ANDANTE New memory Technologies for AI applications

ANDANTE

Introduction

- Leverage embedded non-volatile memories (eNVM) to develop new low-power architectures for edge AI applications.
- Explore new memory technology options that will allow for In Memory Computing (IMC) . The core operations in Artificial Neural Networks (ANNs) are matrix vector multiplications (MVMs). Minimizing the data movement between the compute and memory blocks (non-Von Neuman computing) has had great success towards energy and performance optimization. This has primarily been achieved through IMC techniques.

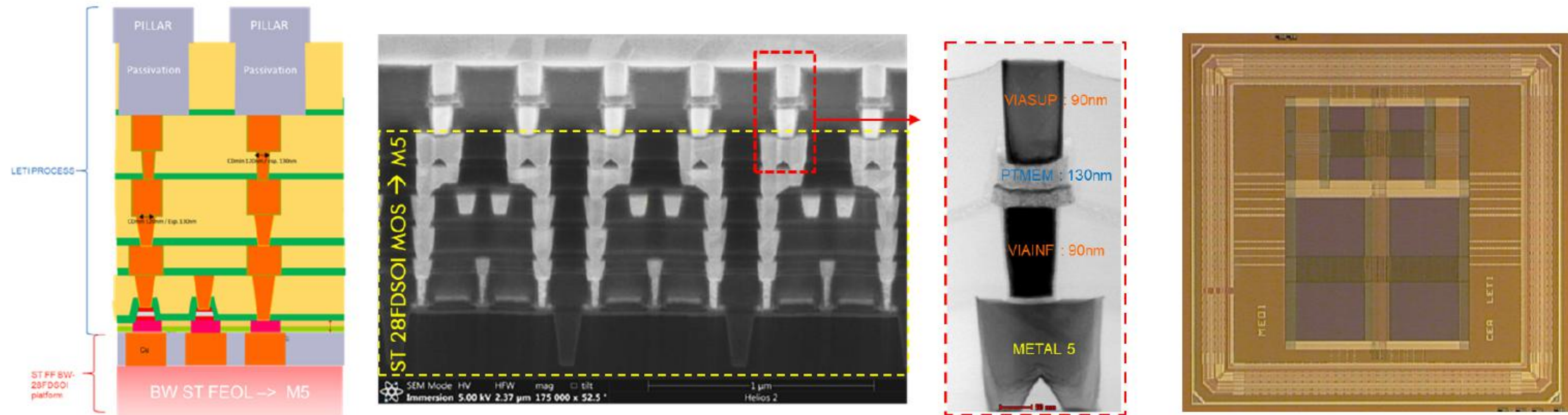
Main Goals

1. Solutions – including process development kits - to be employed within ANDANTE
2. Serve pathfinding solutions for increased on-die embedded memory for neural network acceleration upon ultra-low power consumption.

This poster focuses on enabling and demonstrating the scalability of new memory technologies to larger scale AI applications for ANN. Three memory technologies are explored and presented below.

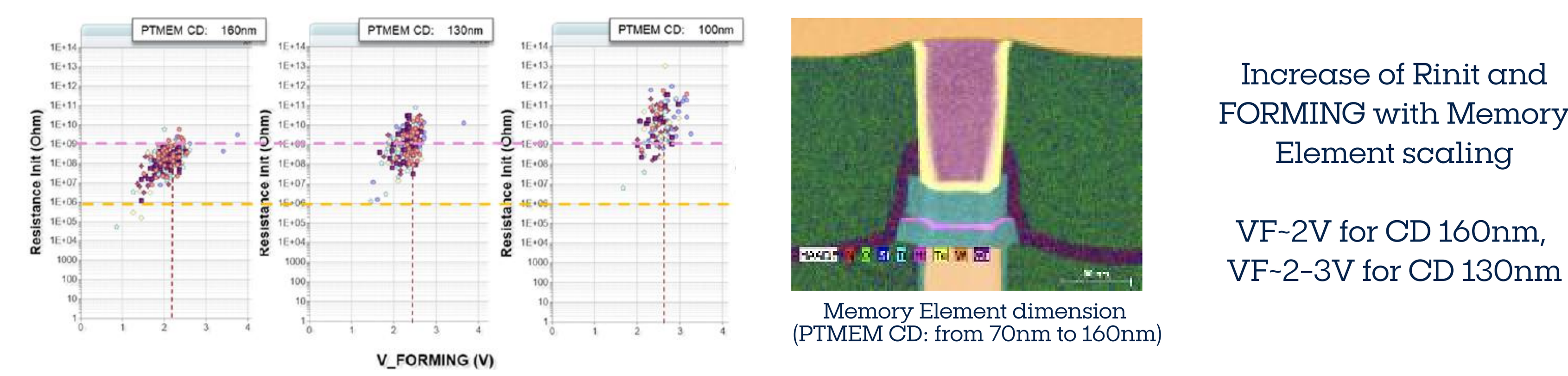
OXide-based resistive RAM (OXRAM) for AI applications

Objectives:
- enable Large Scale integration of RRAM technologies for applications based on ANN
- demonstrate the scalability of 1T1R OXRAM by decreasing the cell size
- increase the number of bits per cell by benchmarking OxRAM for Multivalued resistive RAM (2-4 bit/cell)
- investigate 1S1R option (with Back-end selector for replacing the access transistor)



Bitcell size: 0.163µm2

OxRAM Macro 2.5Mb ~1mm^2



Memory Element dimension (PTMEM CD: from 70nm to 160nm)

Increase of Rinit and FORMING with Memory Element scaling

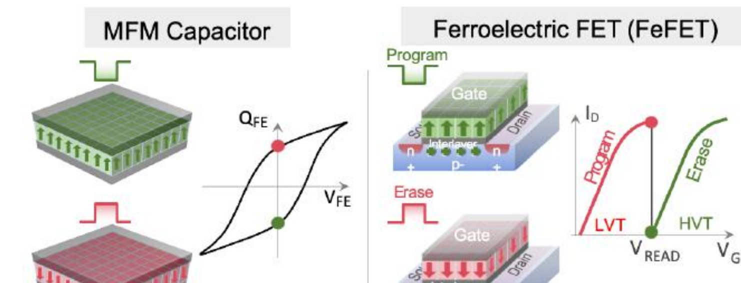
VF~2V for CD 160nm, VF~2-3V for CD 130nm

Ferroelectric FET (FeFET) in 22nm FDSOI and 28nm

Goal : demonstrate of FeFET based an analog in-memory compute (AiMC) array

Objectives :
- adapt the FeFET bitcell for optimized AiMC
- validate cell functionality and scaling route towards 22nm FDSOI
- understand requirements for AiMC applications

Status :
- FeFET based target AiMC operation optimized
- Wafer runs with dedicated AiMC arrays currently ongoing

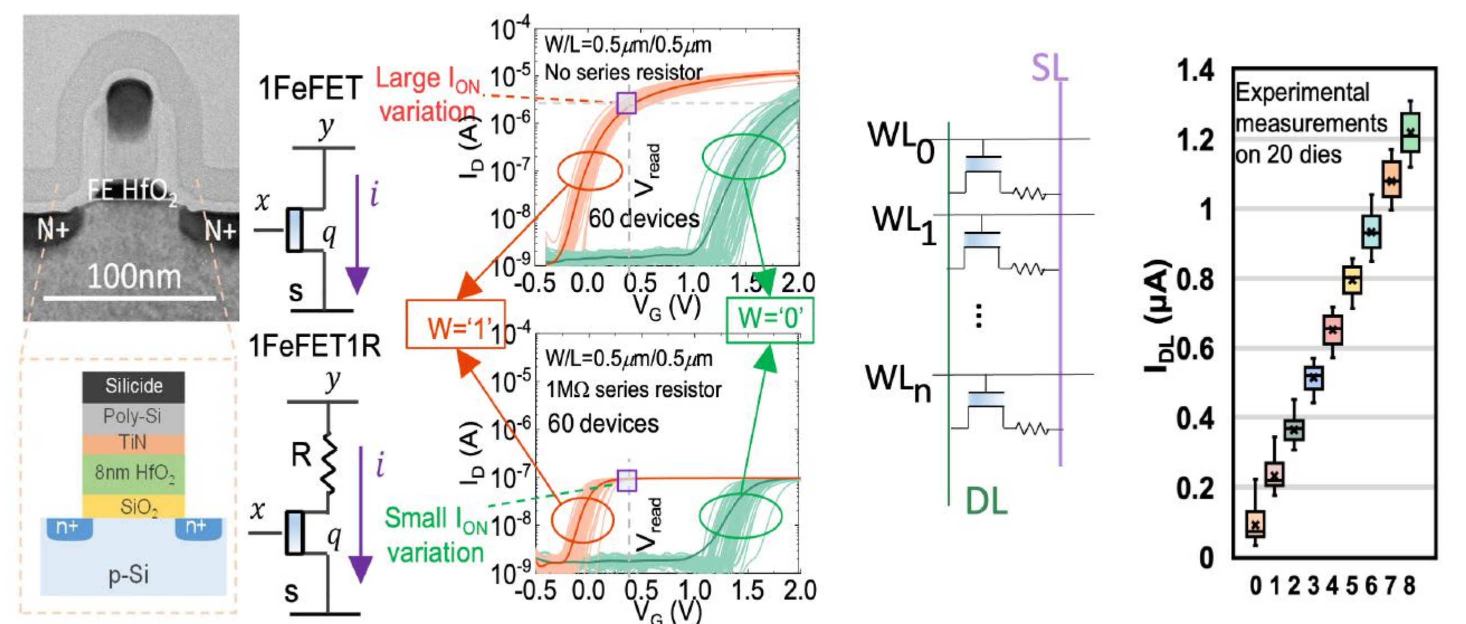


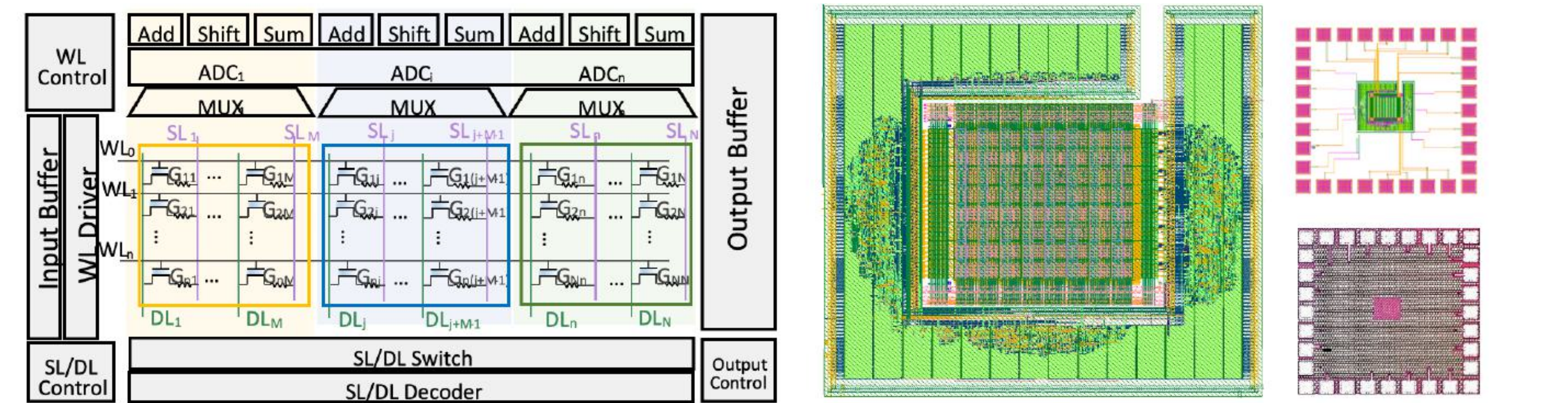
Polarization states in FeFET can be used to store information.

	FeFET
Cell structure	1T
R_{ON}/R_{OFF}	~10⁴
Read scheme	Non-dest.
Write voltage	<4 V
Write energy	~10fJ
Write speed	~10ns
Write endurance	>10⁶

Ferroelectric FET (FeFET) for analog in-memory compute (FhG)

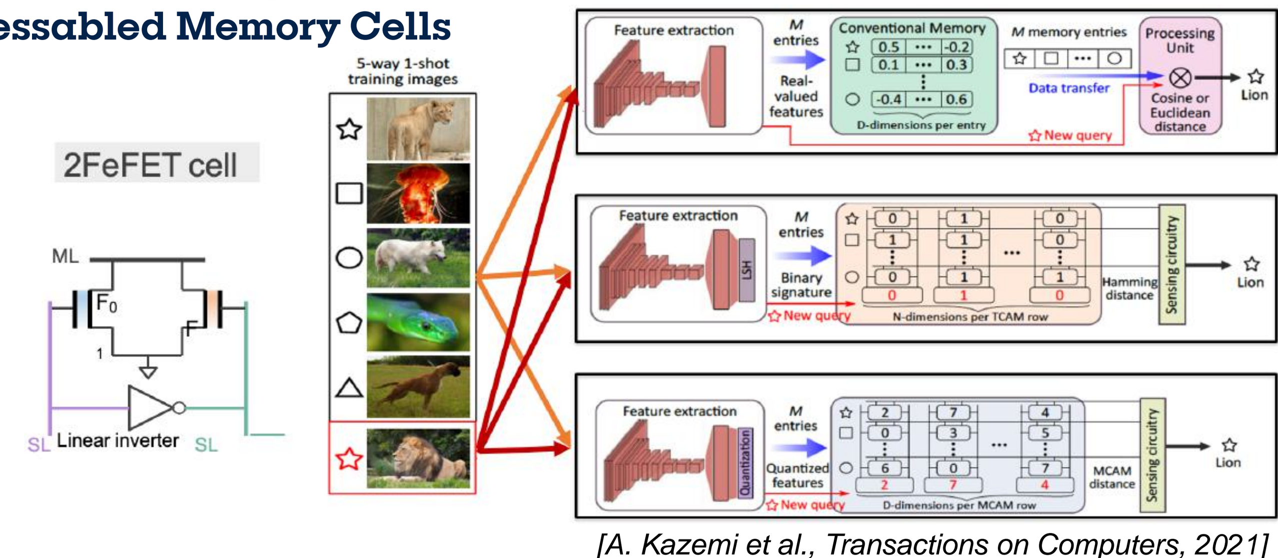
1F1R bitcell concept for accurate accumulation





[S. De et al., TED, 2022] [S. De et al., EDL, 2022]

Few Shot Learning with Multi-level Content Adressabled Memory Cells

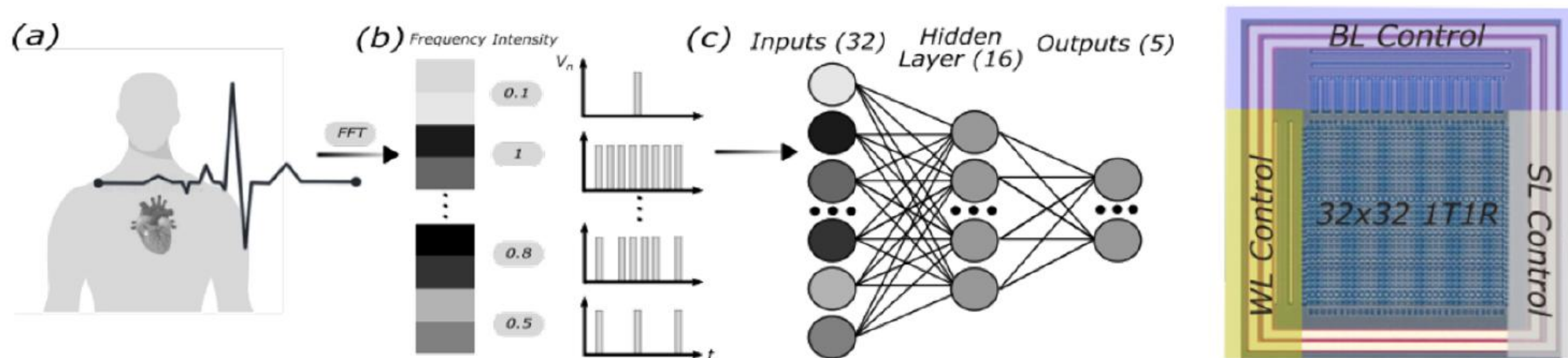


[A. Kazemi et al., Transactions on Computers, 2021]

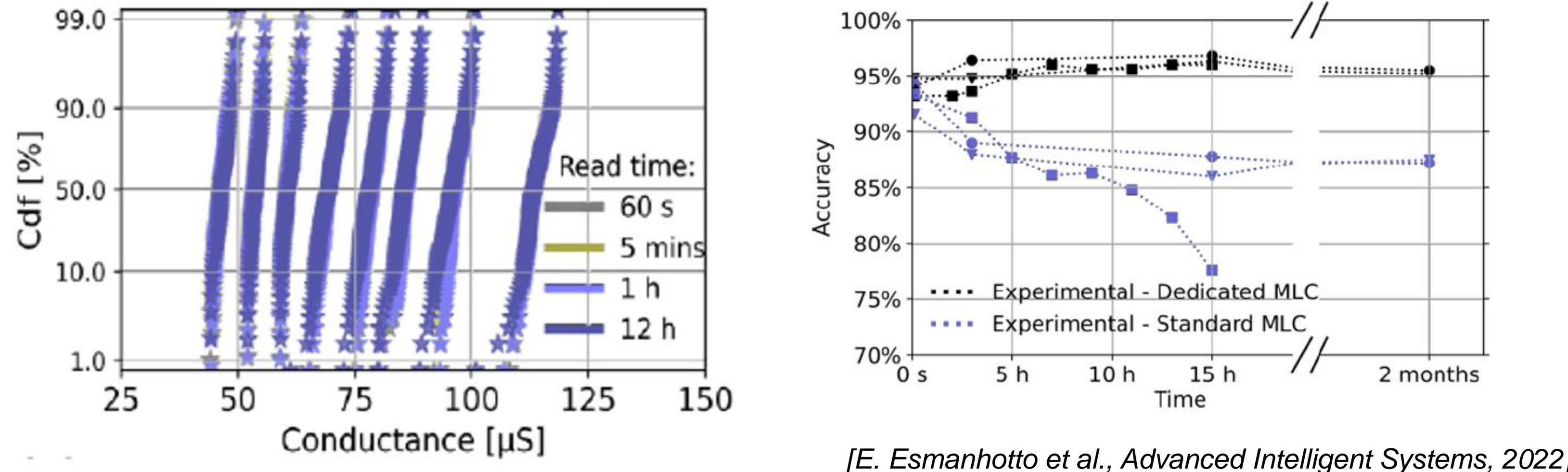
OxRAM for analog in-memory compute

Non-volatile multi-level analog programming and experimental demonstration of on-chip In-Memory computing

In contrast to SRAM or DRAM, OxRAM can be programmed to intermediate states between their lowest and highest resistance values, allowing memorizing the synaptic weights of a neural network. A two-layer perceptron inference based on an RRAM crossbar array has been demonstrated with a new multilevel programming algorithm.
Next Figures show the implementation of an artificial neural networks (ANNs) with analog weights and binary stochastic neurons: The AI application is to identify the heart arrhythmia from ECG recordings



Next Figures show the results of "smart" dedicated multilevel programming technique tested on a complete neural network enabling stable accuracy over time.



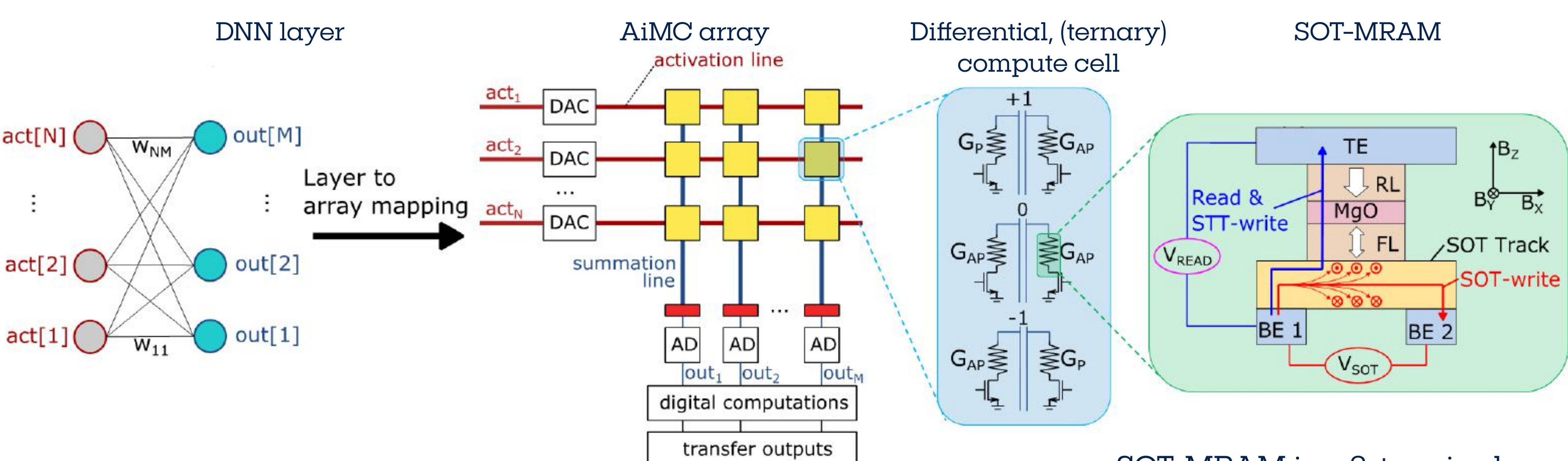
[E. Esmanhotto et al., Advanced Intelligent Systems, 2022]

Spin-Orbit-Torque Magnetic RAM (SOT-MRAM) for analog in-memory compute

Goal: demonstrate an analog in-memory compute (AiMC) array

Objectives :
- adapt the SOT-MRAM stack and cell to AiMC
- validate cell functionality
- understand requirements for AiMC applications

Status :
- SOT-MRAM junction resistance modified to target AiMC operation conditions
- Full eNVM & back-end modules integrated on dedicated test vehicle
- Wafer run with dedicated AiMC stack is currently ongoing



A traditional deep neural network (DNN) layer can be mapped on an analog array to perform the multiply-accumulate (MAC) function in a very energy and area efficient way

The analog MAC is performed by summing currents of different (resistive) weights. To be energy efficient, this requires the device resistance to be very high (in MΩ range).

SOT-MRAM is a 3-terminal magnetic memory with decoupled read/write path. The thickness of the MgO insulating layer can be tuned to increase resistance in the read part without affecting the ability to write. Providing excellent flexibility

Spin-Orbit Torque MRAM: a 3T device

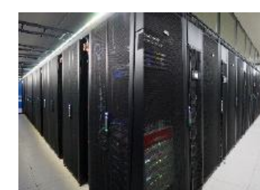
Charge-to-Spin conversion materials:
Non magnetic materials with large SO coupling create ***spin current perpendicular to charge current*** *(Spin Hall effect, Rashba field)*

→ **In-plane write**

- **Decoupled read/write**
  → "Unlimited" endurance
  → Negligible read disturb
- **Spin current geometry**
  → Limited incubation time
  Reliable sub-ns write

TMR Read — RL, MgO, FL — BE 1, SOT Write, BE 2

I.M. Miron, K. G. et al. *Nature* 476, 189 (2011)
Liu et. al, *Science* (2012)
Manchon et. al, Rev. Mod. Phys. 91, 035004 (2019)
Garello. et. al. *APL* 212402 (2014)
Cubucku et. al, IEEE Trans. Mag. (2018)



- SOT tackles HD/HP-SRAM replacement
- Decoupled read/write path provides flexibility for other usage such as in analog in-memory compute

ANDANTE

AI for New Devices And Technologies at the Edge

andante-ai.edu linkedin.com/company/andante-ai mario.diaznava@st.com (coordinator)

ECSEL Joint Undertaking

"This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement № 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Germany, Netherlands, Portugal, Spain and Switzerland"

Scan Me to visit website