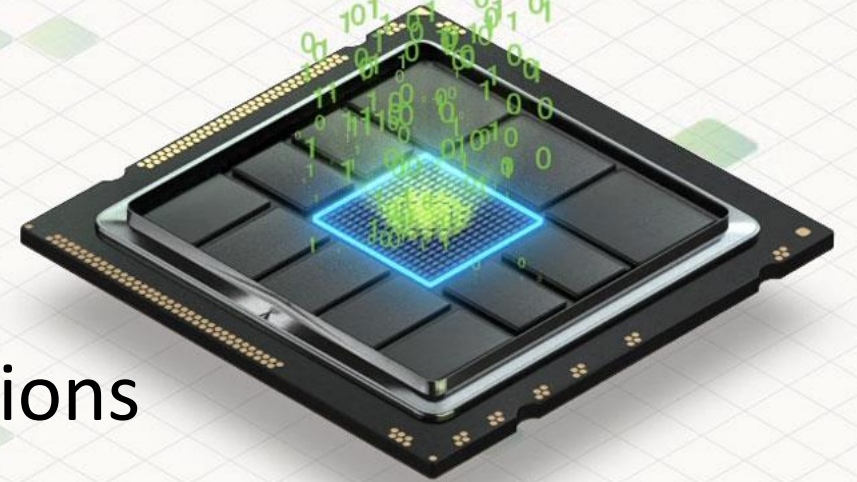




Controlling ReRAM Device Properties to Address Neuromorphic Computing Specifications

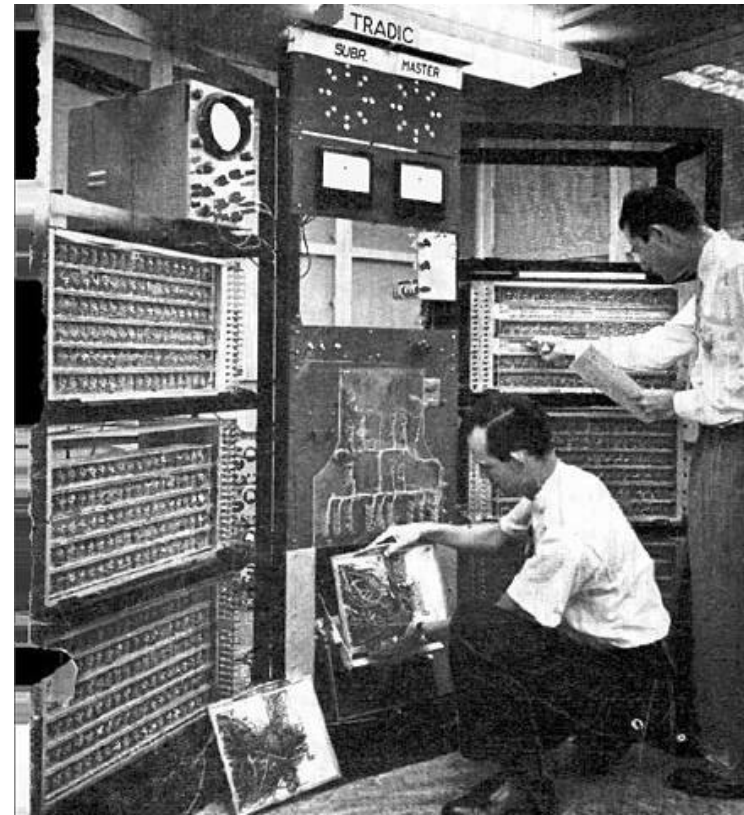
By Gabriel Molas,
Chief Scientific Officer, Weebit Nano
IEEE Senior Member



AI Revolution

70 Years Ago

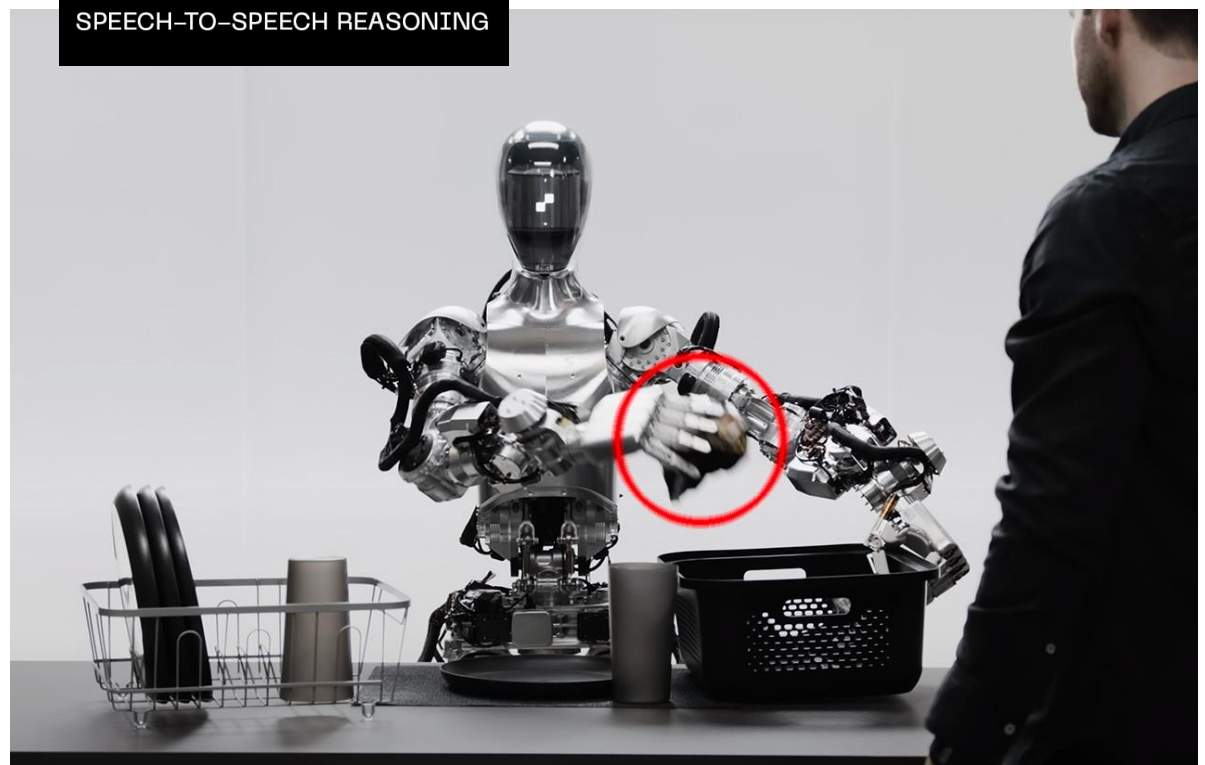
First transistorized computer (Bell Labs)



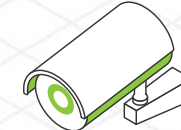
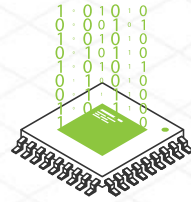
Now

FIGURE 01 + OPENAI (Figure AI)

FIGURE 01 + OPENAI
SPEECH-TO-SPEECH REASONING



ReRAM Fits Edge AI Requirements



	Mixed-Signal / Power Mgmt	IoT / MCUs	Edge AI	Automotive	Aerospace & Defense
Back-end-of-line tech for easy analog integration	○				
Cost-efficiency	○	○	✓	○	
Ultra-low power consumption	○	○	✓		
Robustness in high temp / extreme environments	○	○		○	○
Scaling advantage at 28nm and below		○	✓	○	
High Endurance		○		○	○
Small footprint to store very large arrays			✓	○	
Longevity		○		○	○
Roadmap to neuromorphic computing			✓		

Outline

❖ ReRAM characteristics

- ◆ Device basics
- ◆ State-of-the-art

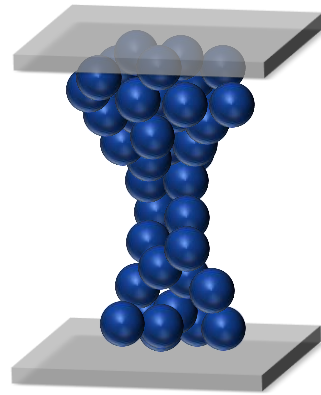
❖ ReRAM device challenges for Neuromorphic circuits

- ◆ Multi-level operations
- ◆ Relaxation and fluctuation mitigation

❖ Roadmap for ReRAM in AI circuits

- ◆ Embedded memory for synaptic weight storage
- ◆ In-memory computing
- ◆ New concepts

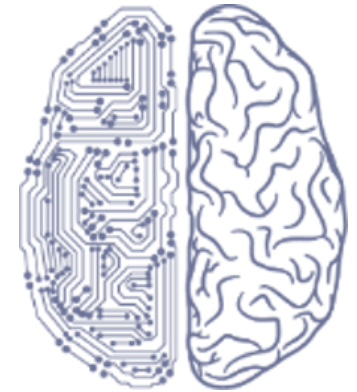
❖ Conclusions



ReRAM characteristics
for AI



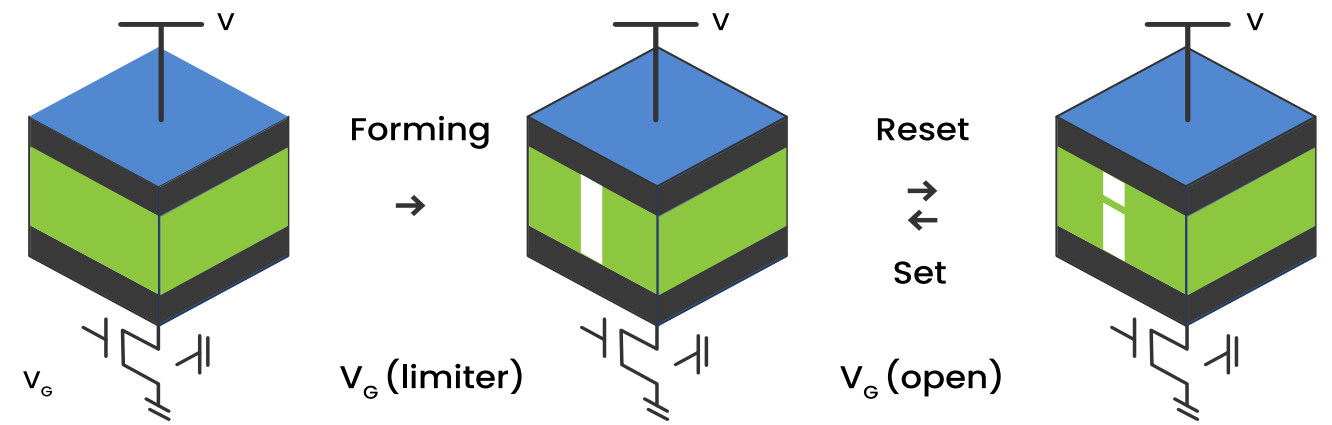
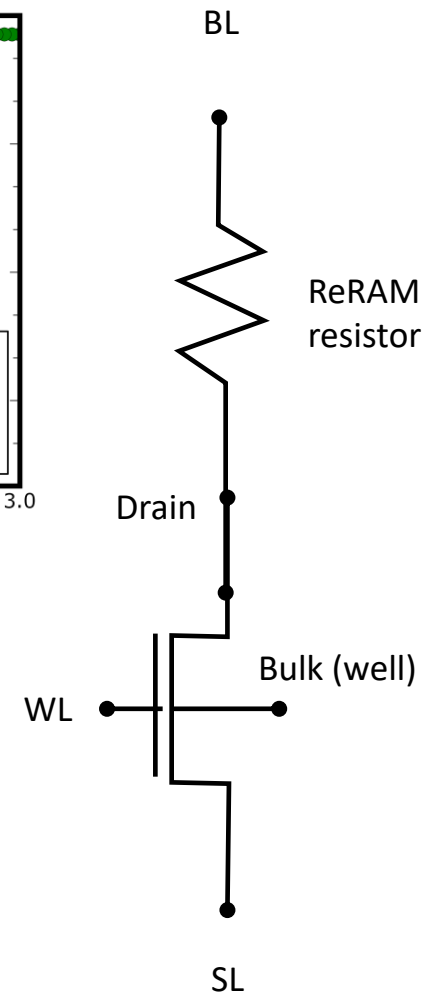
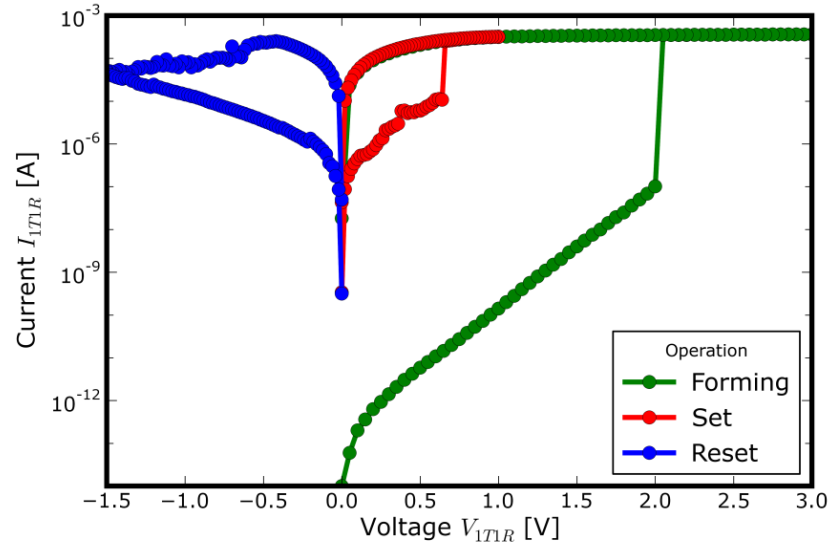
ReRAM device
challenges



Roadmap for ReRAM
in AI circuits

ReRAM Basics

- ❖ Based on the reversible formation of conductive filament (V_O based) in resistive layer
- ❖ Resistive layer is transition metal oxide
- ❖ Scavenging layer to trap oxygen / create vacancies
 - ◆ Active electrode
 - ◆ Sub-stoichiometric oxide
- ❖ 1D concept (scalable)
 - ◆ HRS increases as the inverse of the cell area
 - ◆ Area independence of LRS
- ❖ Bipolar: SET at $V > 0$, RESET at $V < 0$



Low Resistive State (LRS) High Resistive State (HRS)

ReRAM State-of-the-Art

- ❖ Scaling addressed down to 12nm
- ❖ Automotive specs. fulfilled

	Node	Bitcell	Size	Retention	Endurance	Applications
Intel VLSI 2019	22nm	0.0486 μm^2	7,2Mb	10yrs 85°C	10kc	NA
Infineon IMW 2022	28nm	NA	800kB testchip	15yrs 175°C	10-100k	Consumer and industrial
Infineon IMW 2023	28nm	NA	2MB	1000h@175°C for 160°C	125kc	Automotive
TSMC IEDM 2023	12nm	0,02x μm^2	32Mb	10yrs 105°C	10kc	Industrial



Significant Progress in Recent 12 Months

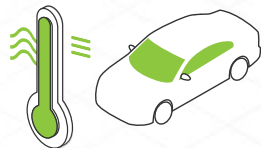


Weebit ReRAM fully qualified at 125°C in SkyWater S130

NOV 2023

FEB 2024

Demonstrated extended automotive performance: 150°C; 100K cycles



APR 2024

Demonstrated ReRAM module prototype on GlobalFoundries 22FDX® wafers

MAY 2024

Partnering with Efabless to broaden user base at SkyWater



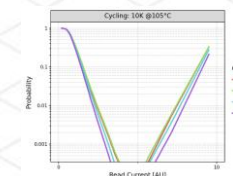
JUL 2024

Taped out module in DB HiTek 130nm BCD process



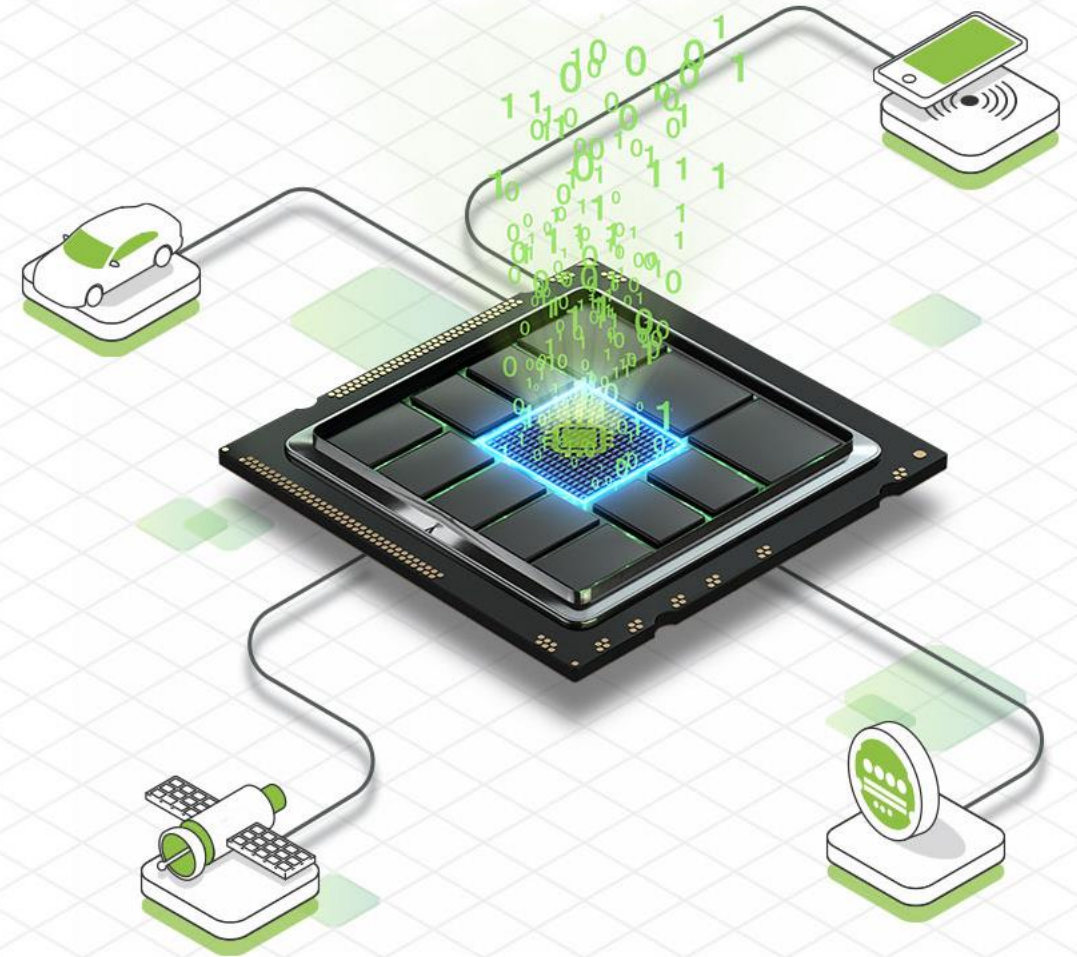
AUG 2024

Initial results of prototype implemented on GlobalFoundries 22FDX® wafers



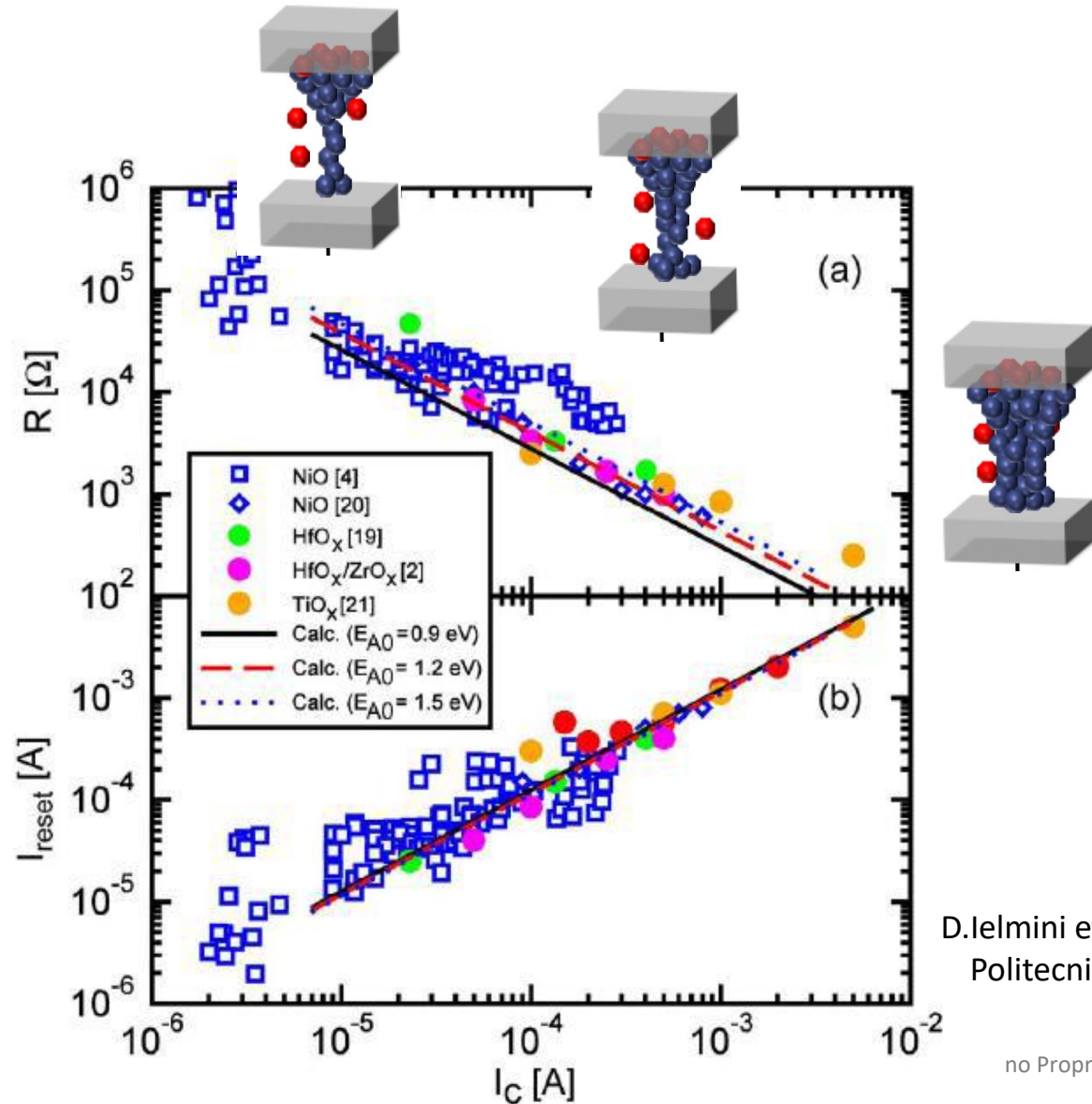
Outline

- ❖ **ReRAM characteristics**
 - ◆ Device basics
 - ◆ State-of-the-art
- ❖ **ReRAM device challenges for Neuromorphic circuits**
 - ◆ Multi-level operations
 - ◆ Relaxation and fluctuation mitigation
- ❖ **Roadmap for ReRAM in AI circuits**
 - ◆ Embedded memory for synaptic weight storage
 - ◆ In-memory computing
 - ◆ New concepts
- ❖ **Conclusions**



Control of Program Resistance – SET

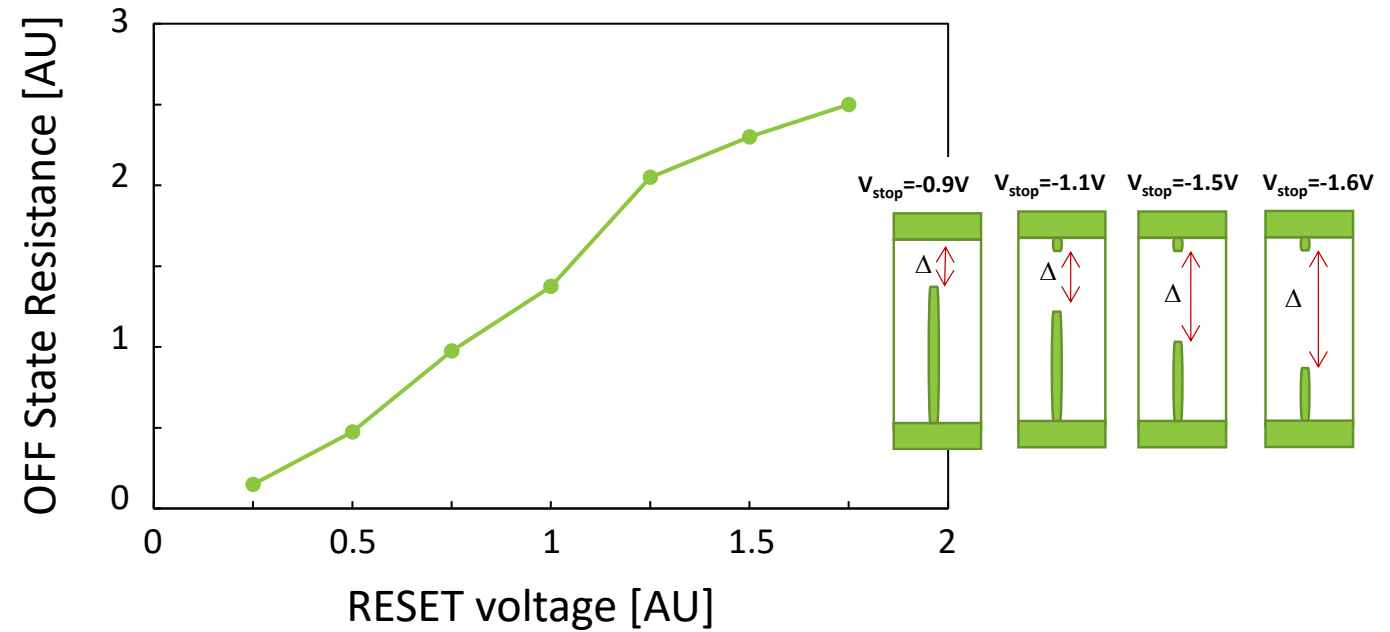
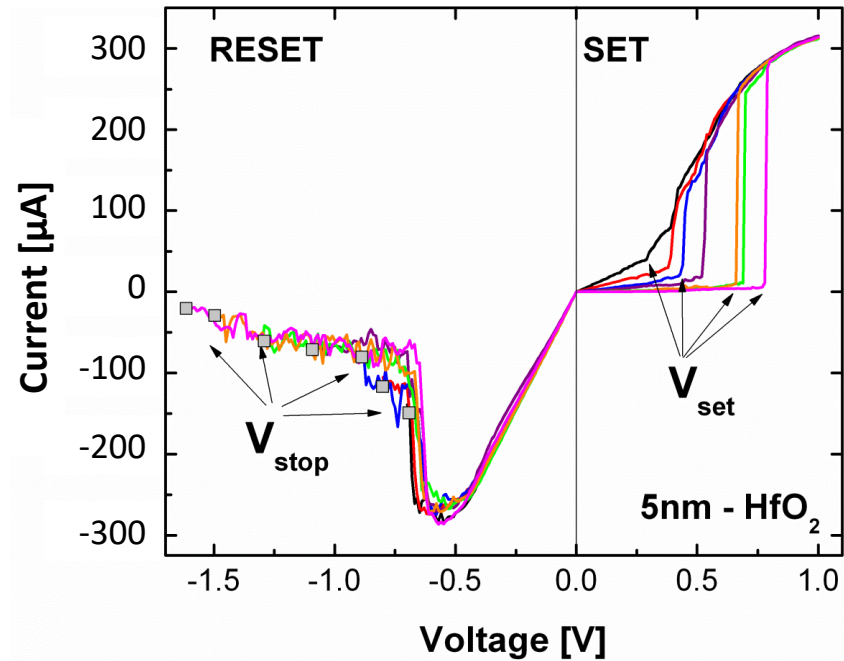
- ❖ ReRAM resistance controlled by filament size
 - ◆ Resistance after SET controlled by programming current
 - ◆ High current → Large filament → low R_{ON}
 - ◆ Low current → Small filament → high R_{ON}



D.Ielmini et al, TED 2011, Politecnico di Milano

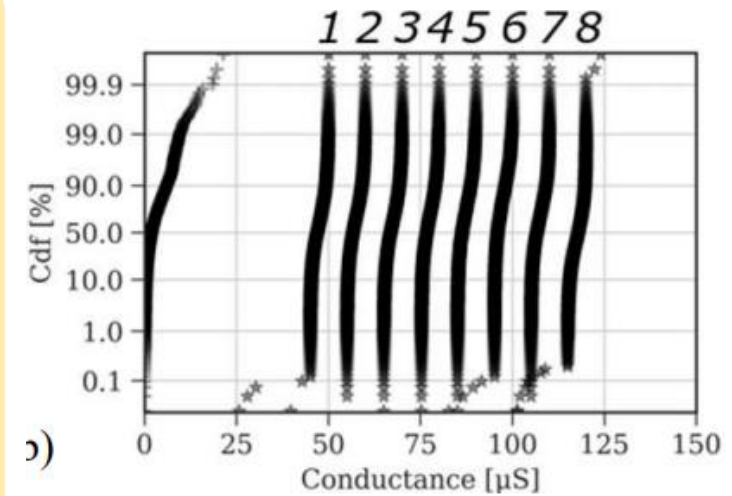
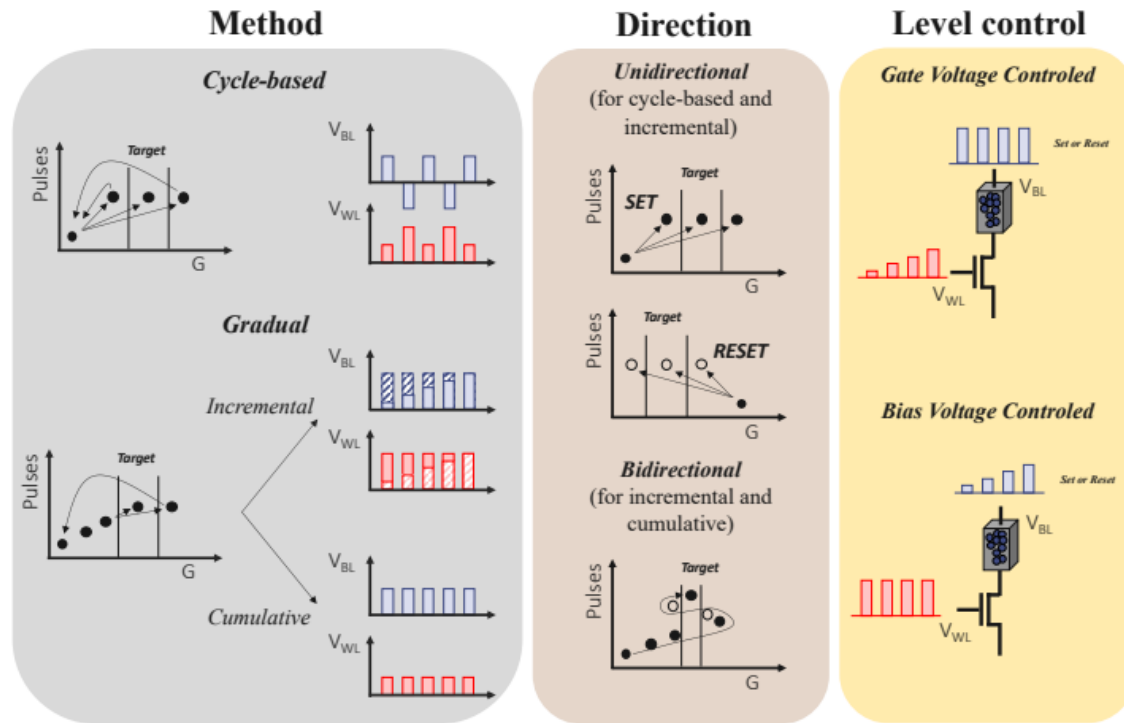
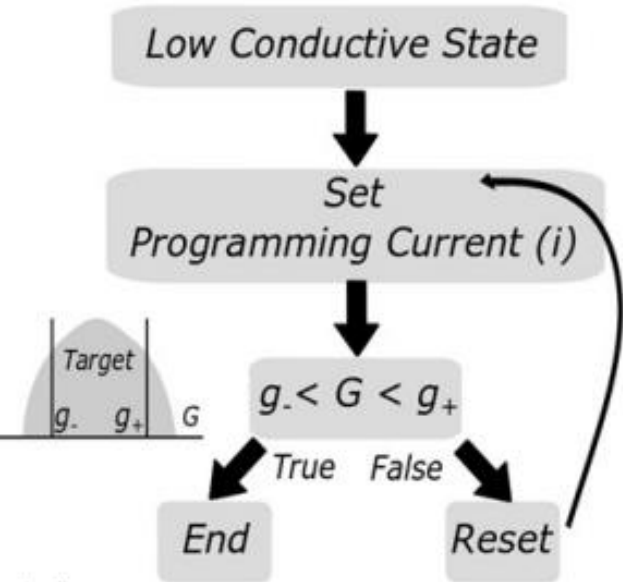
Control of Program Resistance – RESET

T. Diokh et al., Leti, STM, IRPS 2013



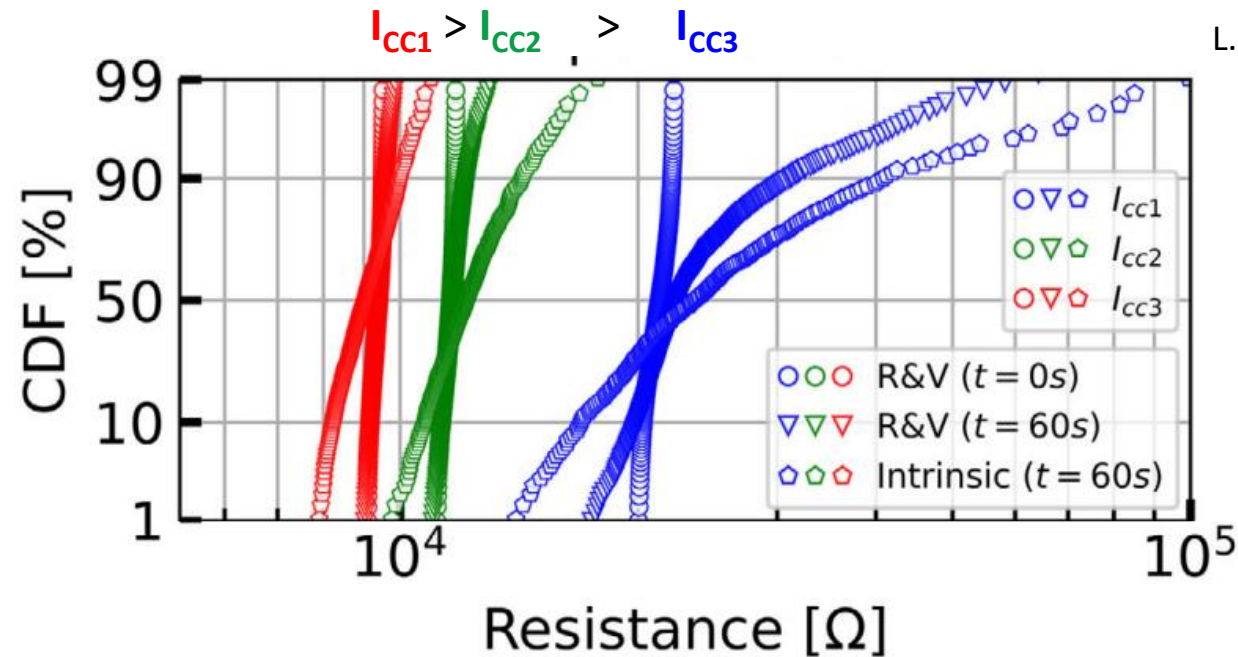
- ❖ ReRAM OFF state resistance R_{HRS} is controlled by RESET voltage
 - ◆ Vreset too low: small margin
 - ◆ Vreset too high: large margin but earlier degradation

Program and Verify for Better Resistance Control



- ❖ Lots of algorithms exist in the literature; playing with current, voltage, sequence... Aiming at resistance distribution control
- ❖ When the resistance after programming is not in the expected range, the cell is reprogrammed

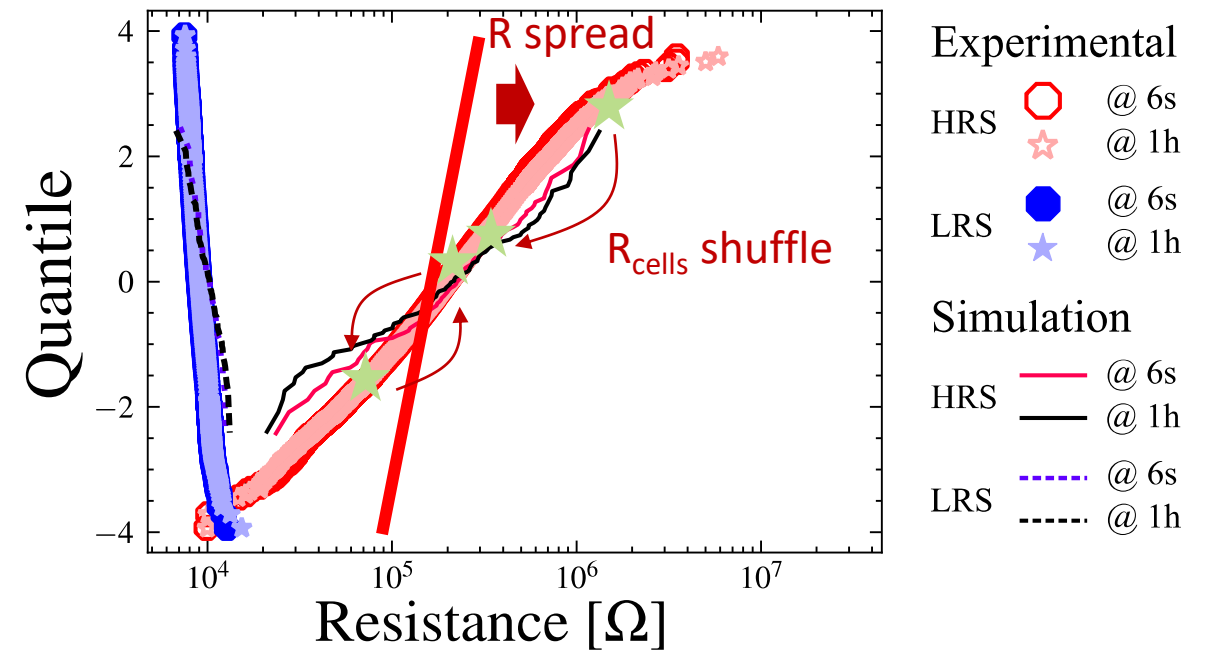
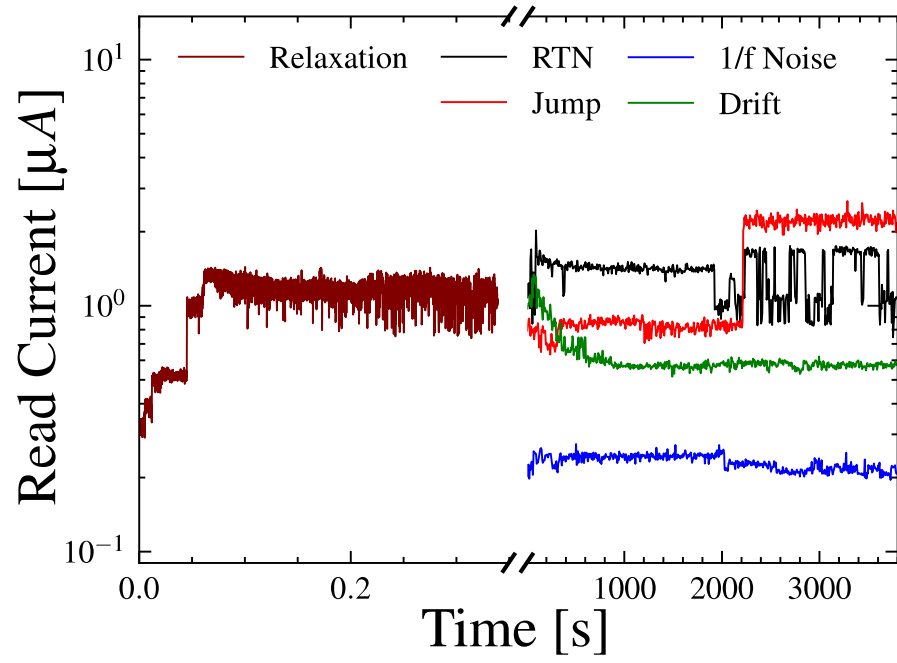
ReRAM Resistance Spread Post Programming



L. Reganaz, PSSA 2022, Leti

- ◆ Program and Verify algorithm allows control of the resistance value...
- ◆ ... But distribution spread after programming \rightarrow Physical origin?

ReRAM Resistance Spread Post Programming

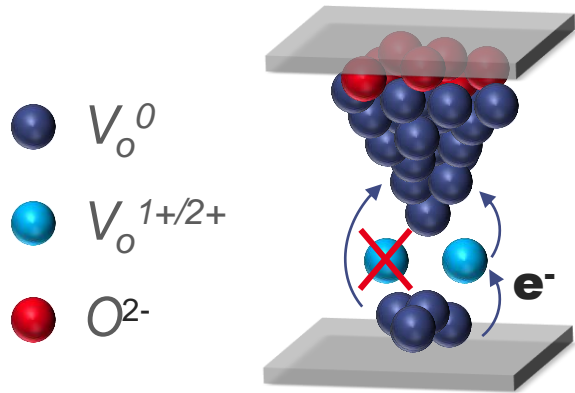


- ❖ After ReRAM switching, single cell level resistance variations are measured; 2 phases can be distinguished:
 - ◆ Relaxation <1s range
 - ◆ Fluctuations (anytime)
- ❖ After relaxation (>seconds), single cell fluctuations do not affect overall distribution
- ❖ These measured ReRAM resistance variations are analyzed by means of KMC simulations; physics discussed in the next slides

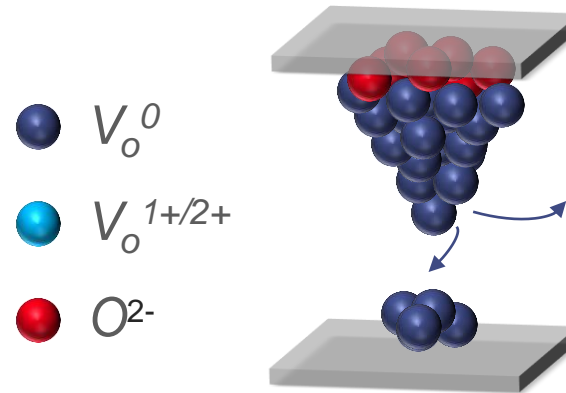
ReRAM Resistance Variations – Fluctuations (anytime)

L. Reganaz, IRPS 2023
Leti – Weebit

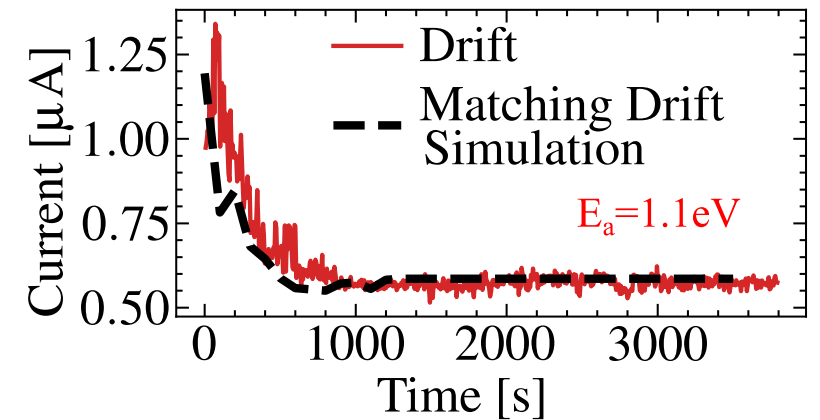
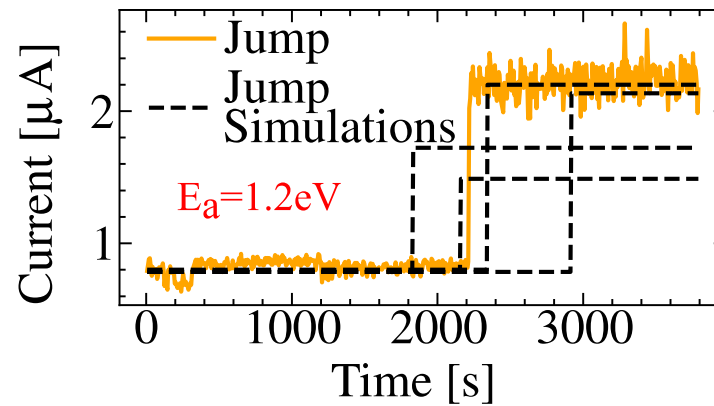
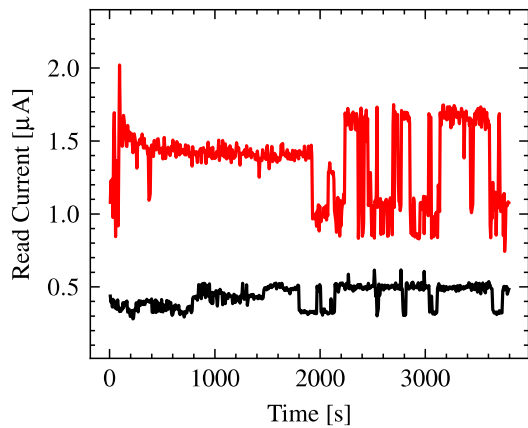
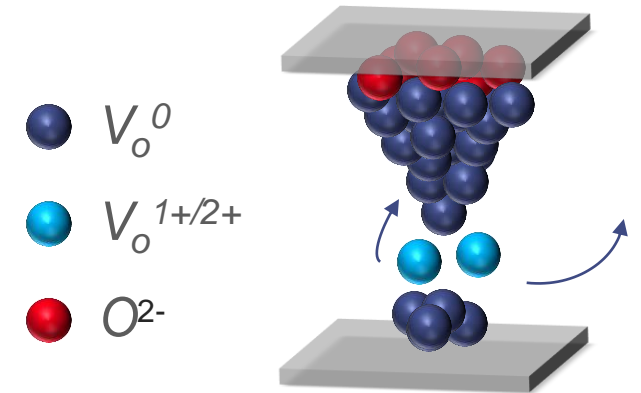
V_o position and charge \rightarrow RTN



V_o migration & ionization

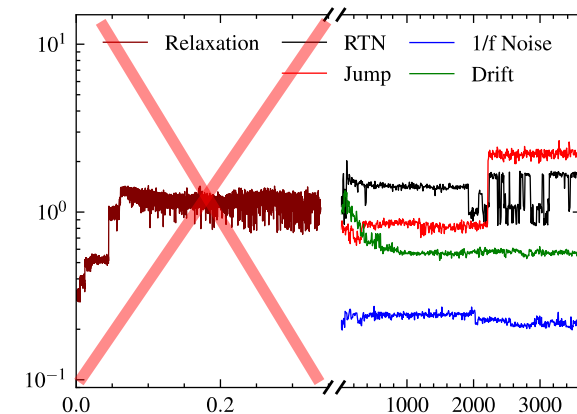
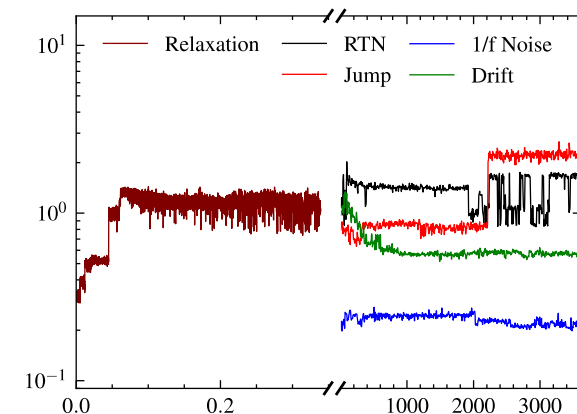
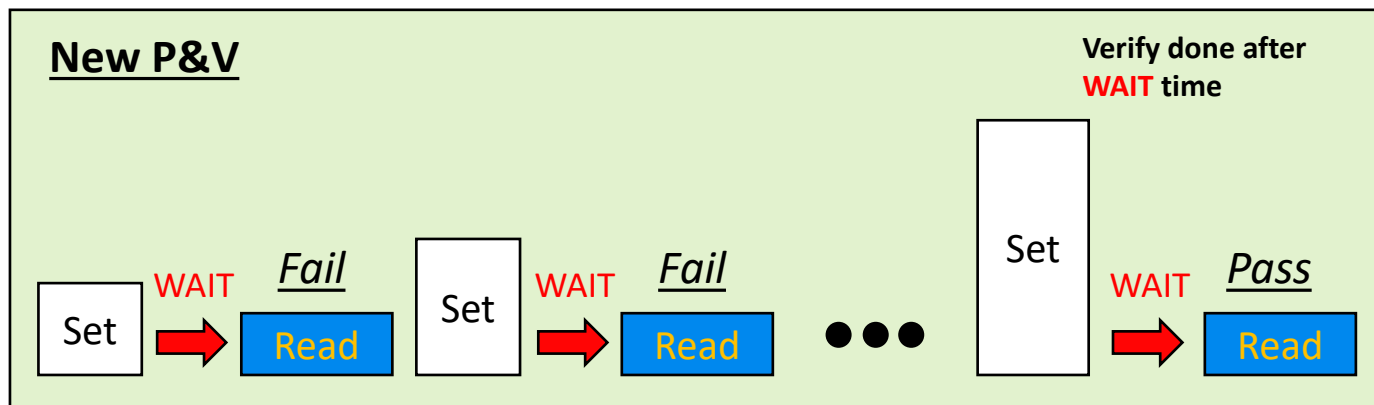
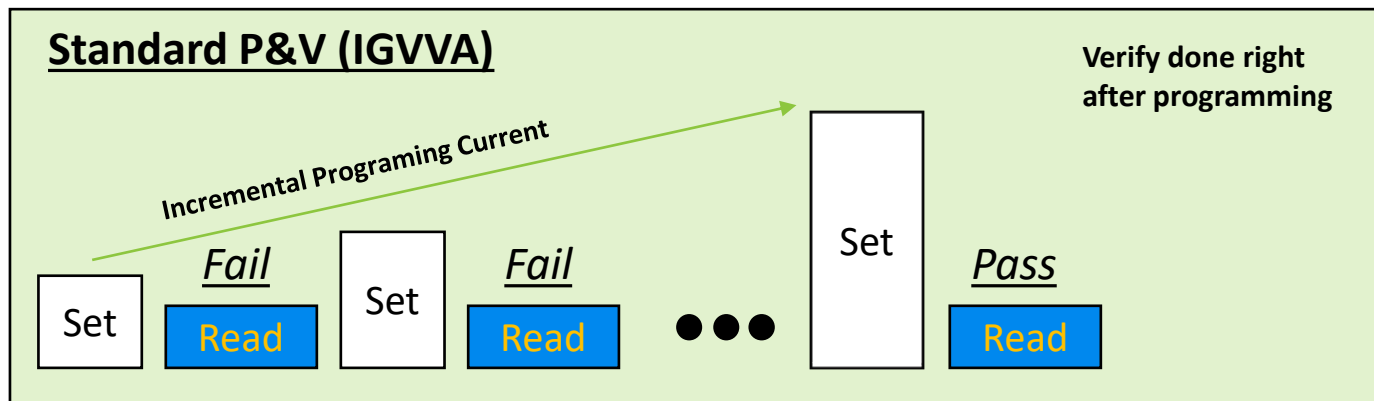


V_o migration & neutralization



Algorithms with Delay to Improve ML – Concept

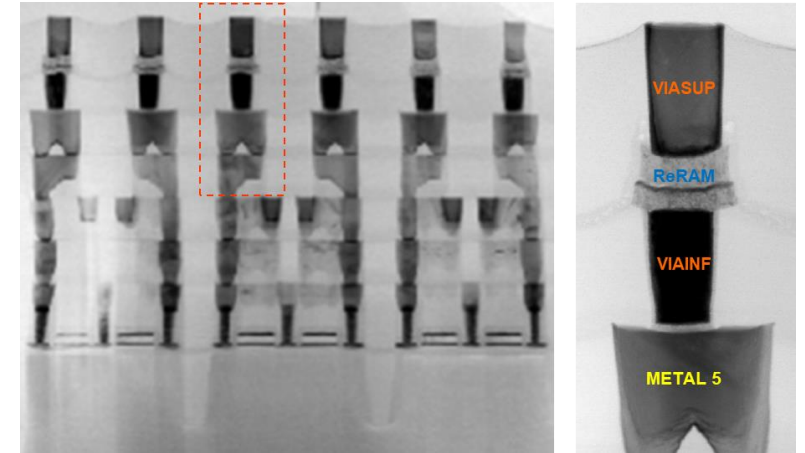
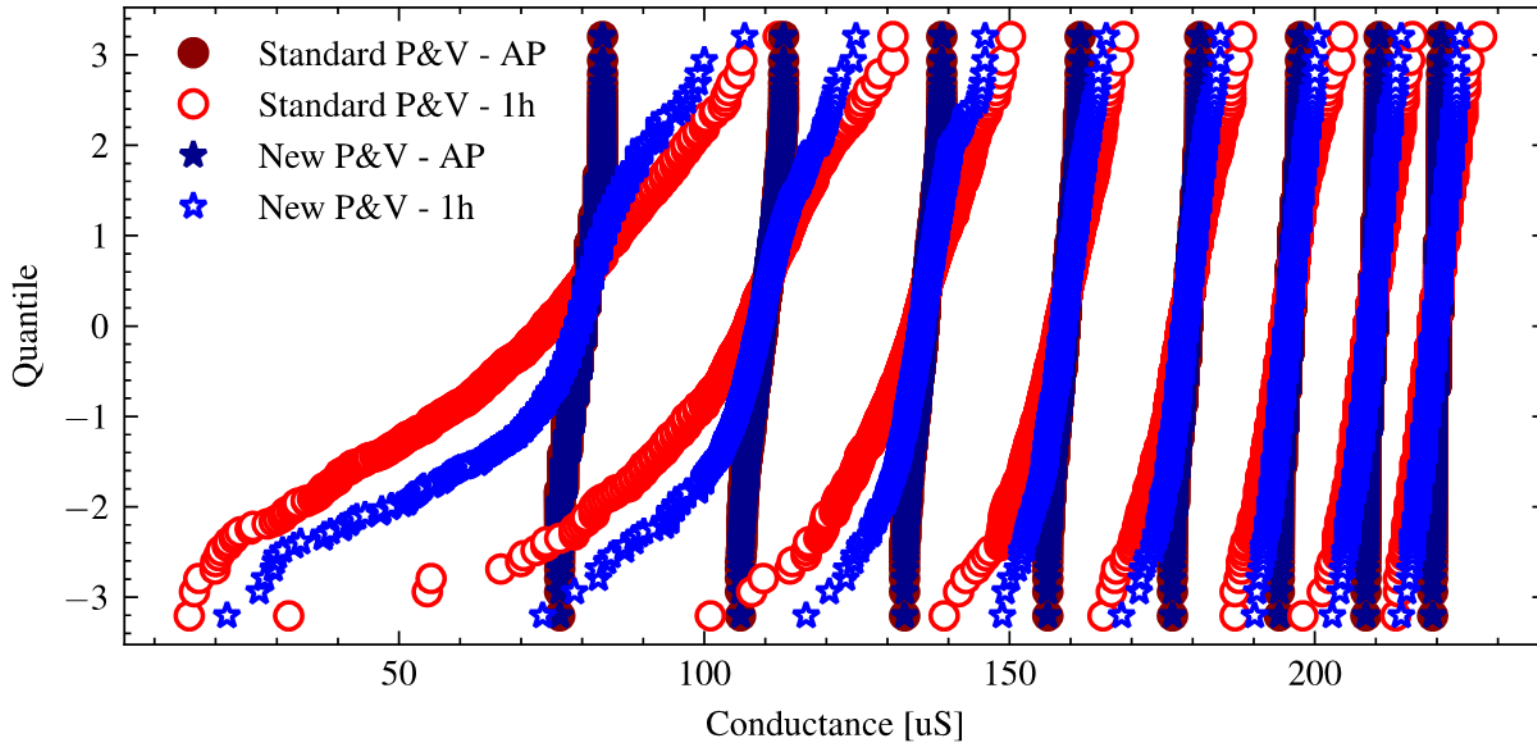
L. Reganaz, SSDM 2023
Leti – Weebit



Part of variability
that can be
removed

Algorithms with Delay to Improve ML – Experimental Results

L. Reganaz, SSDM 2023
Leti – Weebit

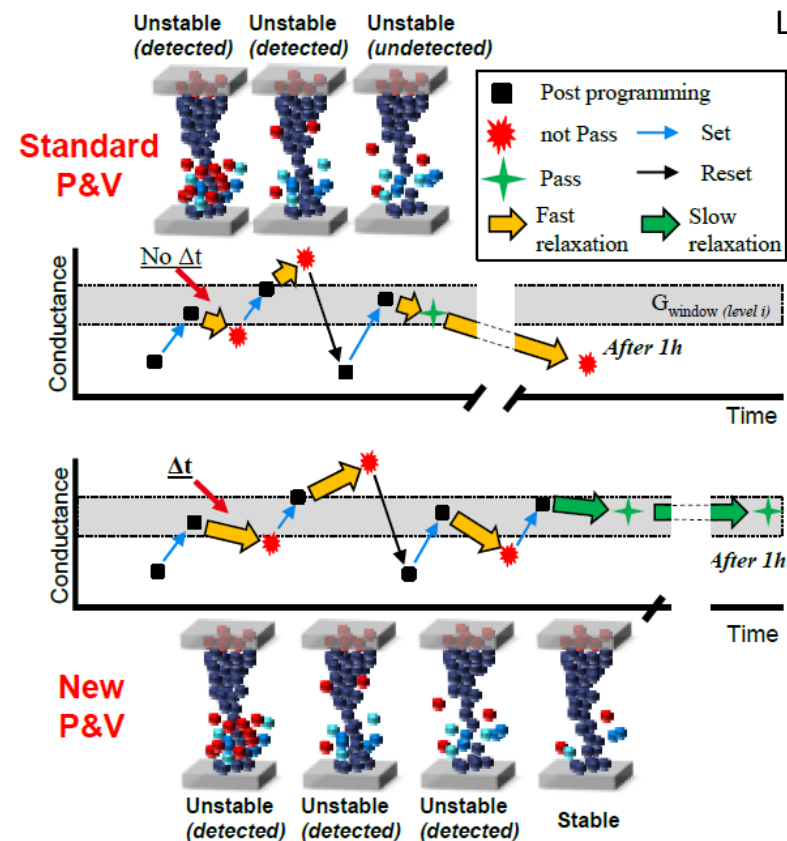
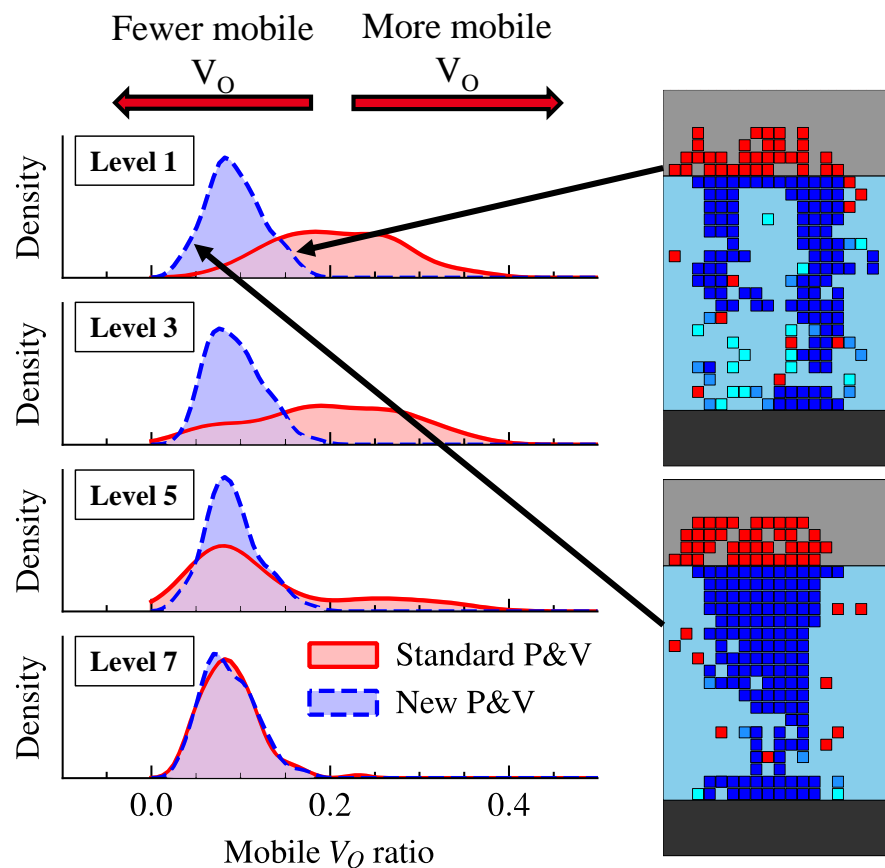


○ 16kb 1T1R array in 28nm FDSOI BEOL

- ◆ By adding 1min verification delay to standard P&V, we can achieve better MLC stability
- ◆ What's the physical explanation?

Conductance Instability from Highly Mobile V_O

L. Reganaz, SSDM 2023
Leti – Weebit



- ❖ New P&V also allows reduction of the number of high mobile V_O because due to the waiting time we can “select” only the conductive filament with fewer high mobile V_O

Outline

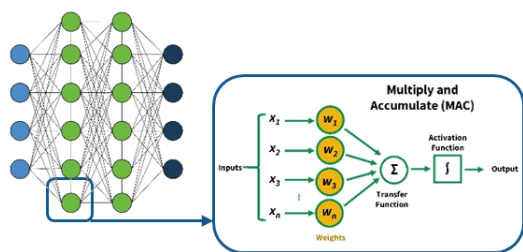
- ❖ **ReRAM characteristics**
 - ◆ Device basics
 - ◆ State-of-the-art
- ❖ **ReRAM device challenges for Neuromorphic circuits**
 - ◆ Multi-level operations
 - ◆ Relaxation and fluctuation mitigation
- ❖ **Roadmap for ReRAM in AI circuits**
 - ◆ Embedded memory for synaptic weight storage
 - ◆ In-memory computing
 - ◆ New concepts
- ❖ **Conclusions**



ReRAM for Neuromorphic

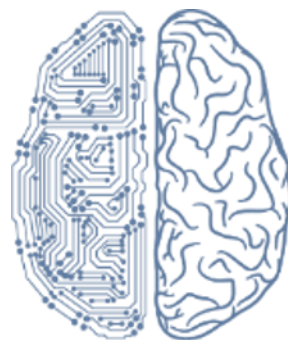
ReRAM perfectly fits in brain inspired systems, in a timely manner:

Embedded, Edge AI



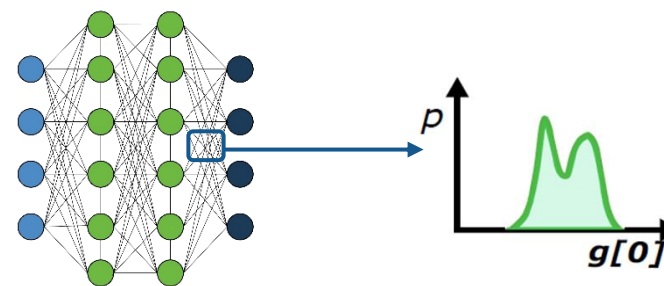
ReRAM used as eNVM to store synaptic weights of the NN

In-Memory Compute:
AI and ML



Computing is done within the ReRAM itself

New Concepts



New concepts take advantage of ReRAM specificities

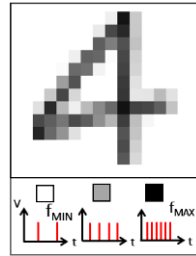
Fully Integrated Spiking Neural Network using Weebit ReRAM as a Synaptic Device

Proof-of-concept SNN performing recognition of MNIST digits

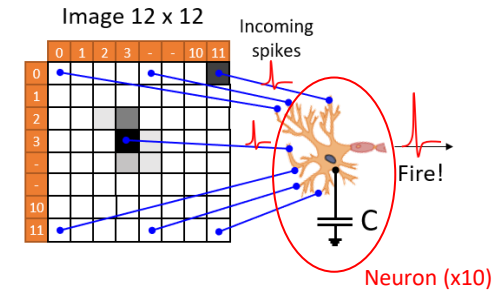
(Off-line learning)



Digit drawn on tablet application



256 grey levels Converted in spike frequency



Fully-connected, single-layer perceptron

Collaboration with



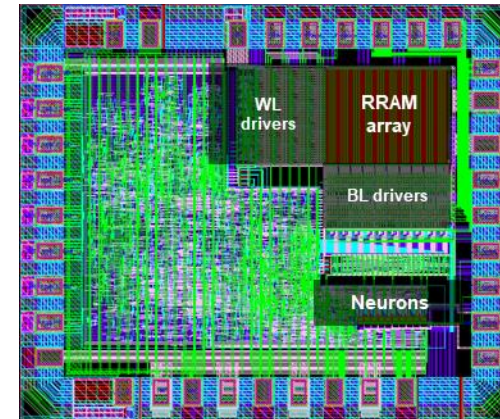
A. Regev, AICAS 2020, Weebit, Leti

❖ Synapse design

- ◆ Synapse emulated with 8 ReRAM (4x excitatory / 4x inhibitory)

❖ System performance

- ◆ 160 spikes, on average, to recognize an image
- ◆ Measured energy consumption: 180 pJ/syn. ev. (3 pJ at synaptic level)
- ◆ 82% of accuracy obtained – to be compared with the maximum theoretical accuracy 88% (topology-limited)

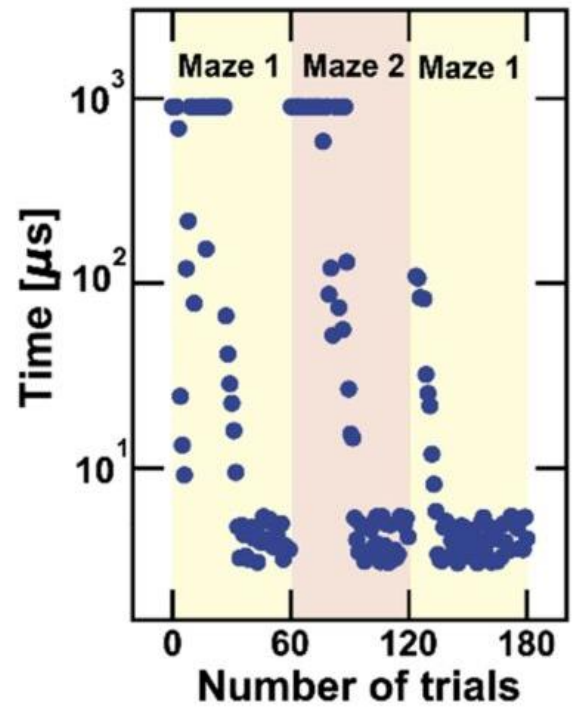
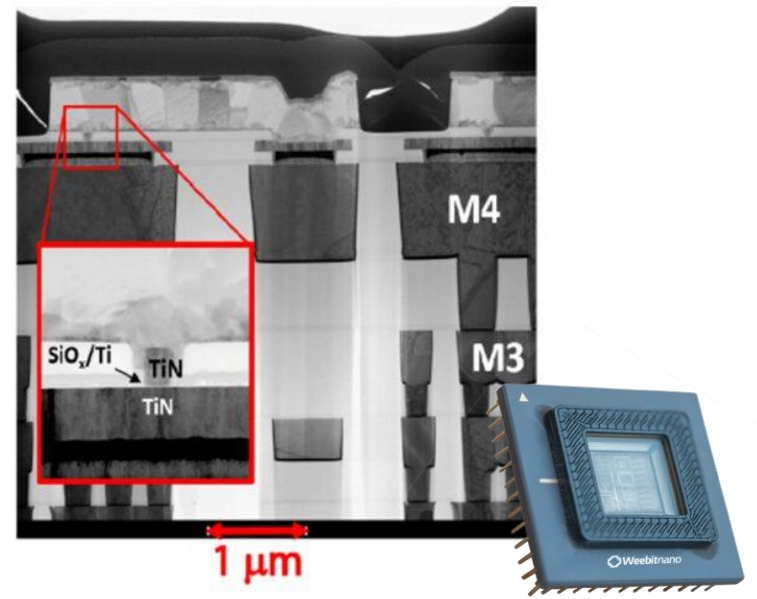



System details

CMOS process	130 nm
Cu interconnects	5
RRAM number	11,5k
RRAM configuration	1T-1R
Chip size	1.8 mm ²

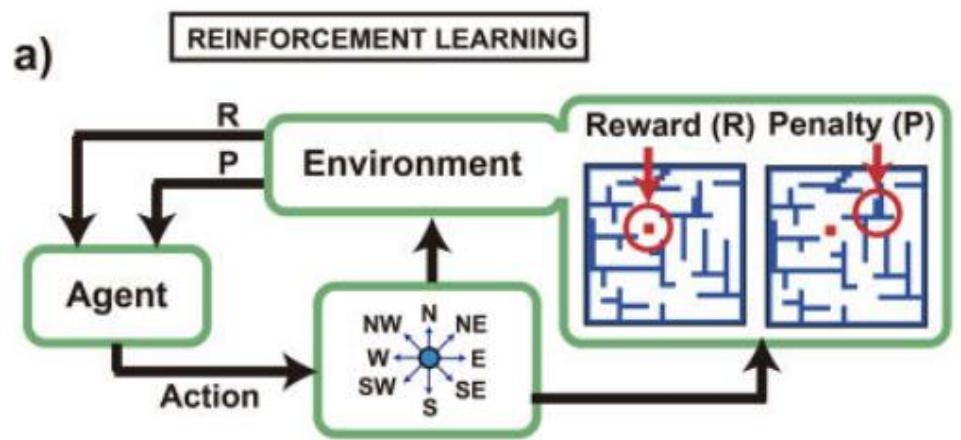
- ❖ **Bio inspired system with hardware Plasticity**
 - ◆ Why is it important? Allows faster adaptation to changing environment, adjusting its state based on specific inputs (as in the case of biological synapses)
 - ◆ How to achieve it? ReRAM conductance can be changed with a few electrical parameters

- ❖ **Weebit's ReRAM devices were used to:**
 - ◆ Store information and adjust the strength of connections between neurons
 - ◆ Tested in the autonomous exploration of an evolutionary two-dimensional dynamic maze



Collaboration with  **POLITECNICO MILANO 1863**

S. Bianchi, Nature, 2023, Politecnico di Milano [18]

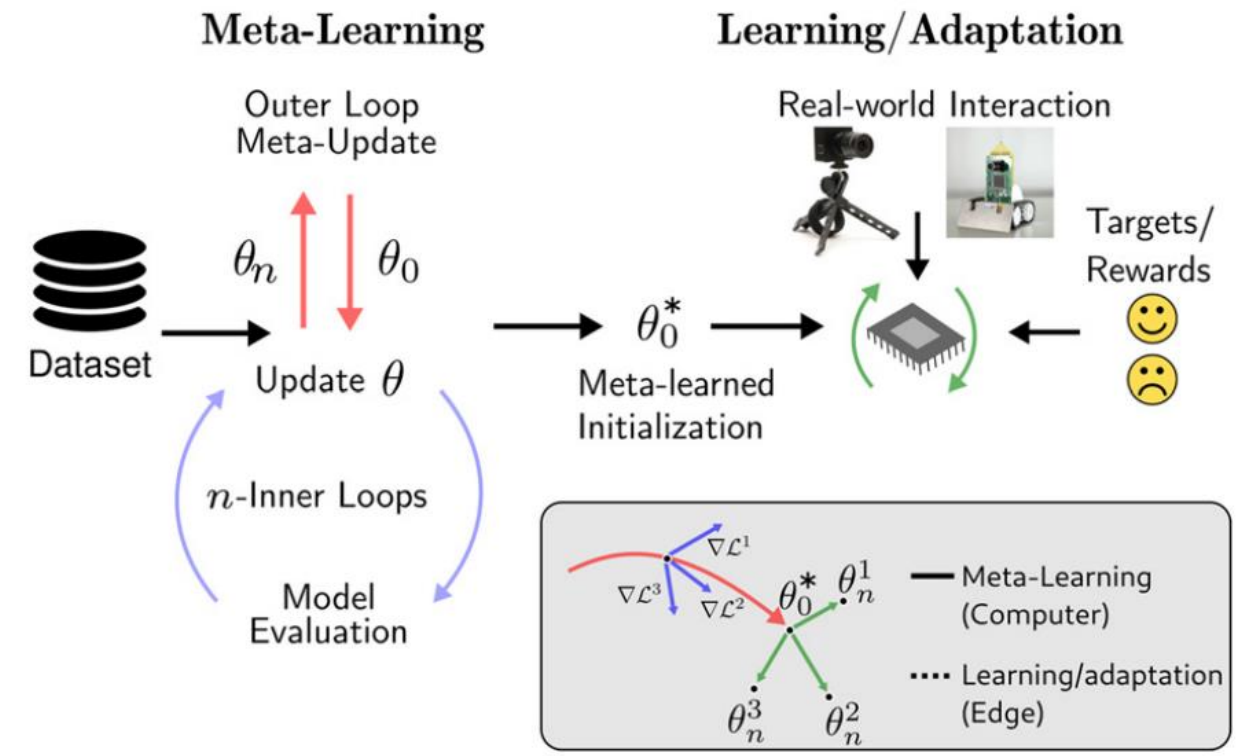


Model-Agnostic-Meta-Learning


Standalone system capable of learning, adapting and acting locally at the edge

- ❖ OFF-chip:
 - ◆ MAML model trained on server (over than 50k iterations on GPU)
 - ◆ Output off-chip:
 - Optimized net parameters
- ❖ ON-chip:
 - ◆ Standalone ReRAM device able to quickly learn new tasks
 - Compatible with ReRAM endurance
 - ◆ Confidentiality assured between customers and between customer and company

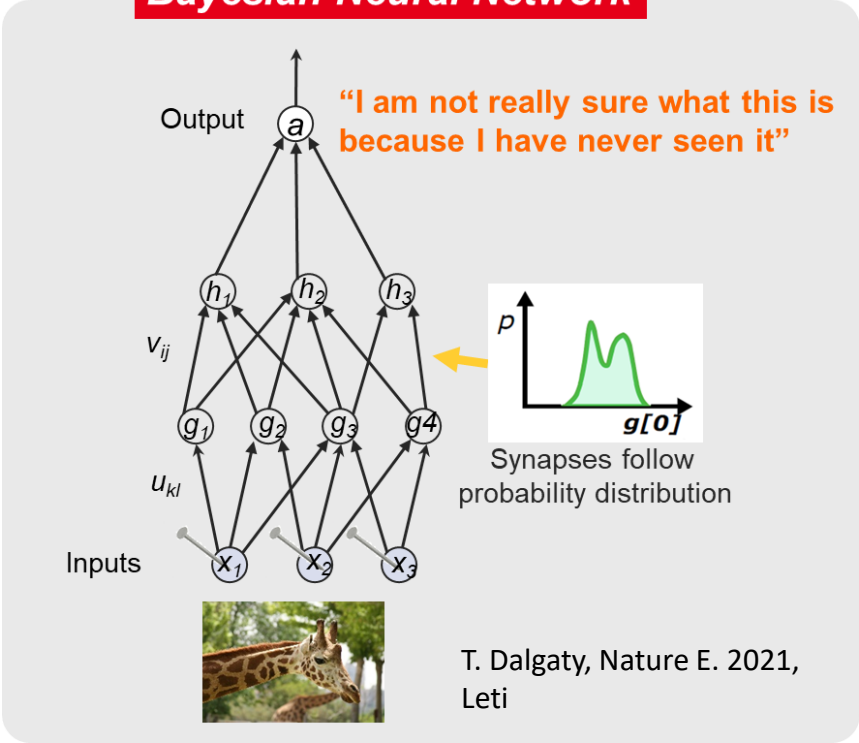
K. Stewart, NCE 2022, UC Irvine [20]



How to Take Advantage of ReRAM Specificities

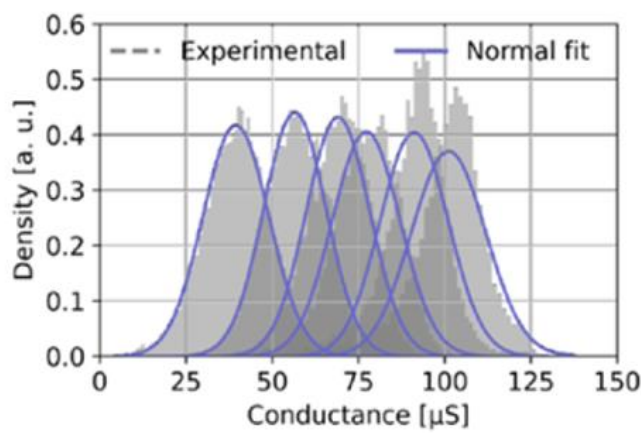
Collaboration with 

Bayesian Neural Network



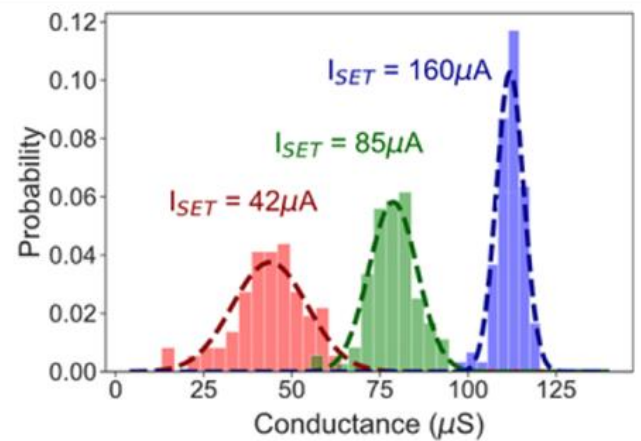
ReRAM’s intrinsic variability naturally produces Bayesian Neural Network

Device-to-device (D2D)



E. Esmanhotto *et al.*, Adv. Intell. Syst., 2022

Cycle-to-cycle (C2C)



T. Dalgaty *et al.*, Adv. Intell. Syst., 2021

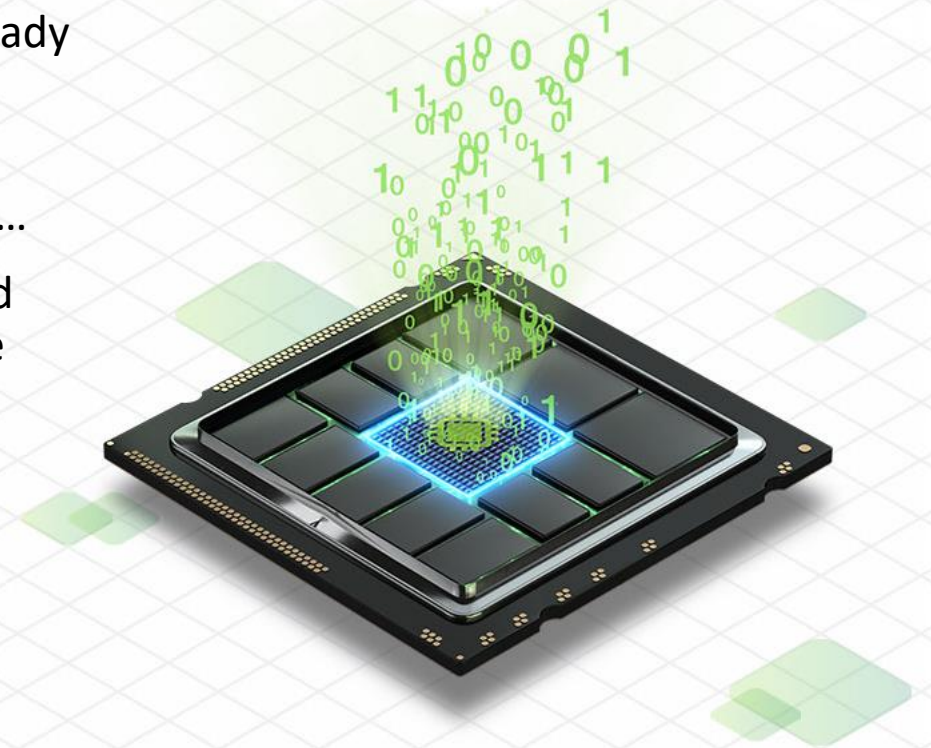
Some models require randomness to compute based on probabilities



Stop fighting non-idealities

Conclusions

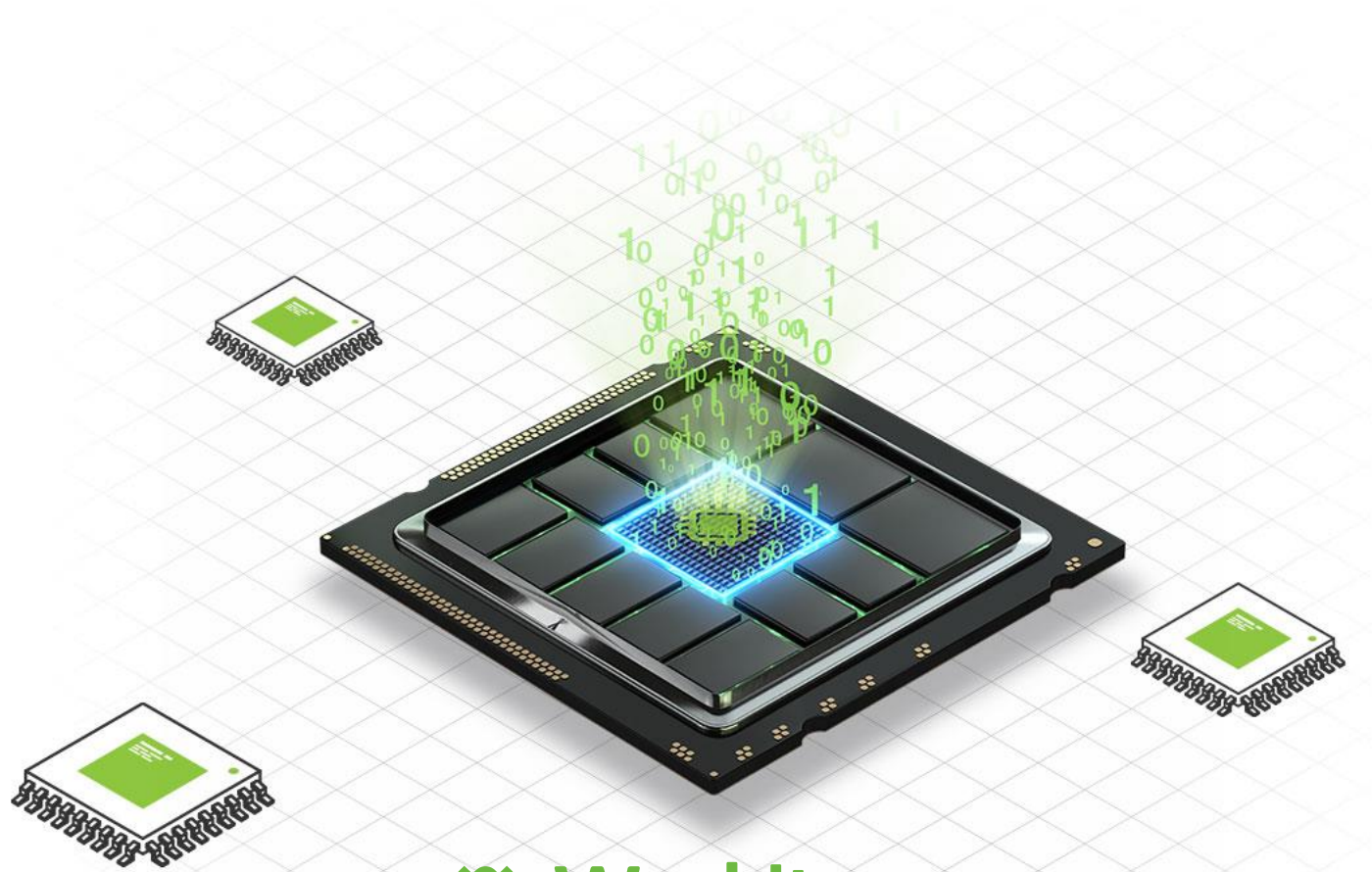
- ❖ ReRAM has progressed from an emerging device to a market ready technology, replacing embedded flash for advanced nodes
- ❖ ReRAM characteristics are perfectly suited to AI applications, thanks to high speed, cost efficiency, small footprint, scalability...
- ❖ ReRAM will enter the AI roadmap step-by-step (1) as embedded memory, (2) as computing device in IMC circuits, (3) as the core device for new concepts
 - ◆ For synaptic weight storage and IMC applications, accuracy can be improved thanks to technology optimization and adapted program algorithms to tackle ReRAM relaxation and fluctuations...
 - ◆ Another approach is to take advantage of ReRAM specificities (like stochastic character) to imagine new brain inspired concepts



Thank You!



POLITECNICO
MILANO 1863



 **Weebitnano**
THE NEXT NVM IS HERE

Thank You!



POLITECNICO
MILANO 1863

