UCLouvain

Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)

icteam

S'attaquer aux problématiques de test et de fiabilité des accélérateurs IA analogiques et mixtes

Martin Andraud

FETCH'26

FAMES

greener electronics

FD-SOI · NVM · 3D · RF · PMIC

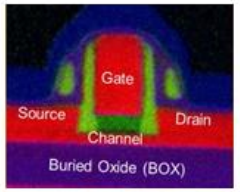
The FAMES pilot line

FAMES
greener electronics
FD-SOI · NVM · 3D · RF · PMIC

- FAMES is an EU pilot line, promoting semiconductor technologies in Europe
- Driven by CEA-Leti
- **The role of my group:** demonstrating AI acceleration systems based on eNVMs

FD-SOI

10nm and 7nm

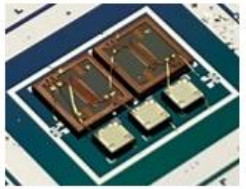


Radiofrequency components

switches, filters, and capacitors

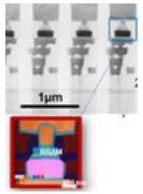
Small inductors for DC-DC converters

Power Management Integrated Circuits (PMIC)



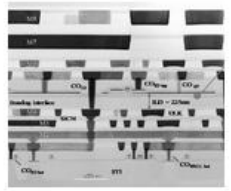
Several types of embedded non-volatile memories

OxRAM, FeRAM, MRAM and FeFETs

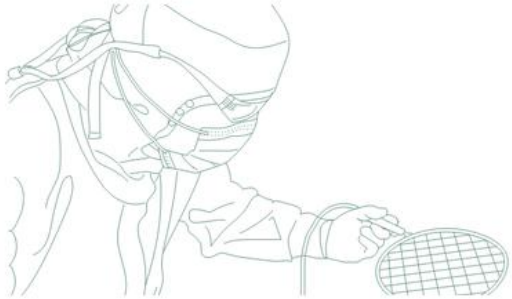


3D integration

heterogeneous and sequential



Sustainability and eco-design



Outline

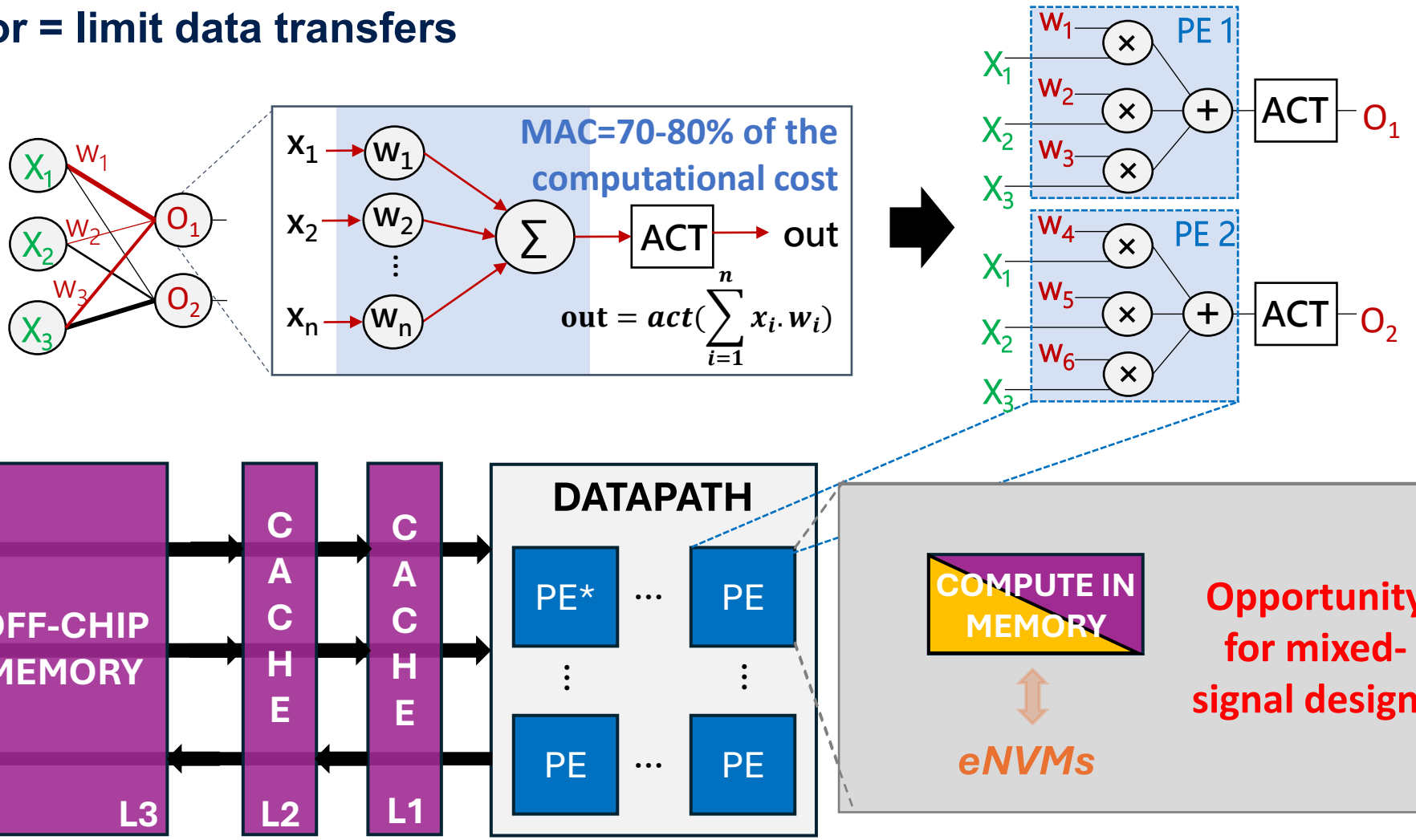
- AI accelerators: from digital processors to Compute-In-Memory (CIM):
  - Design overview (with a focus on **mixed-signal** topologies)

- Our work on reliable mixed-signal CIM systems:
  - How to evaluate the **accuracy** of mixed-signal CIM**?**
  - **Self-calibration** for noise and temperature compensation
  - **Did we do a good design?**
- Conclusions and perspectives

UCLouvain

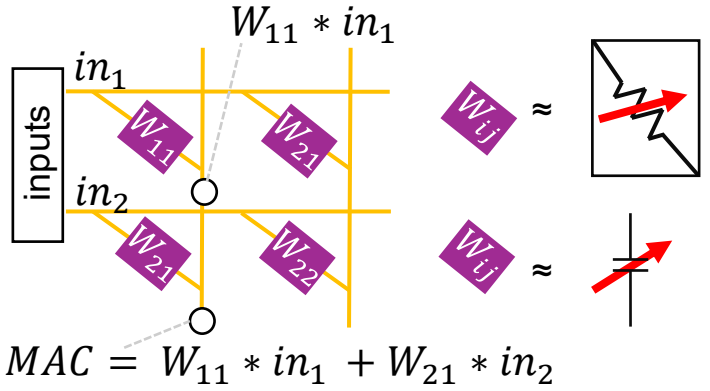
A typical digital AI accelerator

- **AI accelerator = limit data transfers**

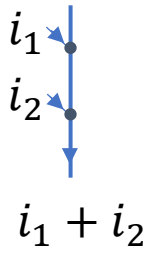


Analog compute in memory arrays

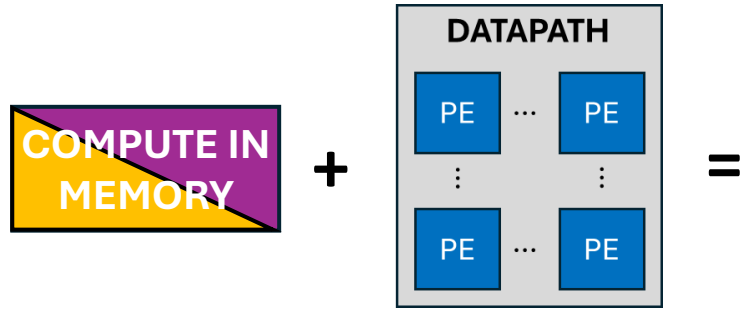
- Based on basic **voltage** and **current** laws

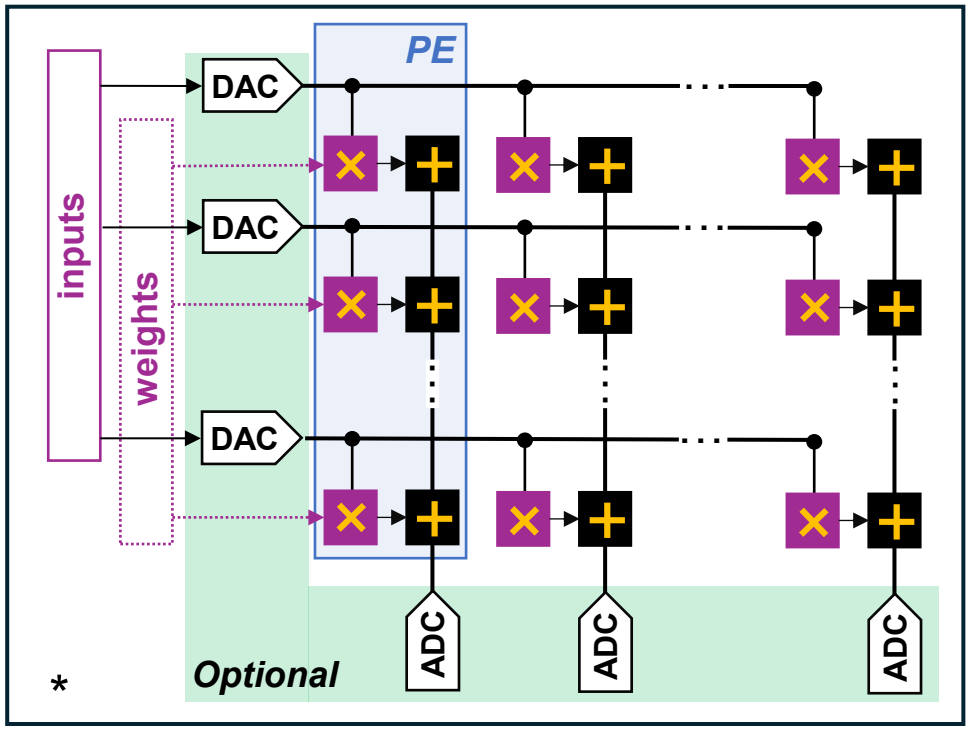


$I_1=(1/R_1).in_1=W_{11}.in_1$



$i_1 + i_2$



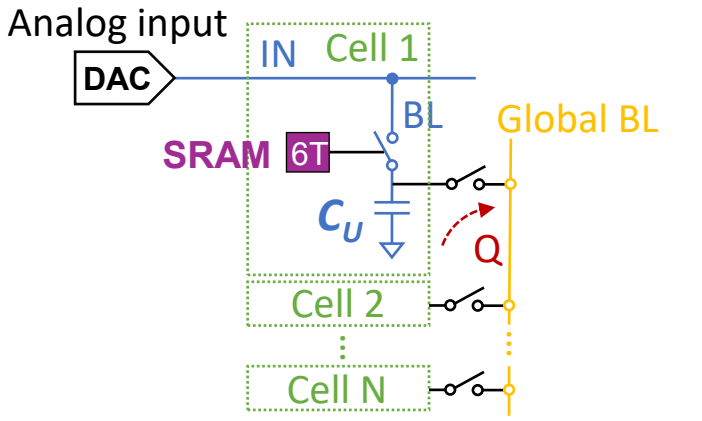


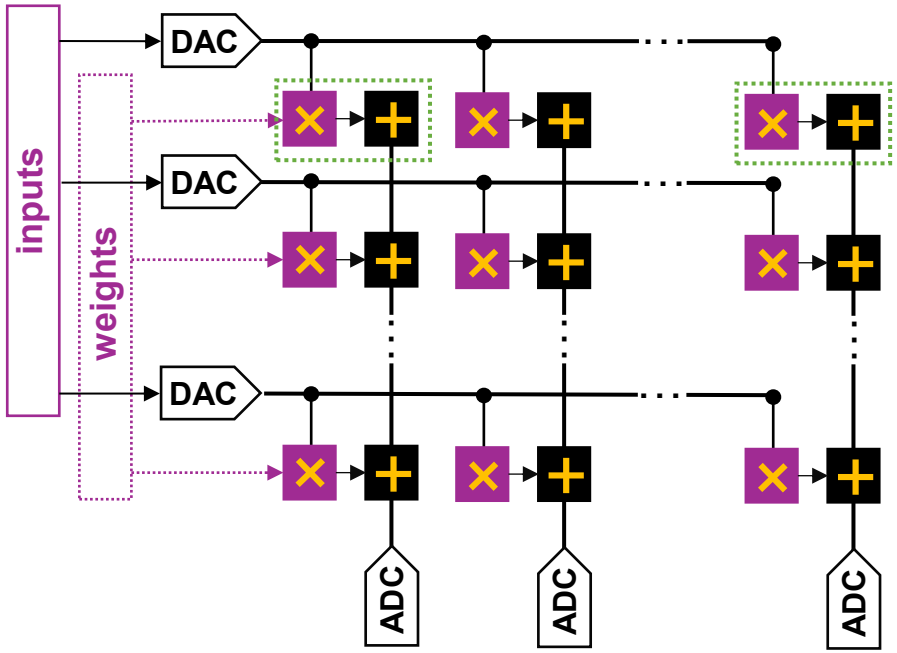
- How to measure the **performance** (op=MAC)?
  - How fast you go (ops/s) ↗ array size, clock
  - How efficient you are (ops/s/W) ↘ cell power
  - How much area it takes (ops/s/mm^2) ↘ cell size

Different types of mixed-signal CIM

CMOS-based
(capacitance or resistance)

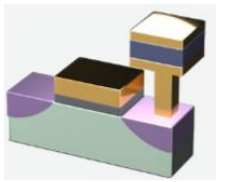
(Charge-based example [1])





eNVM – based
(emerging non-volatile memories)

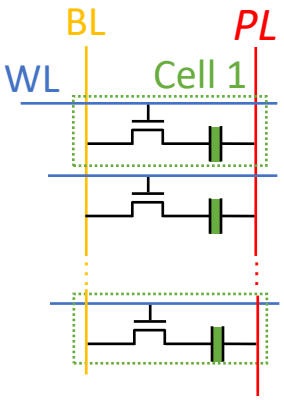
(Example using ferroelectric eNVMs [2])



1T-1C FeRAM (feCAP)

*-Two polarization states, storing "1" or "0", programmed through **PL line***

*- Reading the charge value on **BL** (destructive)*



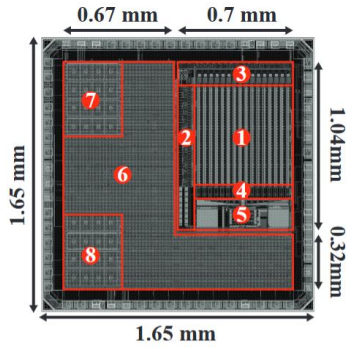
[1] N. Verma et al., "Compute-In-Memory: advances and prospects," SSC mag, 2019.

[2] J. Duan et al., "A full spectrum of 3D ferroelectric memory architectures shaped by polarization sensing," preprint, *2025*.

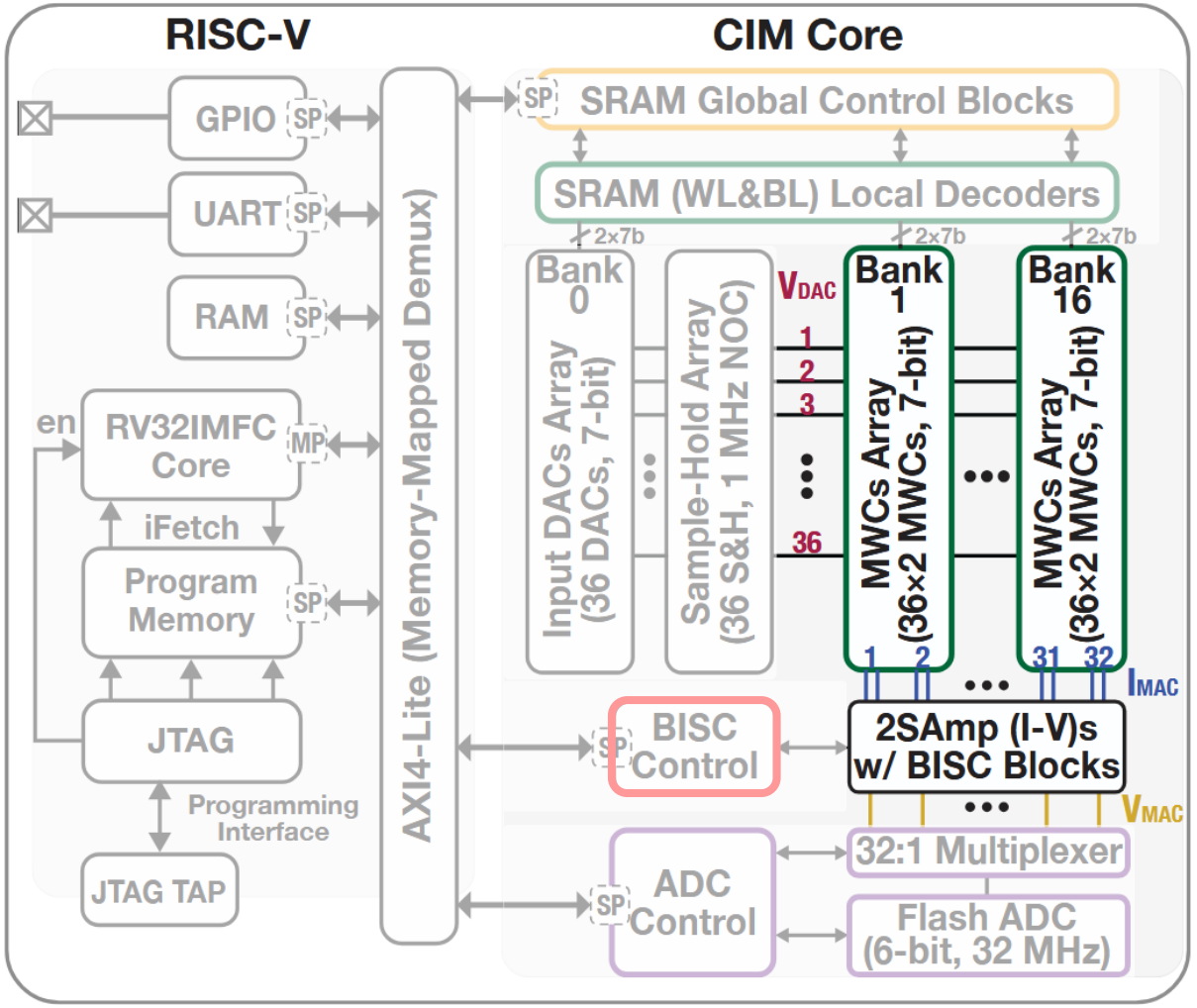
- **Ok, but what about computational accuracy ?**

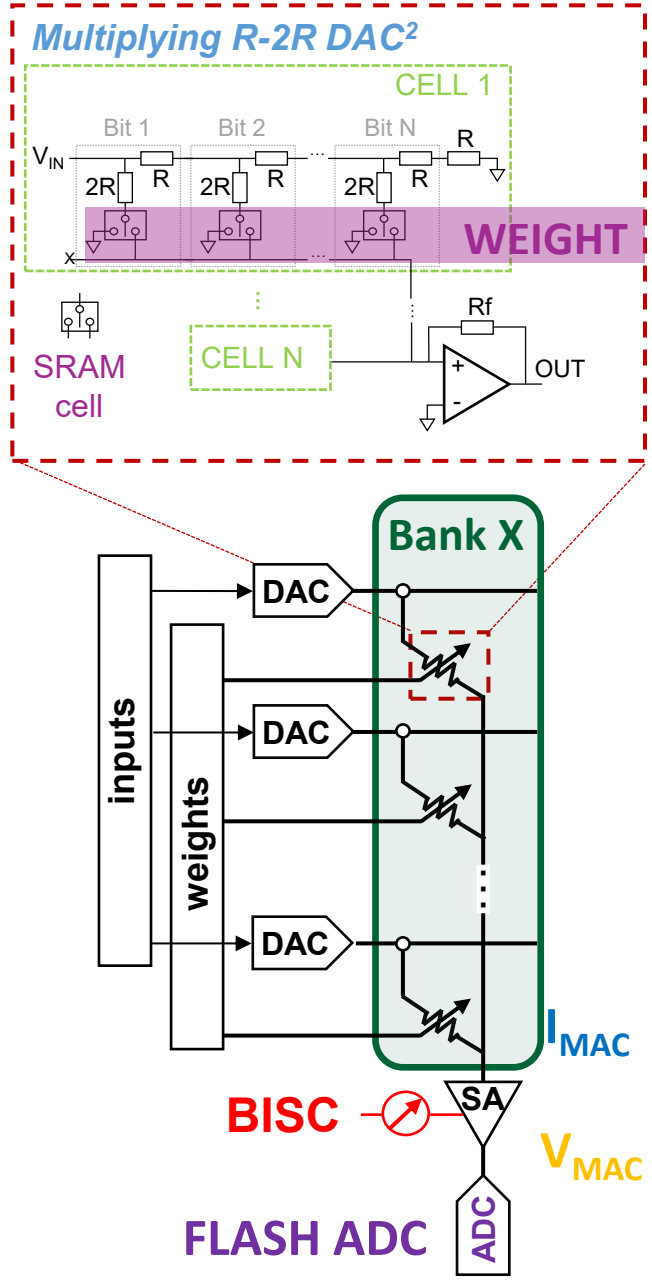
UCLouvain

Case study: A-core CIM chip



- GF22nm
- 36x32 CIM block
- 6-bit accuracy
- RISC-V control
- Built-in self-calibration





Measure the accuracy your ANA-CIM core

- We can the define the **compute SNR (SNR_T)**
- How to optimize the **compute SNR**?
  - *Better ADC SNR? Increase resolution*
  - "Best" design: analog-limited*[1] :*

👍 $\boldsymbol{Rule\ \#1: SNR_{ADC} = SNR_A + 9dB}$

- What is a "practical" value for **SNR_A**?
  - The compute SNR depends on each DNN layer

$$10\text{dB} \leq SNR_T \leq 40\ dB$$

  - Rule of thumb: we try to **limit the accuracy degradation compared to software at 5%**

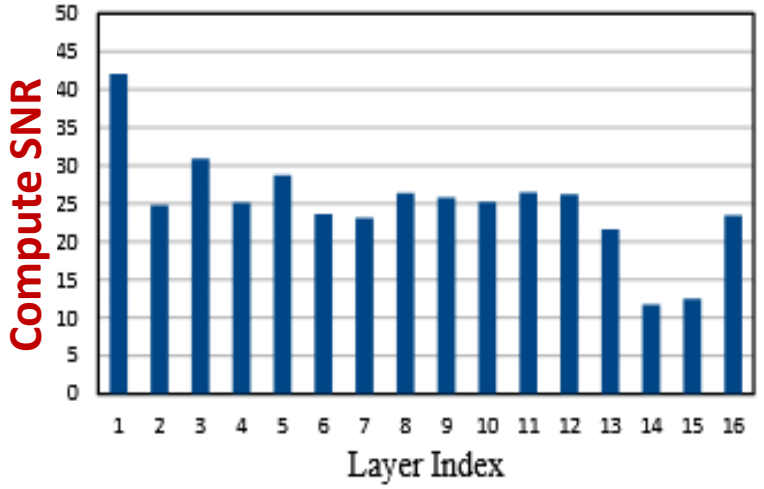
👍 $\boldsymbol{Rule\ \#2: fix\ SNR_T \approx SNR_A \approx 20dB}$

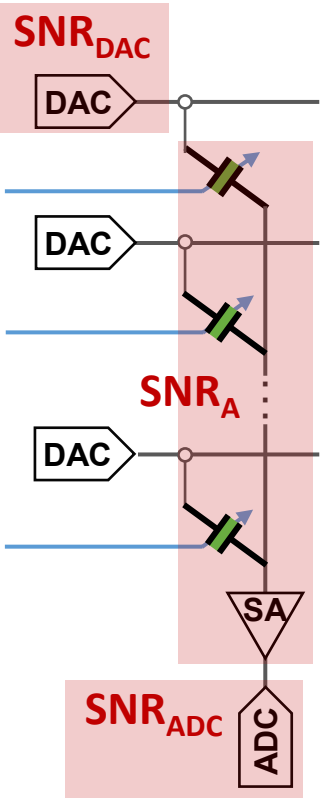
➔ $\boldsymbol{SNR_{ADC} \approx 30\ dB \Rightarrow ENOB = 4.6\ bits}$

$$SNR_T = \left[\frac{1}{SNR_A} + \frac{1}{SNR_{ADC}}\right]^{-1}$$

$$SNR_{ADC} \gg SNR_A$$

Example with VGG 16 [1]





[1] Gonugondla et al., arXiv, 2020

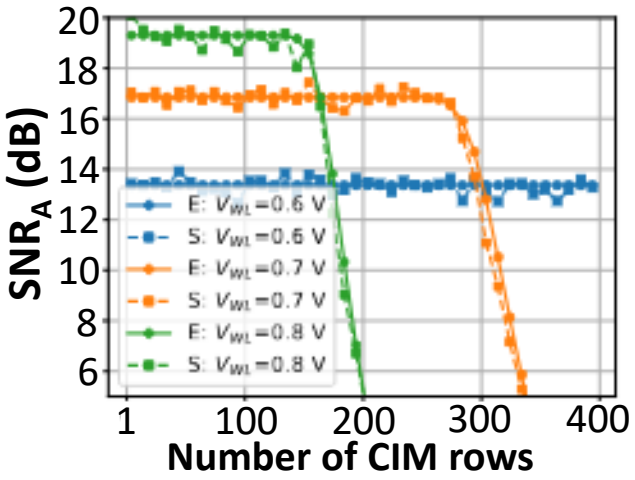
UCLouvain

Mixed-signal CIM: a need for calibration

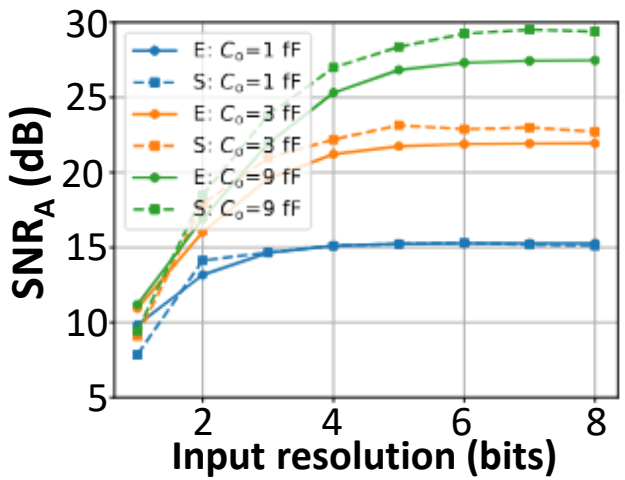
CMOS CIM

eNVM CIM

[1]

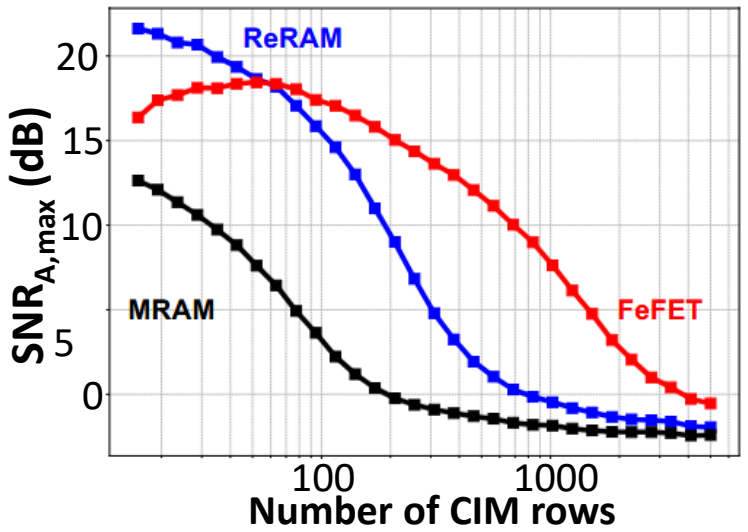


Limited by architecture & noise in charge-sharing or current accumulation



Mitigation with typical design trade offs (increase cap for better SNR)

[2]



Limited by process variations, mismatches, and thermal noise

👎 **It is difficult to attain more than 20dB of SNR without calibration**

[1] Gonugondla **et al.,** arXiv, 2020
[2] Roy & Shanbhag, JXCDC'24

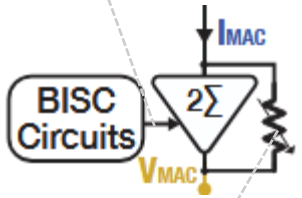
UCLouvain

Built-in Self Calibration: principle

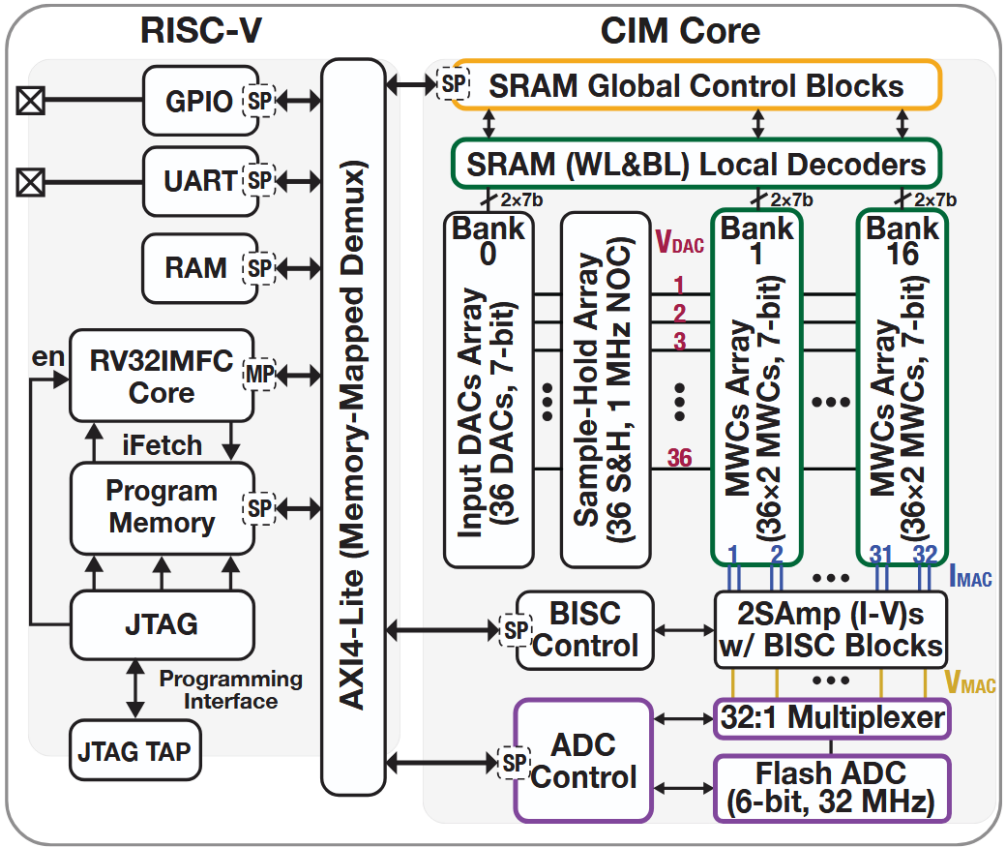
Offline (off-chip)

Characterize the link between gain and offset errors with the calibration voltage of the amplifier

$$V'_{CAL} = V_{CAL} - \frac{\hat{\epsilon}_{tot} - \beta_D}{\alpha_D C_{ADC}}$$



$$R'_{SA} = \frac{\alpha_D R_{SA}}{\hat{g}_{tot}}$$



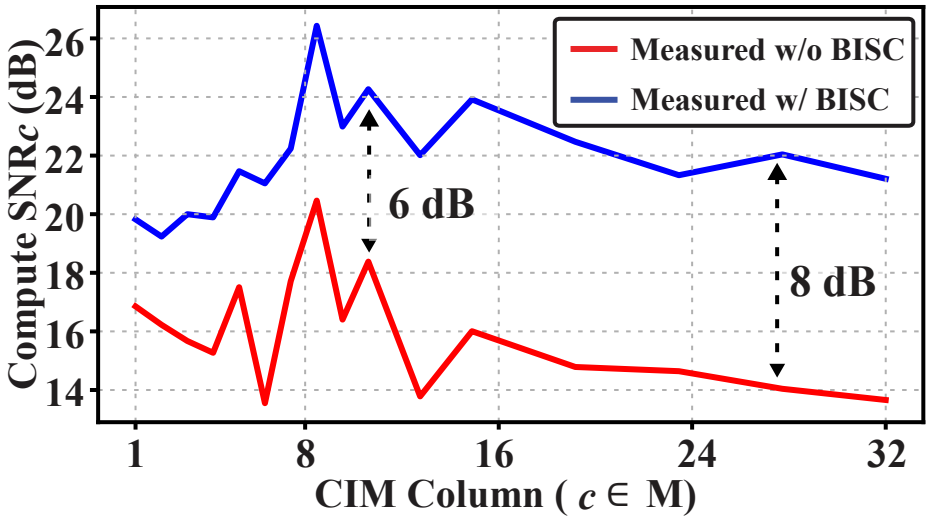
Online (on-chip)

1. *Generate test vectors (a set of known MAC operations) with the RISC-V processor*
2. *Calculate the error on the whole chain and estimate the compute SNR (SNR_T)*
3. *Apply correction factors to optimize SNR_T*

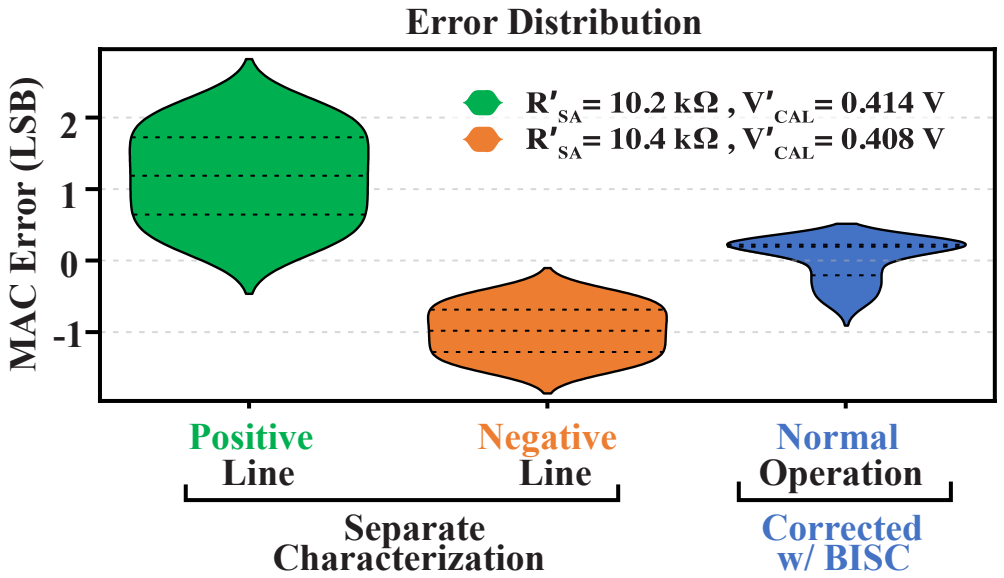
UCLouvain

SNR calibration

- Before calibration
  - Low compute SNR (14-20dB)
  - Large error (+2 LSB…)



- After calibration
  - Reduced computation error
  - The compute SNR is between **19-26dB**
  - The error is contained inside **± 1 LSB** after ADC quantization

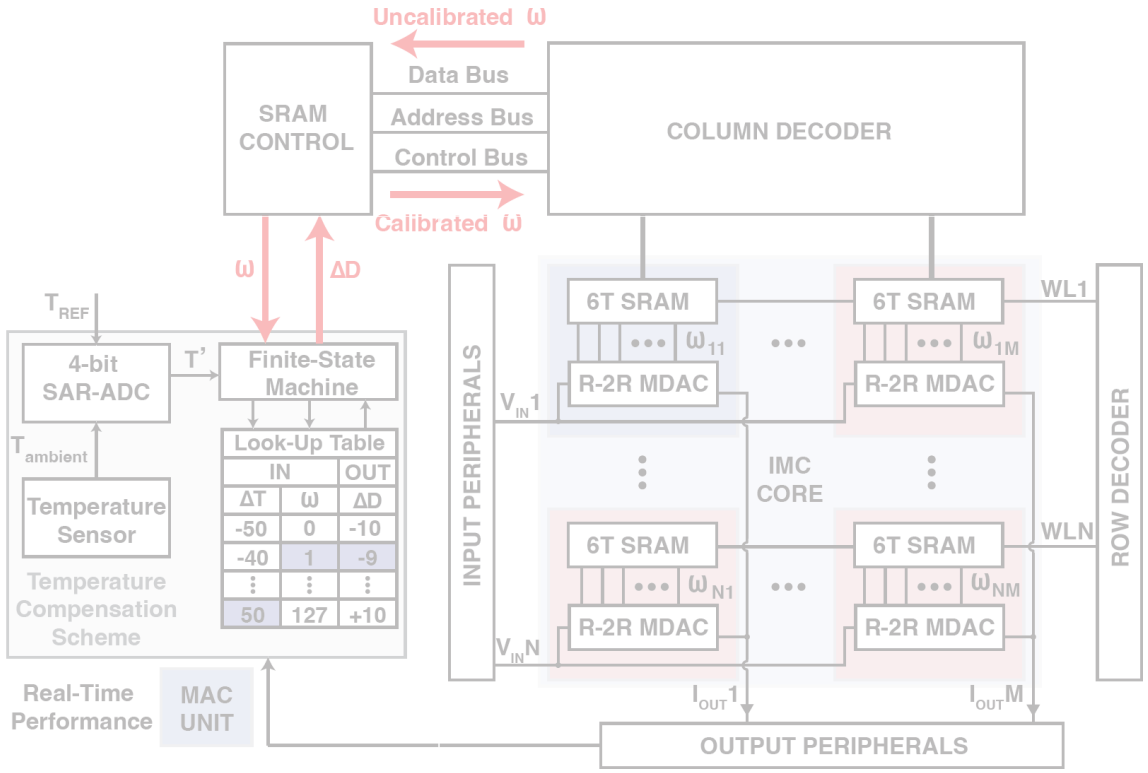


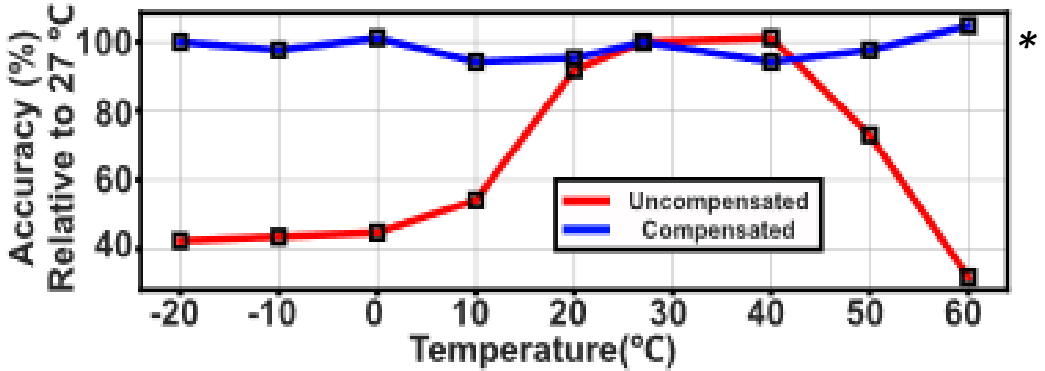
Temperature compensation

- Temperature **greatly** impacts the accuracy of mixed-signal CIM cores
  - Error build up as each component has a sightly different Temp. coeff.

$$R(T)=R_0\,[1+\alpha(T-T_R)]$$

- Our strategy: rewrite the stored weights to compensate for these variations
  - On-chip temp. sensor + LUT + mem write
  - Very effective (almost no accuracy loss)
  - Can be costly if frequent rewrite (but temperature variations are slow)





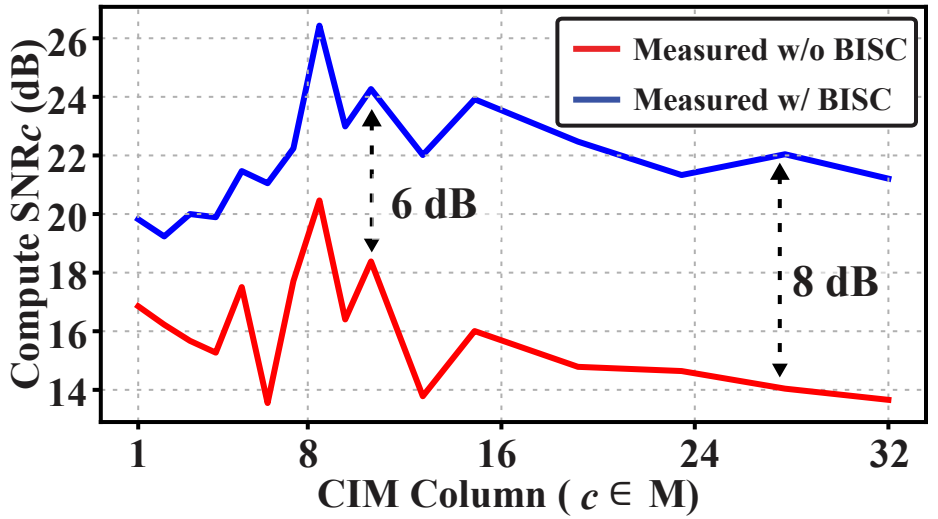
** Measured for MNIST task*

Did we do a good design?

- This depends how we look at it:
  - "Required accuracy": **yes**, we have a compute SNR of 20dB+ for most columns
  - System-wise?
    - SNR_A=34 dB (5.4 ENOB)
    - SNR_{ADC}=31dB (4.9 ENOB)

SNR_T=26,7dB

Syst. ENOB: 4.1



~~Rule #1: $SNR_{ADC} = SNR_A + 9dB$~~

Rule #2: $fix\ SNR_A \approx 20dB$

- **Lessons learned**
  - To target the "required" 20 dB of SNR, 4-bit weights and activations seem sufficient
    - **Yet, 5% accuracy loss is a lot compared to digital implementations**
  - Take care of your ADC design!
    - **Limiting factor in our case (no good pre-amplifier in the flash ADC)**
  - We have uncalibrated variations left
    - **Mismatch? Interferences between columns?**

Conclusions and perspectives

- Test and reliability of mixed-signal CIM is crucial to ensure the required computation accuracy
  - There are **large differences** between the *required DNN accuracy* and the *CIM system's accuracy*
  - The situation might get worse with **emerging technologies**

Parameter	Software baseline	**"Required" DNN accuracy**	**Optimal design (6b target)**	Current design (no BISC)	Current design (with BISC)
SNR_A (dB)	'40'	20	38	14 - 20	**20**-26
System ENOB (bits)	'6.3'	3	6	2.3 - 3	**3.3** - 4

- Calibration methodologies are required
  - Compensate for noise, mismatches, temperature, process…
  - **Design guidelines** and **existing automated strategies** are there to build your system

UCLouvain

Thank you for your attention!

martin.andraud@uclouvain.be

FAMES
greener electronics
FD-SOI · NVM · 3D · RF · PMIC