ANDANTE

Non-Volatile Memory (NVM) based In-Memory Computing (IMC) solutions

ANDANTE

Motivations

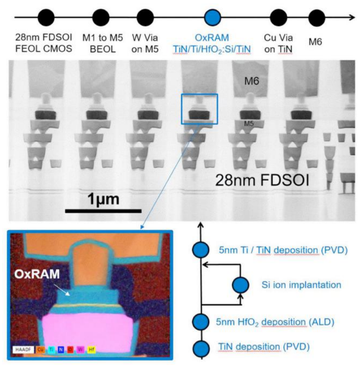
- In data-centric applications, such as AI, memory accesses consume a large part of the total energy dissipation. This is why doing computations in-memory is one of the key solutions for reducing power at the edge
- Furthermore, as the number of neural network parameters increases over time, relying on dense non-volatile embedded memories avoids duplicating memories and loading a considerable number of weights at each start up.

Main Goals

1. Implement and benchmark several In-Memory Computing solutions, leaveraging respectively OxRAM, SOT-MRAM and FeFET technologies.
2. Deliver hard macros and IPs to future Edge AI hardware accelerators

OXide-based resistive RAM (OXRAM) Analog IMC

Objectives:
- enable multi-level abilities of OXRAM technology
- demonstrate the feasibility of analog matrix-vector multiplication



Memory device inserted in the back-end of line

Operation	Initial state	Final state	V_{SL}	V_{BL}	I_{WRITE} control
Forming	$R_{PRISTINE}$	LRS	Ground	High	Yes
Reset	LRS	HRS	High	Ground	No
Set	HRS	LRS = f(I_{WRITE})	Ground	High	Yes

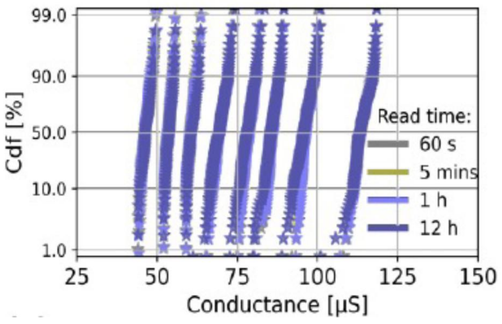
Writing operations
(HRS: High Resistive State; LRS: Low Resistive State)

State	R_{OxRAM}
$R_{PRISTINE}$	Superior to 10 MΩ
LRS	Inferior to 10 kΩ
HRS	Between 20 kΩ and 200 kΩ

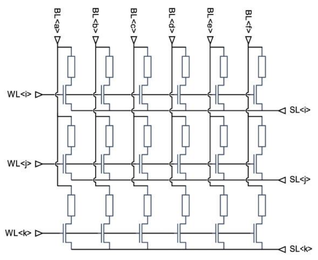
Resistance ranges

OxRAM-based design guidelines for analog in-memory compute
- Drastic limitations of pure crossbar array for RRAM are well known (sneak leakage path): need for an access transistor.
- Unitary bitcells are thus 1T1R cells, composed an OxRAM and a transistor
- In order to process analog matrix-vector multiplication, the bitcells array has to be arranged with orthogonal Bit Line (BL) and Source Line (SL). Voltages are applied at the input and currents are read at the output.
- In a multilevel approach, the constraints on the access transistor On Resistance (RON) are high. Since the LRS range is narrow and the resistance values low, the RON needs to be particularly low, which implies to have a wide and short transistor. The resulting cell size is 0,288 µm x 0,741 µm = 0,213 µm^2.
- The memory density is defined by the number of levels which can be stored in an OXRAM. Because of relaxation mechanisms and process variability, a smart multilevel programming technique must be employed: 8 levels are experimentally demonstrated, leading to a density of 14 bits / µm^2

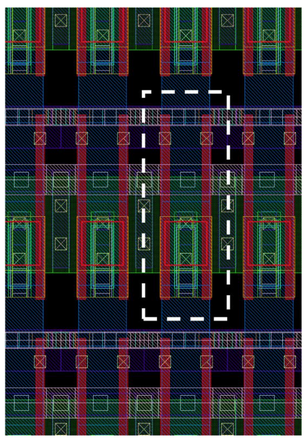
E. Esmanhotto et al., Advanced Intelligent Systems (2022)



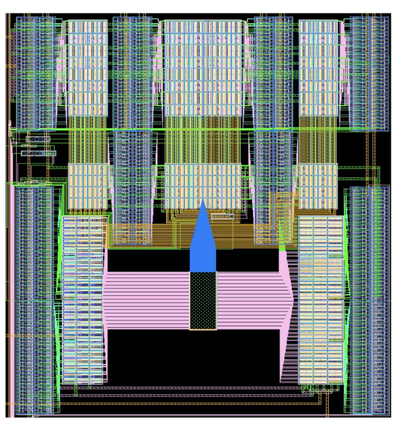
Experimental demonstration of stable 8-level storage



1T1R bitcell array



Bitcell layout (3b/cell)



Conclusion and future work

- OXRAM technology is a promising technology for implementing analog matrix-vector multiplication in heavily quantized neural networks
- The stable storage of 8 levels per cell (i.e. 3b) was experimentally demonstrated
- An analog IMC macro was designed and will be electrically tested

Ferroelectric FET (FeFET) Based Mixed-Signal IMC Accelerator

Fraunhofer

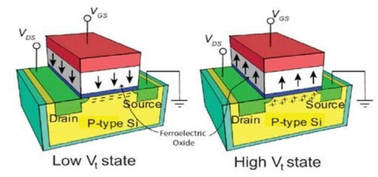
Objectives :
- To demonstrate the operational capability of FeFETs as efficient memory cells within AiMC architecture.
- To address latency and power consumption challenges in deep learning , particularly in IoT devices.
- To deliver a full IP design showcasing its capability in executing matrix-vector multiplications (MVMs).
- To ensure a robust and functional AiMC architecture conducive for efficient machine learning inference.

Building Blocks of FeFET-based Analog In-Memory Computing

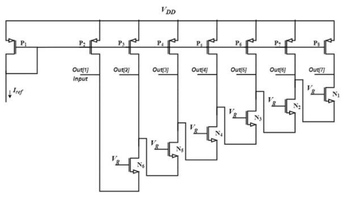
FeFET as Memory Cell: FeFETs in LVT and HVT states serve as pivotal memory cells enabling binary storage—essential for in-memory computations.

3-bit Flash ADC: It is crucial for sensing and quantizing the Multiply-and-Accumulate (MAC) operations, bridging analog computation with digital readout.

Layout of the Segment: It exemplifies a segmented architecture, housing an 8x8 FeFET array connected to a degeneration resistance, to minimize power consumption and mitigate current variability of the FeFET.

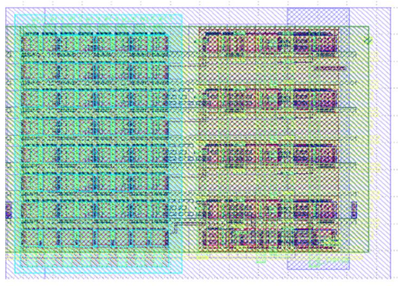


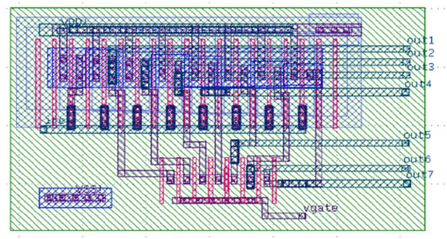
LVT and HVT state of the FeFET



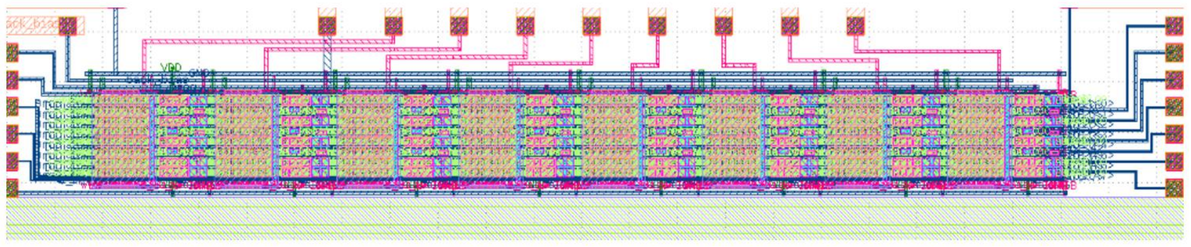
3 bit flash ADC

Layout of the Column Segment: It aggregates eight 8x8 1FeFET1R segments, each connected to a 3-bit Flash ADC. This setup fosters efficient MAC operations and ensures a precise analog to digital transition.

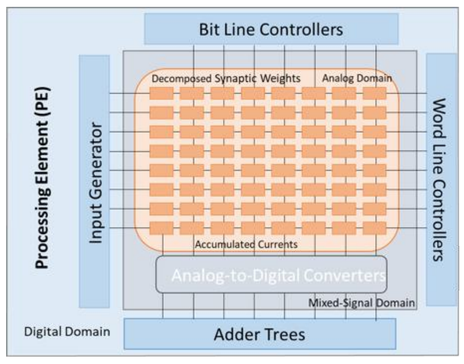




Segment and ADC Layout



Layout of the column segment



Processing Element For Neural Network Acceleration

FeFET Mixed Signal Processing Element:
The FeFET crossbar coupled with digital interfaces like bitline and wordline controllers, input generator, and adder trees, enables a smooth analog to digital transition in deep learning computations.

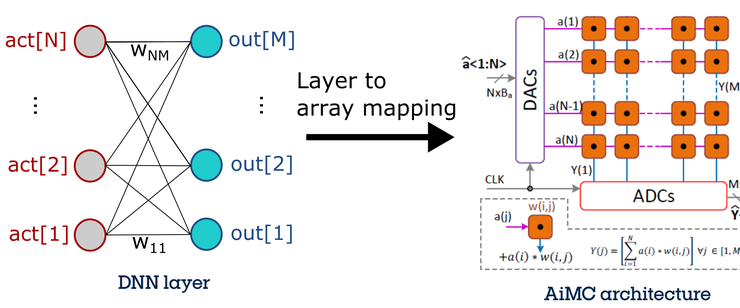
Conclusions and Future Work:
- The FeFET-based AiMC architecture exhibits promise towards efficient and reliable matrix-vector multiplication in deep learning computations.
- The described macro has been taped out, and measured.
- Characterization of the array is to be disseminated in future reports.

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

imec

Goal:
demonstrate an analog in-memory compute (AiMC) array using SOT-MRAM devices performing matrix vector multiplications (MVMs)

Objectives :
Device level scorecard for benchmarking AiMC
Full circuit-level array design



DNN layer

Layer to array mapping

AiMC architecture

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).

AiMC Key Components:
- Compute Cell: SOT-MRAM-based with read and write transistors
- Digital to Analog Converters (DAC): 512 DACs organized into 8 blocks of 64 DACs per block
- Analog to Digital Converters (ADC): Successive Approximation ADC (SAR ADC) presents energy-efficient structures having small area

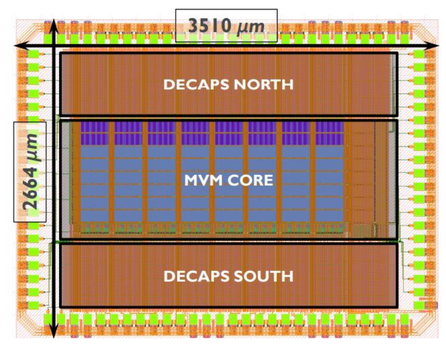
Memory Devices for AiMC

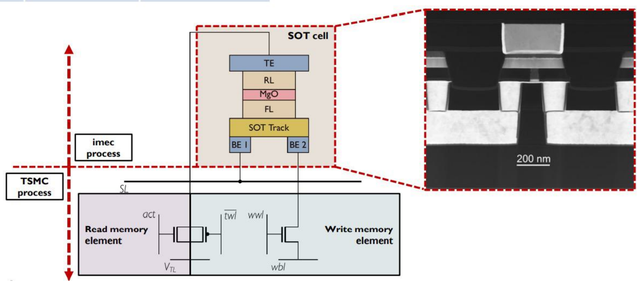
	RRAM	SOT	SRAM	IGZO	Ideal
Type	Res	Res	CS	CS	-
Non-Volatile	Yes	Yes	No	No	Yes
R_{ON}	< 0.1 MΩ	> 1MΩ	-	-	> 1MΩ
I_{ON}	-	-	< 1 µA	< 1 µA	< 1 µA
ON/OFF	Large	Small	Large	Large	Large
Cell area	Medium	Medium	Huge	Small	Small
Variations	High	Low	Low	High	Low
FEOL free	No	No	No	Yes	Yes
Multilevel	Yes	Yes	No	Yes	Yes

SOT-MRAM is a 3-terminal magnetic memory with decoupled read/write path. It presents area and low-leakage advantages over conventional SRAM, while IGZO DRAM would still require periodic refresh, thereby hindering the objective of achieving low-energy operations for ML accelerator.

AiMC with SoT-MRAM:
- Circuit design supporting MVM core with SOT devices comprising the compute cell, DAC and ADC
- implemented in TSMC 40 nm FEOL process and a combination of TSMC-imec layers for BEOL.

Top Level GDSII





SOT based compute cell with read and write transistors

Conclusions and future work:
- Design has been taped out manufacturing is ongoing.
- Future reports will disseminate array characterization.

ANDANTE

AI for New Devices And Technologies at the Edge

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

"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"

ECSEL Joint Undertaking
Electronic Components and Systems for European Leadership

Scan Me
to visit website