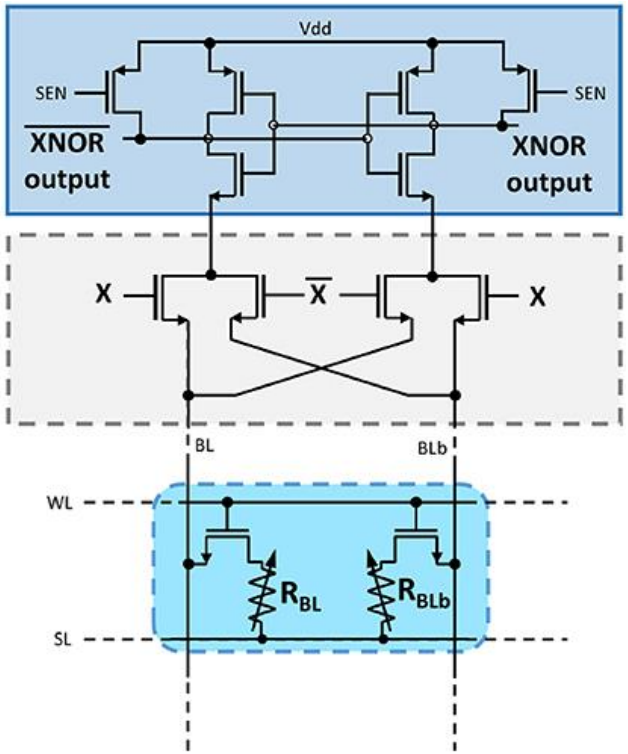
Bocquet et al, IEEE IEDM, p. 20.6.1, 2018; Hirtzlin et al., Front. Neurosci. 13, p. 1383, 2020

Logic in Memory Reading Circuit

- No ECC offers opportunity for in memory operation
- XNOR operation directly in sense amplifier circuit

XNOR → Bitcount → Sign





Sense Amp.

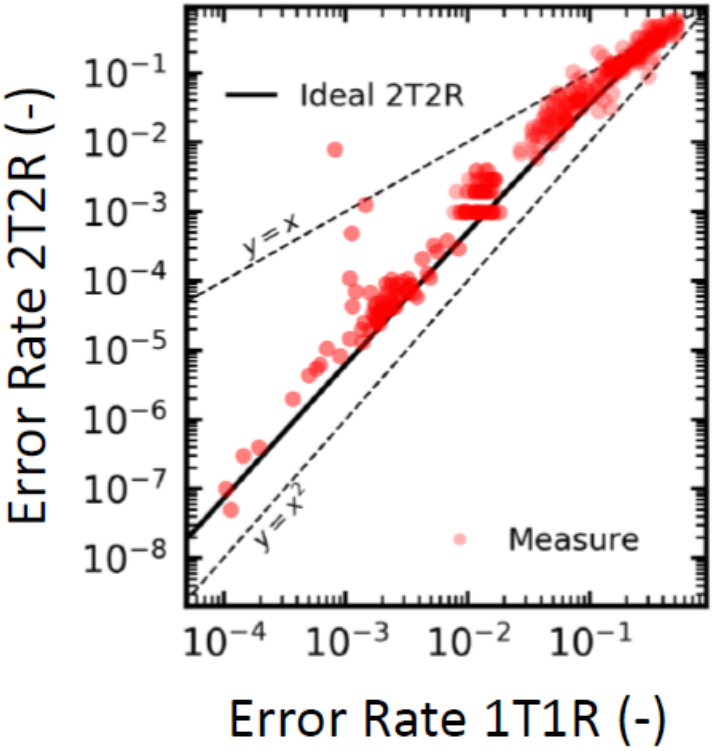
a_{prev0} input

RRAM matrix (2T2R)

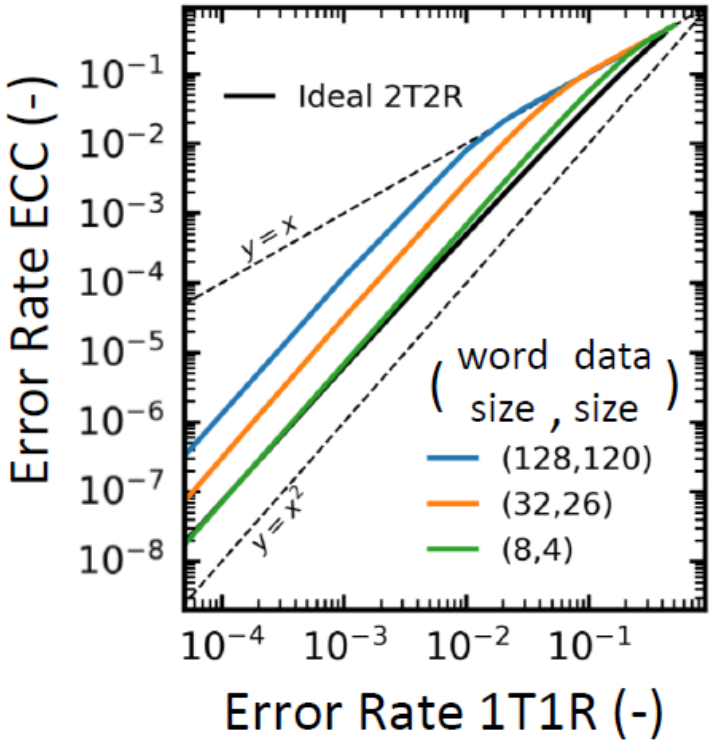
Weisheng Zhao, … Klein, Querlioz …., *IEEE TCAS I:* 2014.
IEEE Guillemin-Cauer Best Paper Award

2T2R Architecture to Limit Bit Errors without ECC Decoding

- Experimental bit error rate

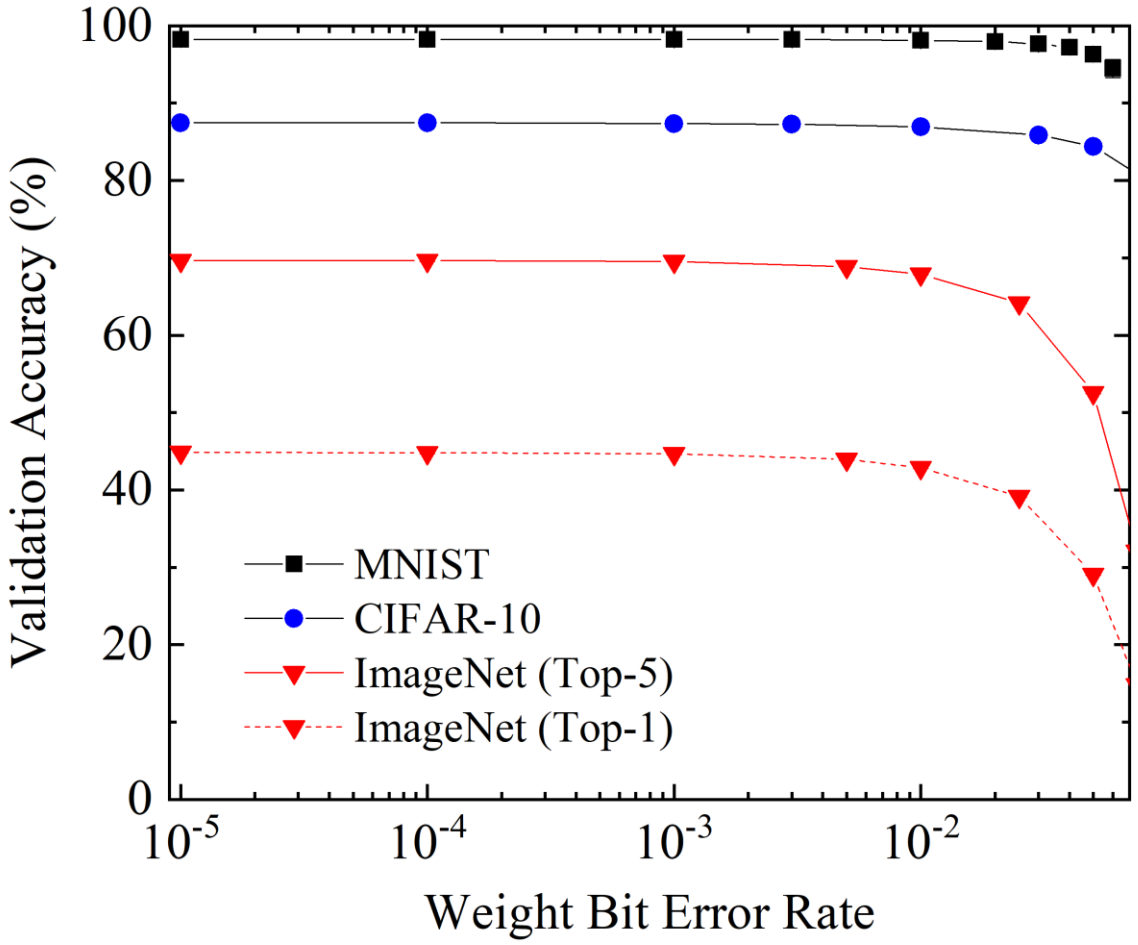


- Error Correction Code SECDED

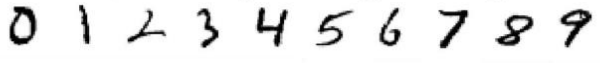


Bocquet et al, IEEE IEDM, p. 20.6.1, 2018; Hirtzlin et al., Front. Neurosci. 13, p. 1383, 2020

Do All Errors Really Need to Be Corrected?

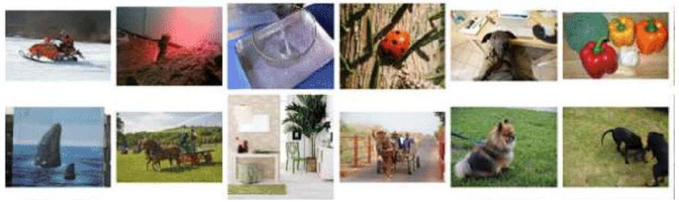


MNIST



CIFAR-10

ImageNet



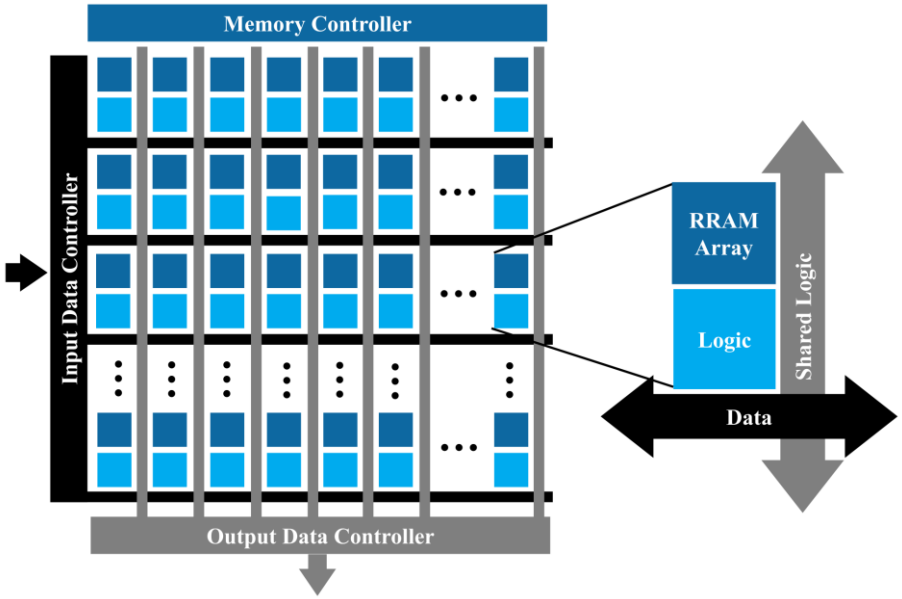
Bocquet et al, IEEE IEDM, 2018; Hirtzlin et al, IEEE AICAS 2019

Benefits of Embracing Errors

	Typical	Optimized Endurance	Optimized Programming Energy
SET Compliance	200 µA	200 µA	200 µA
RESET Voltage	2.5 V	1.5 V	2 V
Programming time	1 µs	1 µs	0.1 µs
2T2R BER	**$<10^{-7}$**	**$<10^{-4}$**	**$<10^{-5}$**
Programming Energy (SET/RESET)	**~ 300 pJ**	**~ 300 pJ**	**~ 25 pJ**
Cyclability	**$>10^{8}$**	**$>10^{10}$**	**$>10^{8}$**

Bocquet et al, IEEE IEDM, p. 20.6.1, 2018; Hirtzlin et al., Front. Neurosci. 13, p. 1383, 2020

New Design: Fully Digital Architecture



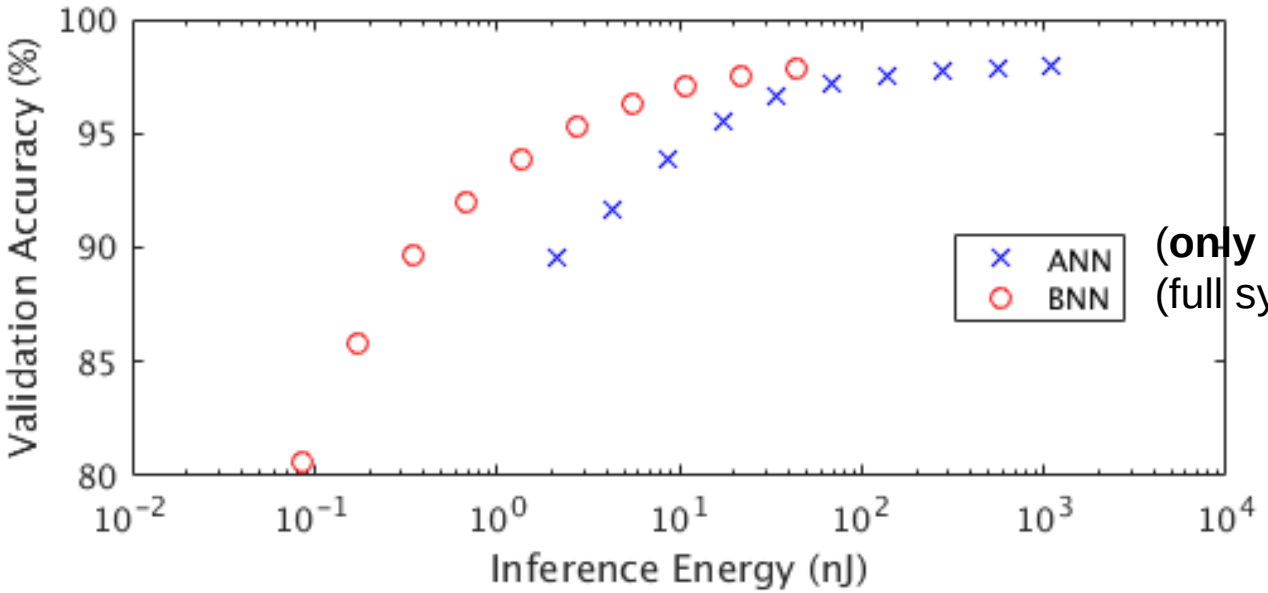
- The sum is achieved with few bits integer digital circuits
  - *We do not use Kirchhoff laws*

Energy Benefits of BNN

Inference Energy for Recognizing a MNIST Digit:

- **CPU / GPU ~mJ**

- **In-memory ASIC:**



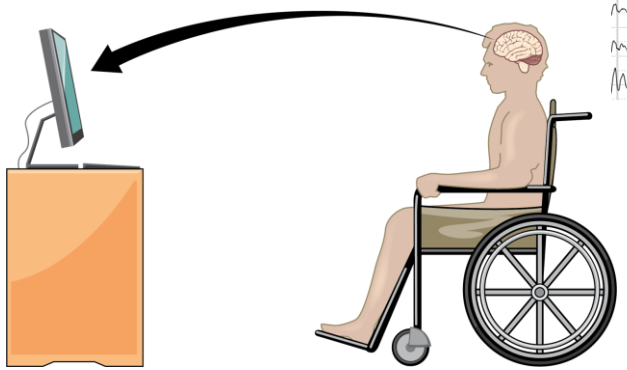
(**only arithmetic operations**, 8 bits)
(full system)

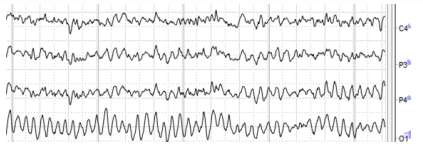
Analysis with Cadence encounter, and
DK of a 28 nm commercial technology

Hirtzlin et al., Front. Neurosci. 13, p. 1383, 2020

Possible Application of In-Memory BNN: Edge Analysis of Medical Signals

Armo





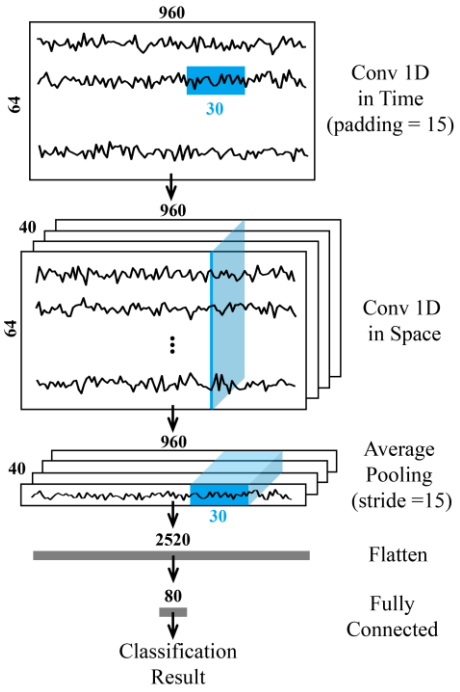
Wikimedia

- Could allow detection of strokes, epileptic seizures, heart attacks or BCI
- Without relying on the cloud (better privacy, security, reliability)

However, AI algorithms have high energy cost

B. Penkovsky et al., DATE, p. 690, 2020

Application to Biomedical Signal Analysis

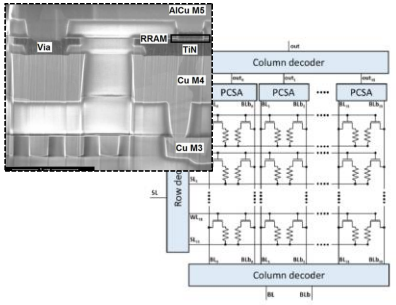


- We tried two tasks
  - ECG electrodes misplacement
  - EEG motor movement/imagery
- 1-D convolutional neural network

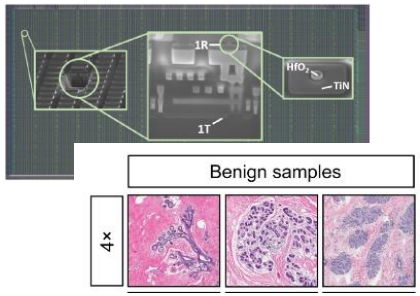
BNN implementation saves 76% memory wrt. 8-bit precison neural network for ECG (58% for EEG), at equivalent accuracy

B. Penkovsky et al., DATE, p. 690, 2020

Memory-Centric Artificial Intelligence



- Circuits that work even if the devices make many errors – *Binarized Neural Networks*

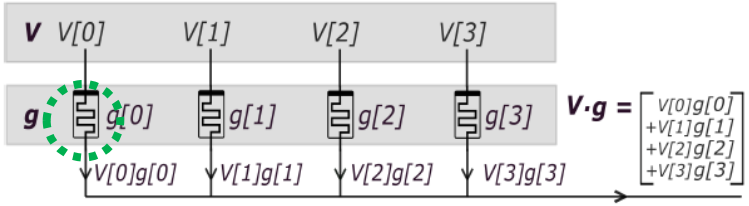


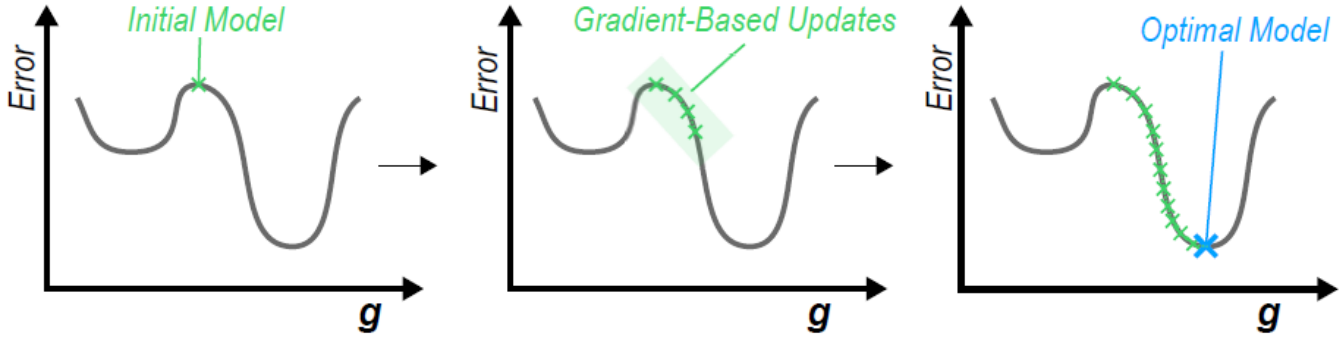
- Exploiting the statistical nature of memristors – *Markov Chain Monte Carlo learning*

What Happens If We Implement Backpropagation with Memristors

New challenge: learning with memristors

Idea: memristor conductance represents synaptic weights (real)



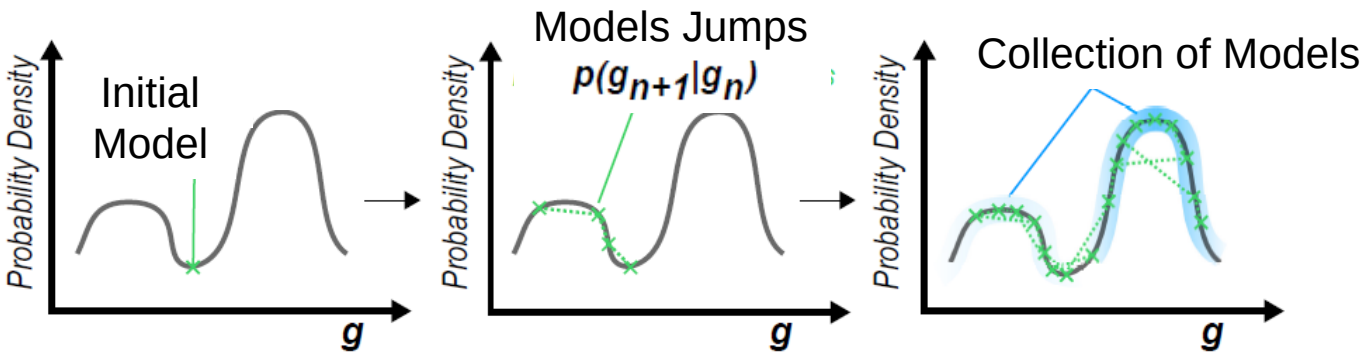


Backpropagation requires very fine tuning of the weight:

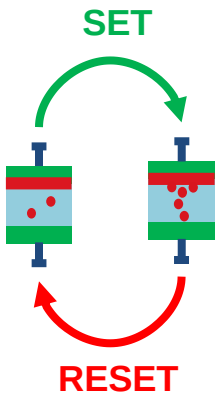
difficult to achieve due to memristor imperfections!

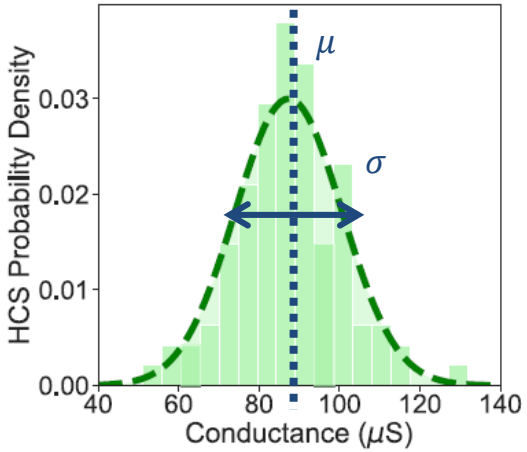
Learning by Embracing the Statistical Nature of Memristors

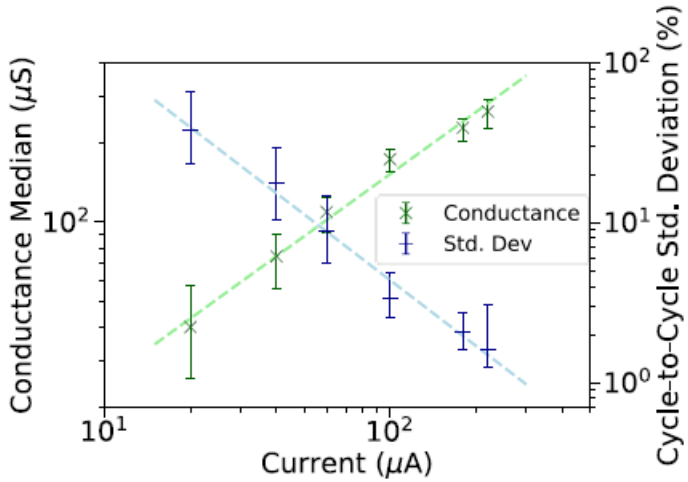
Learning with Metropolis-Hastings Markov Chain Monte Carlo (MCMC)



The jumps $p(g_{n+1}|g_n)$ can be performed easily using the statistical behavior of memristors!

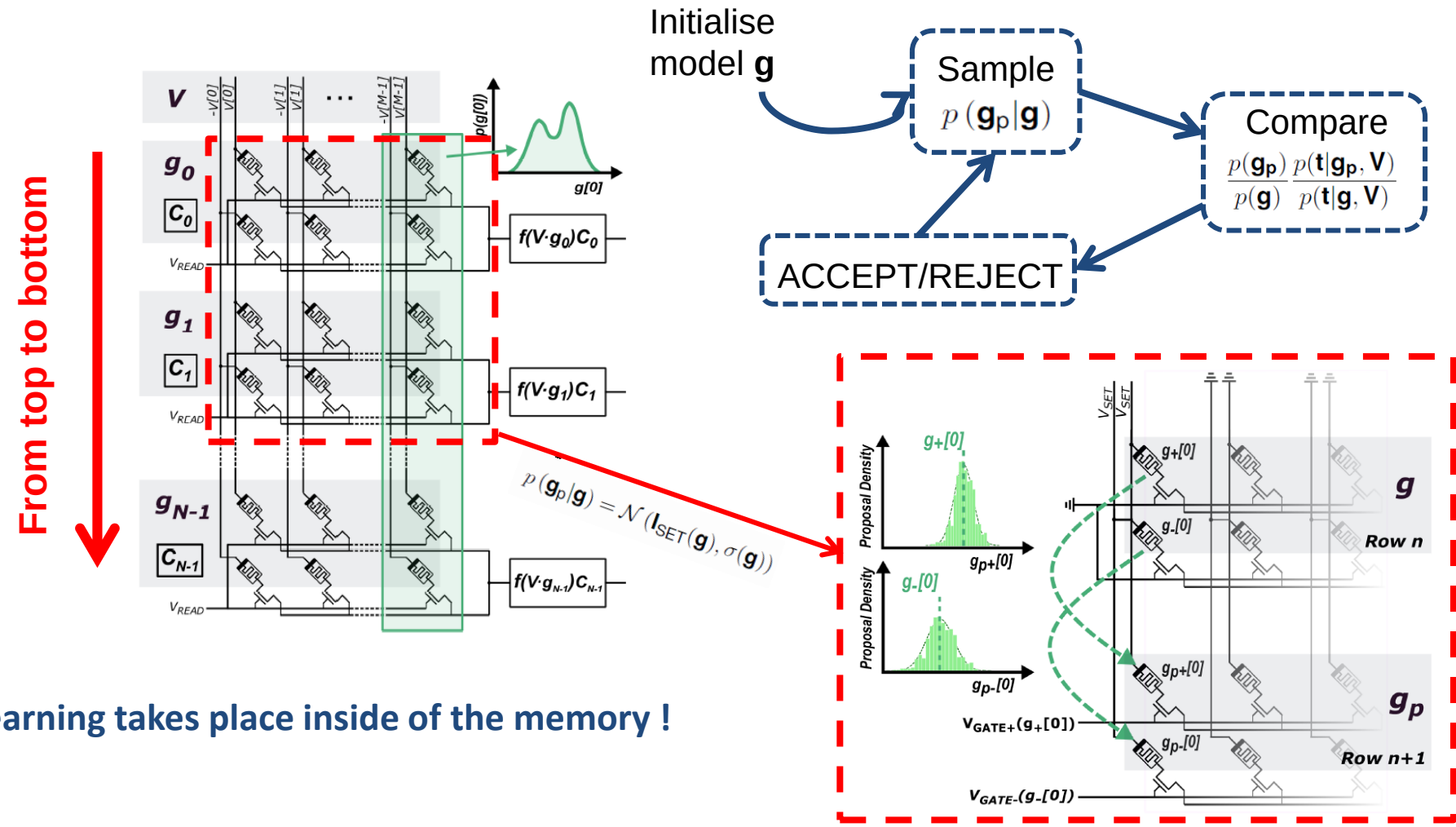






Memristors are ideal for MCMC-based learning!

Memristor-Based MCMC in Practice

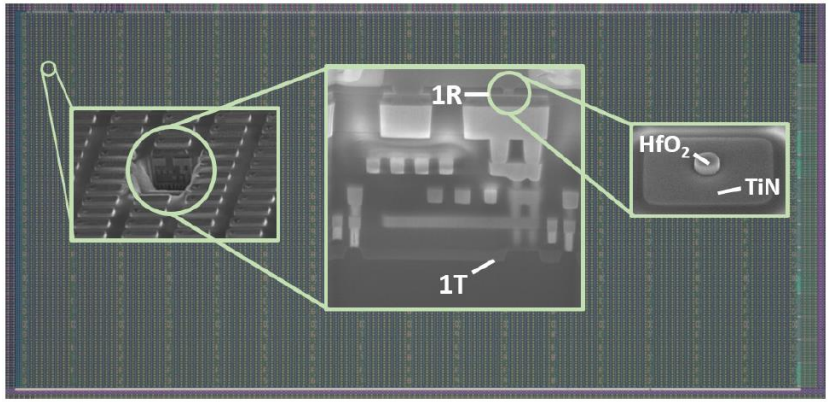


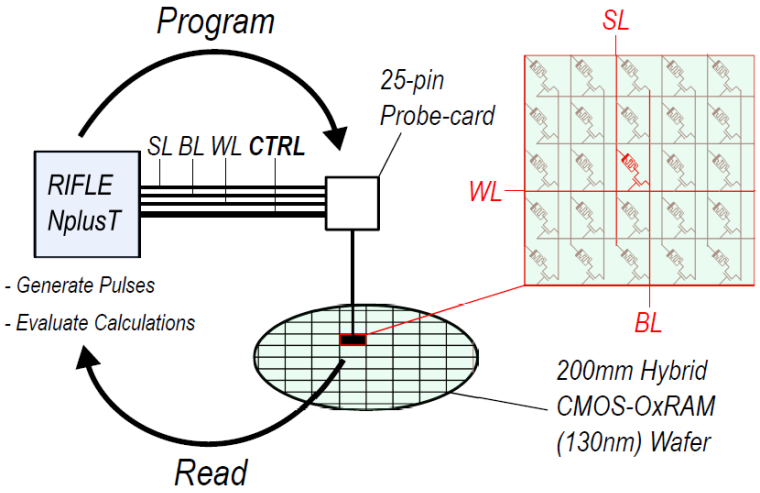
Learning takes place inside of the memory !

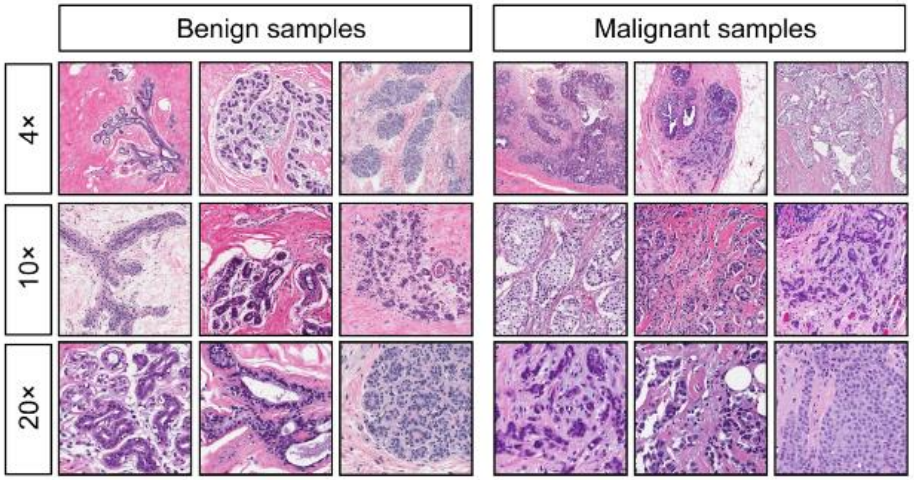
Dalgaty et al, Nature Electronics 4, p. 151, 2021

Supervised Learning with Memristor-Based MCMC

Computer-in-the-loop experiment with an array of 16,384 memristors



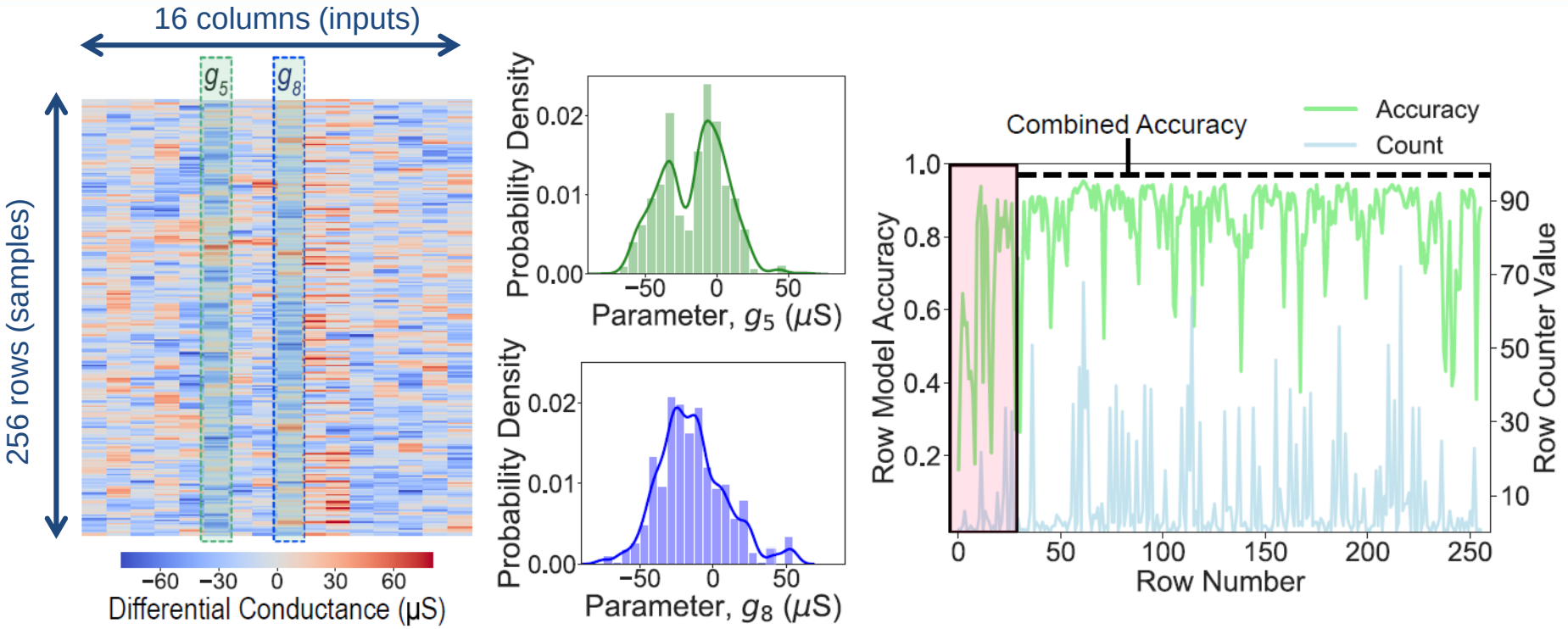




Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, *43*(4), 570-577.

Dalgaty et al, Nature Electronics 4, p. 151, 2021

Supervised Learning with Memristor-Based MCMC

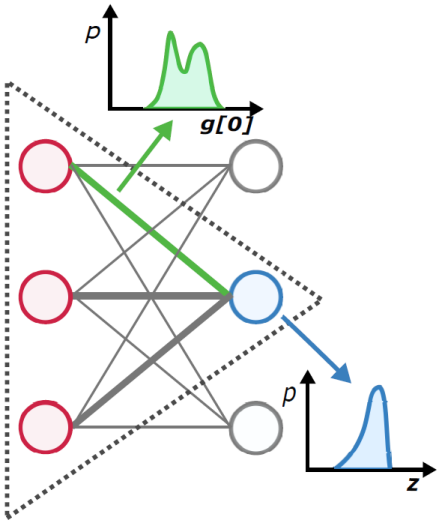


The experimental system was able to detect maligant tissue with 98% accuracy

Dalgaty et al, Nature Electronics 4, p. 151, 2021

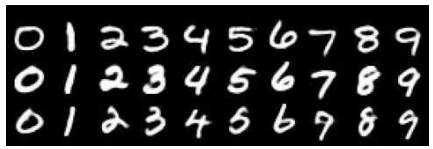
MCMC Learning Leads to a collection of Models: It Provides a *Bayesian* Model

Bayesian model parameters / activations are probability distributions describing uncertainty

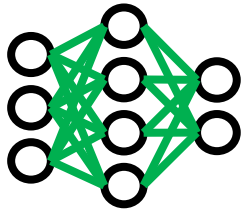


Particularly adapted for the "small data" world, which has a lot of uncertainity

Why might output uncertainty be useful ?

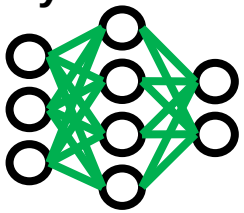


Deterministic model



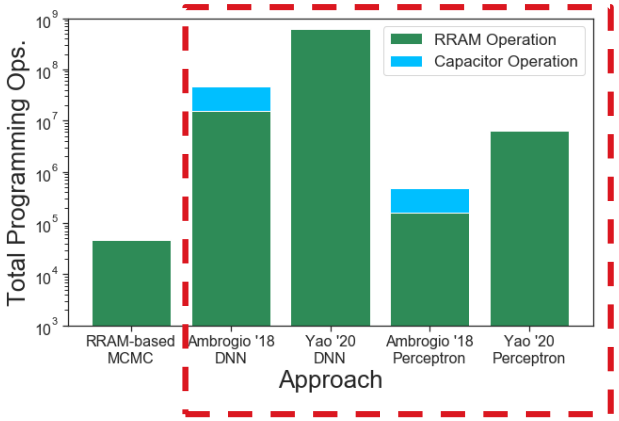
'You have input a 3'

Bayesian model



'The input looks most like a 3… but I am very uncertain about that'

MCMC Learning Is Highly Energy-Efficient



RRAM-based backpropagation

	Step 1 (Model evaluation)	Step 2 (Model acceptance/rejection)	Step 3 (RRAM programming)	Total
Number of repetitions	$500 \times 10 \times 512$	10×512	10×512	
Total energy (130nm)	5.8μJ	$120n$J	1.1μJ	**6.9μJ**
Total energy (28nm)	2.5μJ	$34n$J	1.1μJ	**3.6μJ**

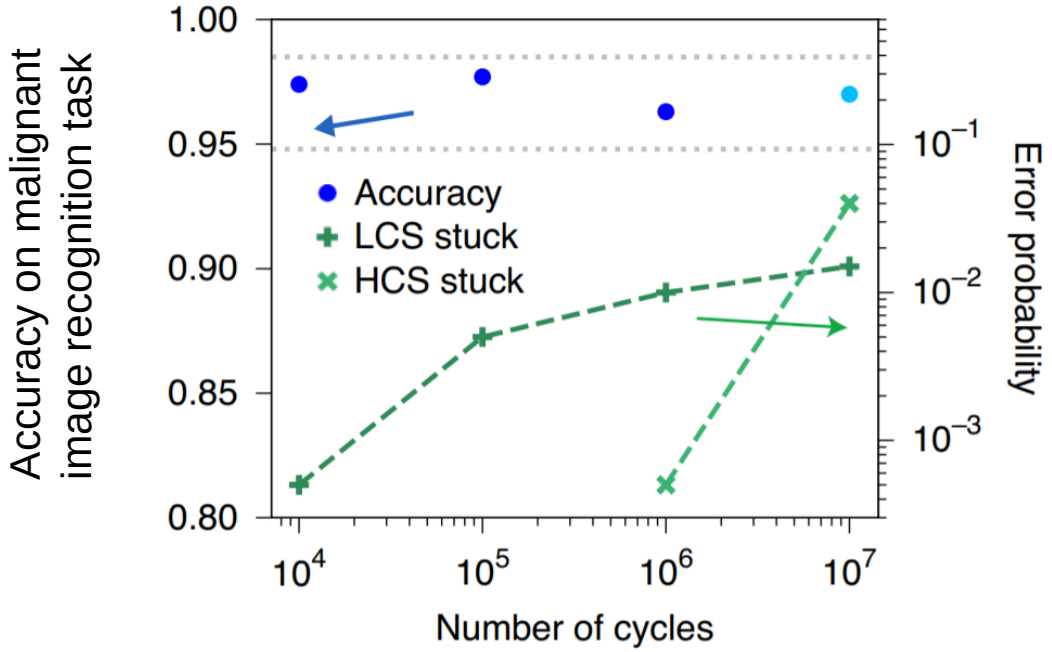
Preliminary mixed-signal design results (full model training)

Intel Xeon processor (7nm) implementation of MCMC sampling required **600mJ**

Dalgaty et al, Nature Electronics 4, p. 151, 2021

MCMC Learning Is Highly Robust to Device Imperfection

RRAM devices are prone to aging

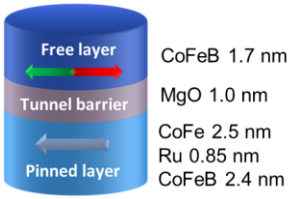


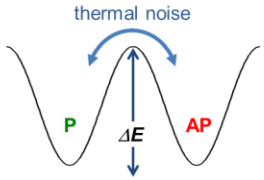
MCMC still works even if devices have been programmed millions of times, and would be unusable as conventional memory

Dalgaty et al, Nature Electronics 4, p. 151, 2021

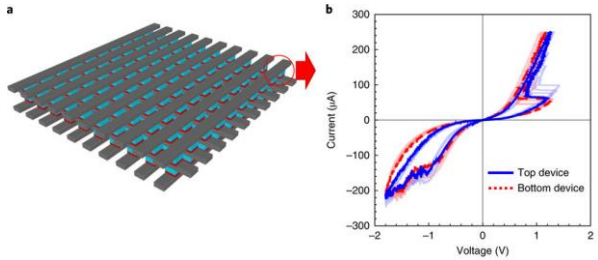
Other Approaches that Exploit Memory Devices Imperfections

- TRNGs and PUFs



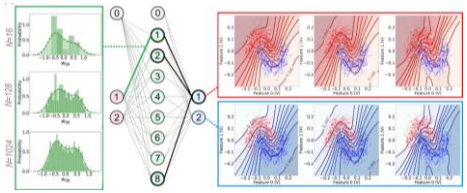


Vodenicarevic et al., Phys. Rev. Appl (2017)



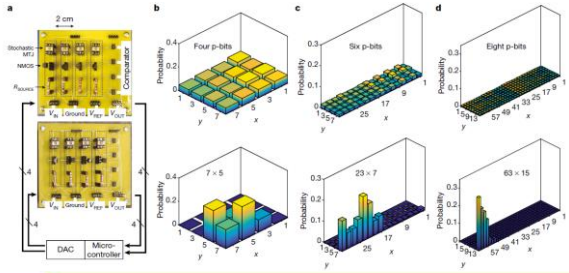
Nili et al., Nature Electronics 1, 197 (2018)

- Bayesian Neural Network Inference



Dalgaty et al., Adv. Intelligent Systems (2021)

- Stochastic Computing-Based Optimization



Borders et al., Nature 573, 390 (2019)

Conclusions

- Binarized Neural Networks excellent candidates for In-Memory Computing with emerging memories
- Embracing bit errors has important benefits in terms of programming energy, cell area and reliability
- The statistical nature of the devices can even be *exploited* for learning Bayesian models

Thank You for Your Attention!

I am hiring postdocs