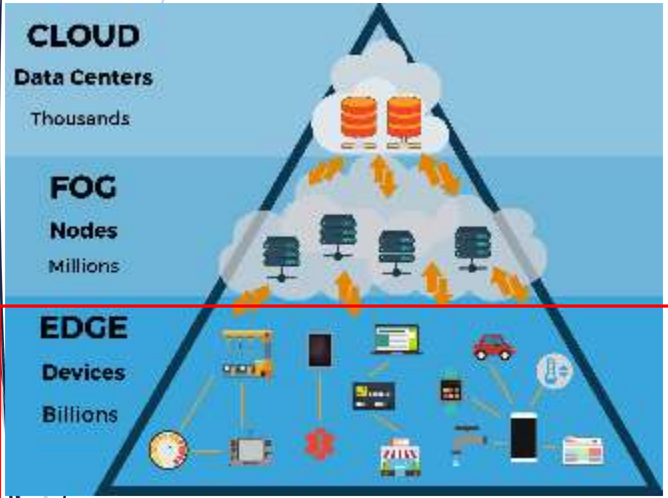
chist-era

UNICO - Unsupervised spiking neural networks with analog memristive devices for edge computing

Context

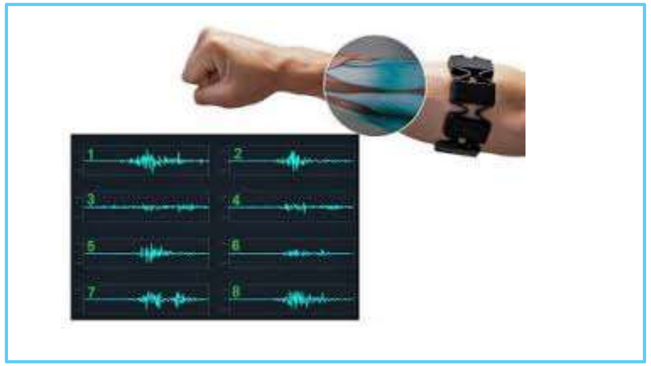
- **UNICO**: Spiking Neural network for Edge Computing Applications



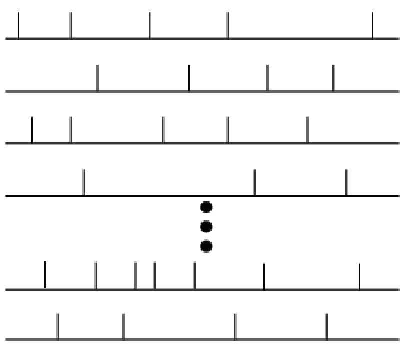
Signals with strong temporal signature from sensors

ToF camera

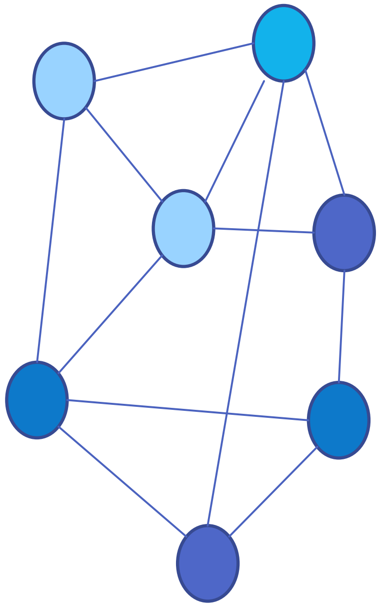
Bio-signals (e.g. EMG)



Spike encoding



Spike computing

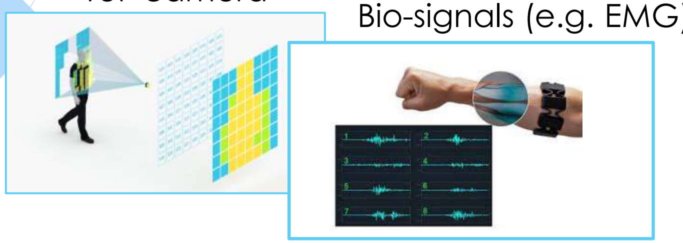


Context

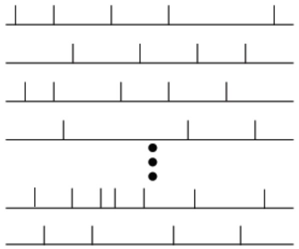
Signals from sensors

ToF camera

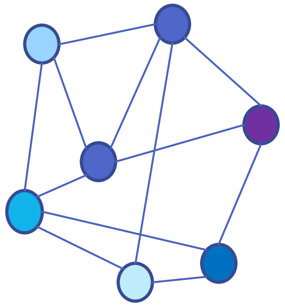
Bio-signals (e.g. EMG)



Spike encoding



Spike computing



- Key challenges and opportunities

SOFTWARE

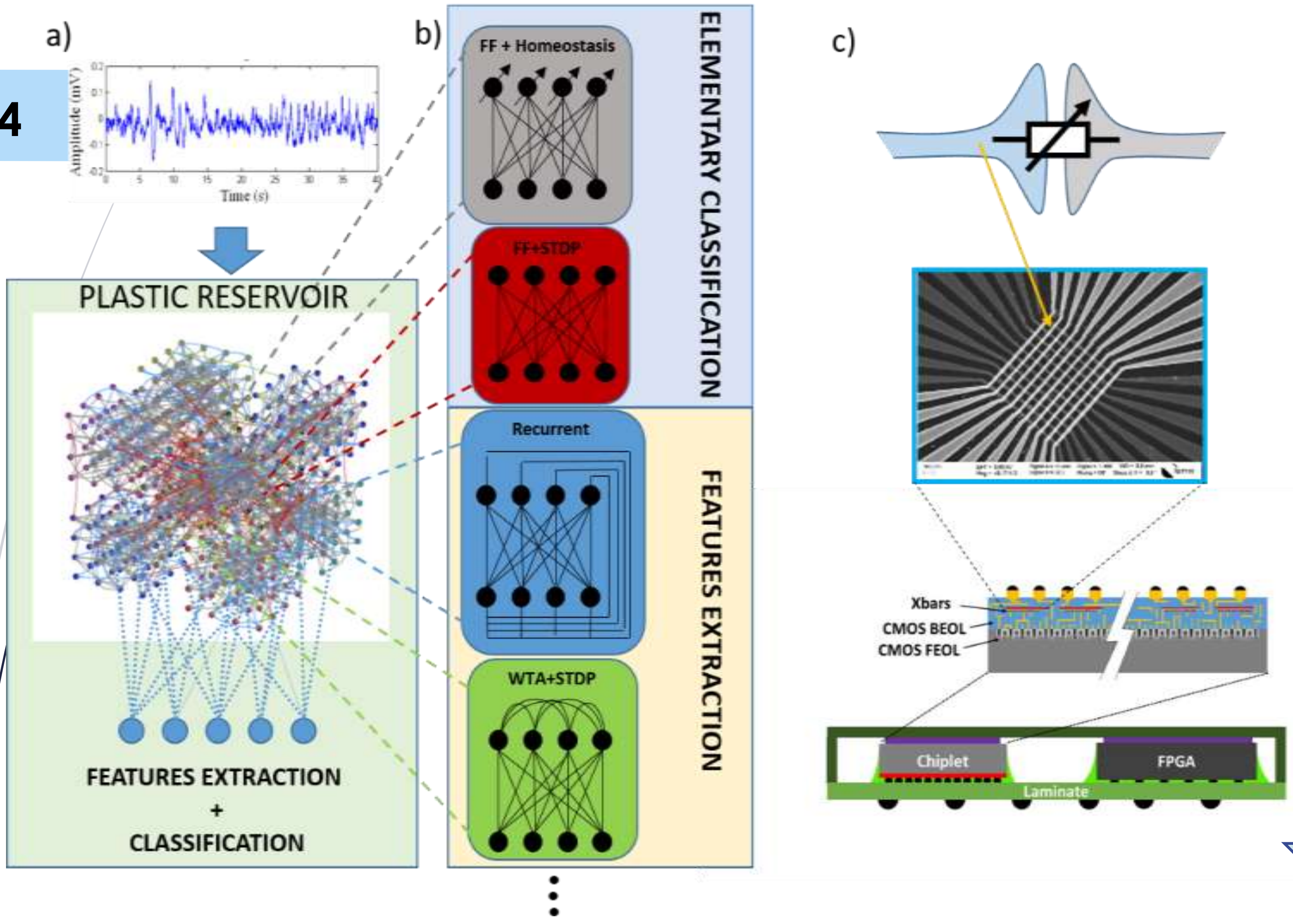
- Identification of spatio-temporal features
- In-sensor computing
- Footprints/Energy constraints (Edge Computing)

- Features enhancement vs full band encoding
- Event sensor vs continuous analog signals (time and frequency)
- Circuit design for encoding

- Benchmarking / perfromances
- SNN definition
- Hardware mapping opportunities and constraints

HARDWARE

The software/hardware strategy

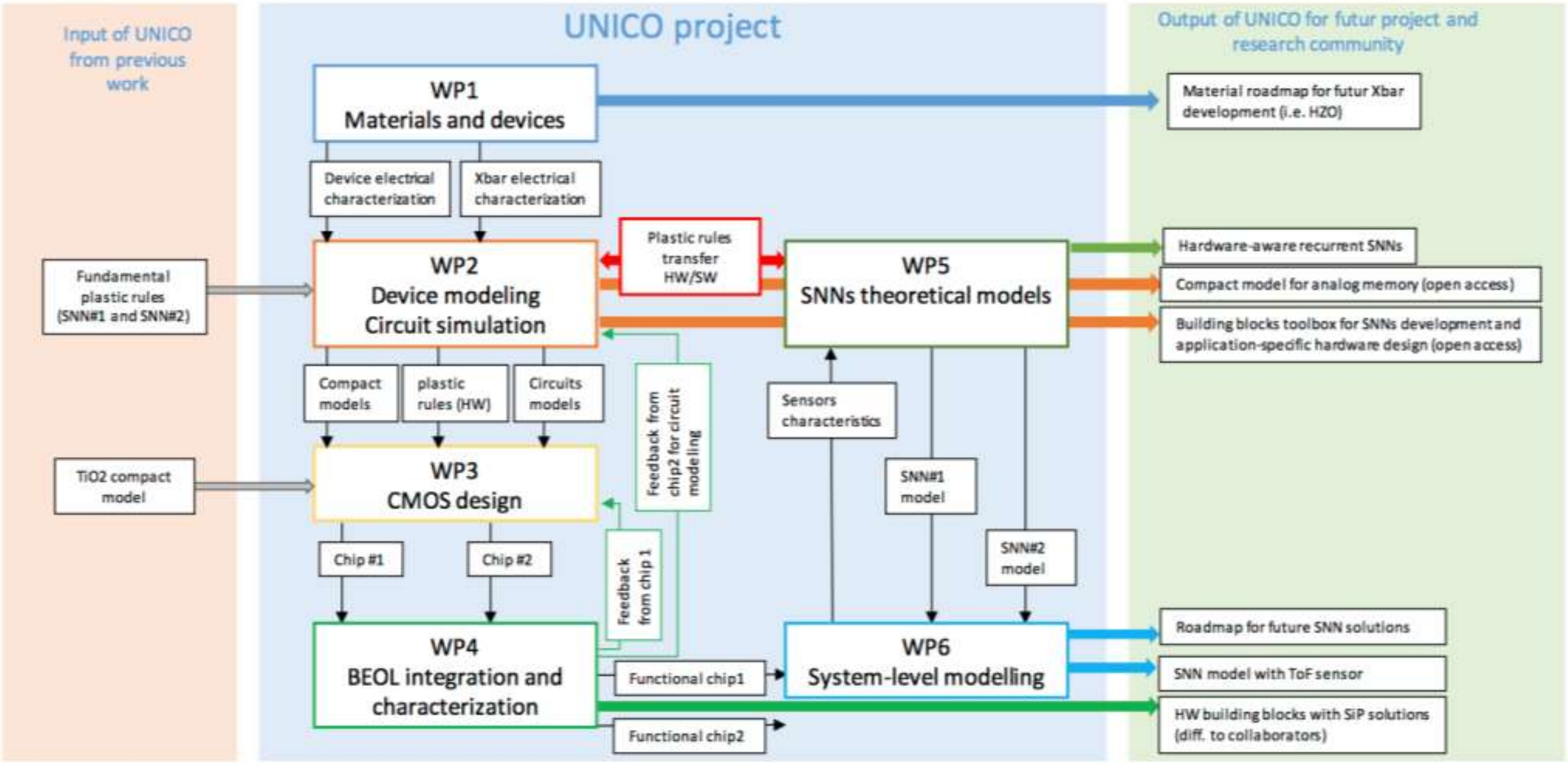


SW

- Find the optimal SNN
- Identification of key building blocks
- Engineering of building blocks
- Characterization of BB performances

HW

- System design
- Circuit design
- Device and circuit characterization
- Material / device fabrication

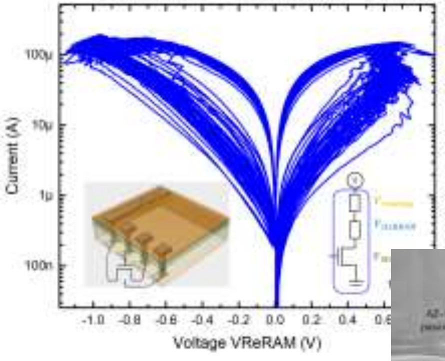

Input of UNICO from previous work
UNICO project
Output of UNICO for futur project and research community
WP1
Materials and devices
Material roadmap for futur Xbar development (i.e. HZO)
Device electrical characterization
Xbar electrical characterization
Plastic rules transfer HW/SW
Fundamental plastic rules (SNN#1 and SNN#2)
WP2
Device modeling
Circuit simulation
WP5
SNNs theoretical models
Hardware-aware recurrent SNNs
Compact model for analog memory (open access)
Building blocks toolbox for SNNs development and application-specific hardware design (open access)
Compact models
plastic rules (HW)
Circuits models
Feedback from chip2 for circuit modeling
Sensors characteristics
TiO2 compact model
WP3
CMOS design
SNN#1 model
Chip #1
Chip #2
Feedback from chip 1
SNN#2 model
WP4
BEOL integration and characterization
Functional chip1
Functional chip2
WP6
System-level modelling
Roadmap for future SNN solutions
SNN model with ToF sensor
HW building blocks with SiP solutions (diff. to collaborators)

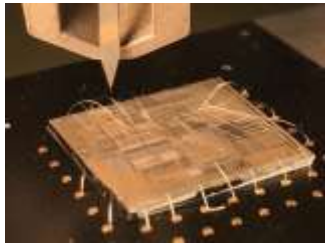
Involvement

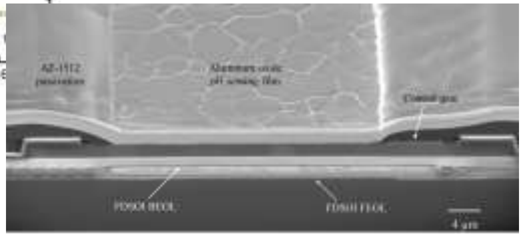
- Université de Sherbrooke (CA, UdeS-Nano & UdeS-ML) Coordinator
- CNRS / C2N (FR)
- CNRS/ Aix-Marseille University (FR)
- IBM-Zurich (CH)
- Jagiellonian University with CEZAMAT PW (PL)

NANOELECTRONIC GROUP
(Pr. D. Drouin)

The NANOELECTRONIC group has strong expertise in nanodevices fabrication, BEOL integration and packaging. The group evolves in a academic / industry transfer oriented environment (C2MI, IBM-Bromont).

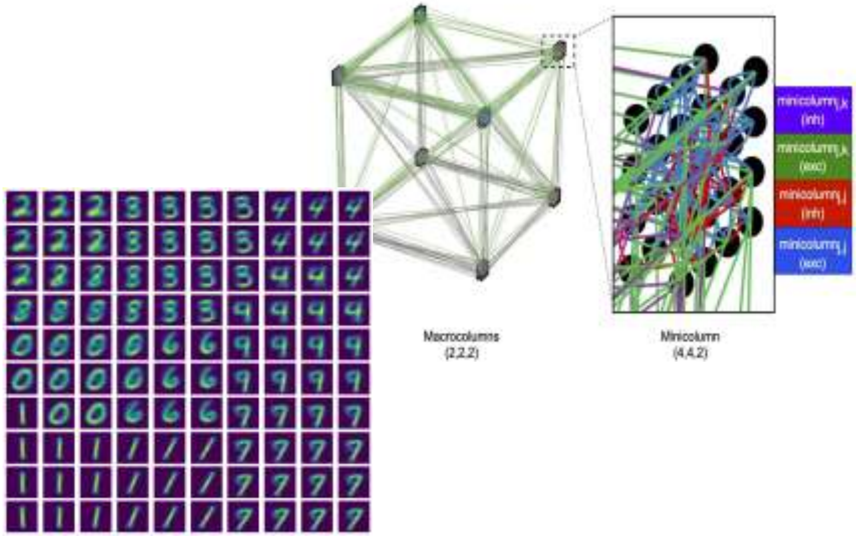




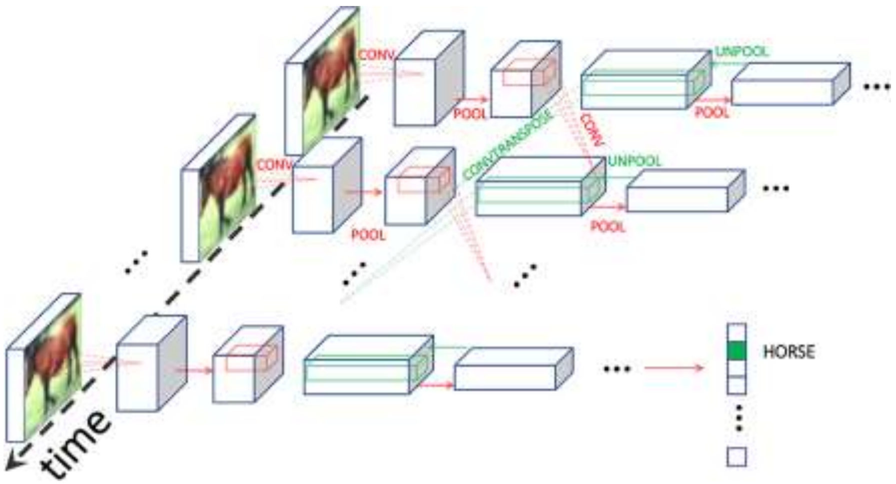


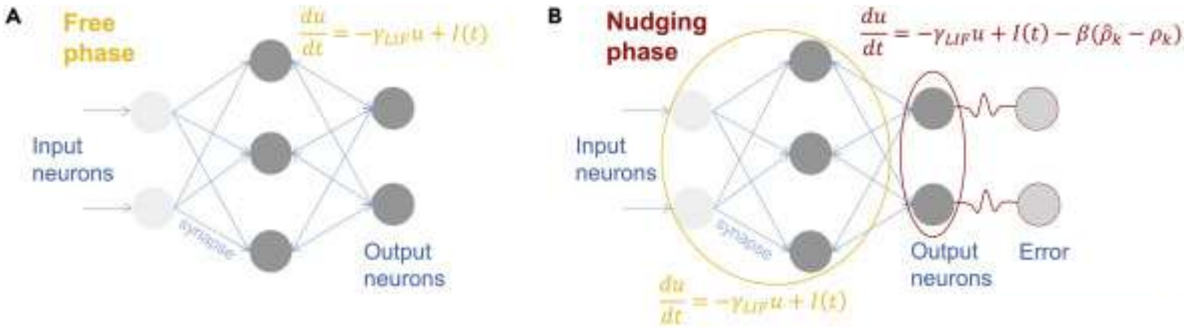
NECOTIS GROUP
(Neurosciences computationnelles et traitement intelligent des signaux)
(Pr. J. Rouat)

NECOTIS is a pioneering group in neuromorphic computing with interest in data analysis, spike computing and exploration of new SNN topologies

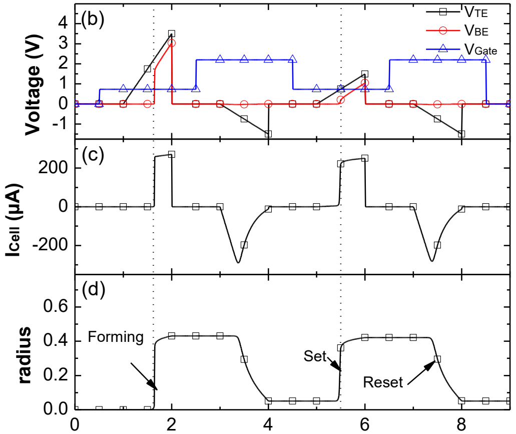


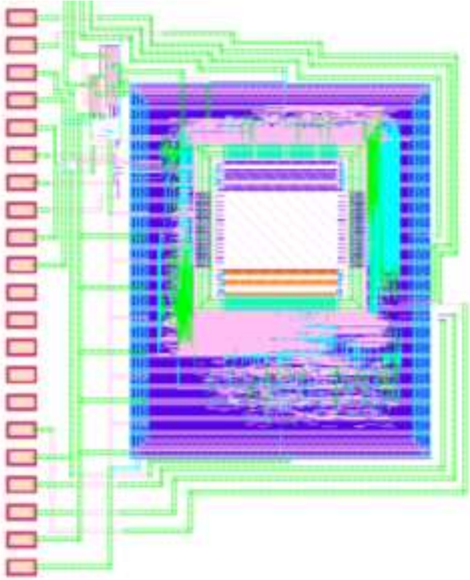
- Université Paris-Saclay (Centre de Nanosciences et de Nanotechnologies / C2N)
- Expert on the development of neural networks specific to neuromorphic/memristive hardware
- Provided assistance for the development of VDSP
- Currently developing an approach inspired by « Equilibrium Propagation » to scale VDSP to deep neural networks
  - *Large-scale simulations and planned small-scale validation on the UNICO demonstrator*



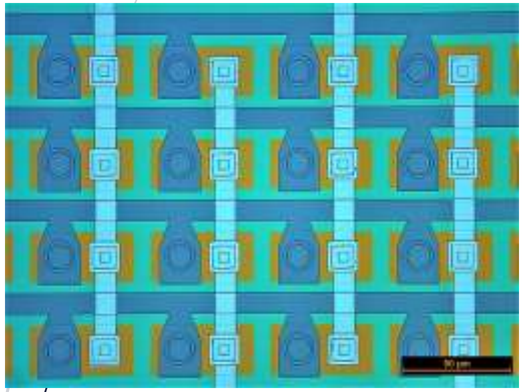


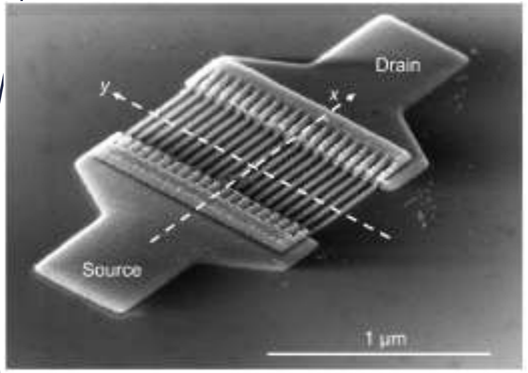
<u>AMU (IM2NP)</u> : The memories team at AMU (IM2NP) works exclusively on non-volatile memory technologies, covering several aspects from material science (e.g. AFM analysis), device electrical characterization, memory cell analytical and compact modeling, and memory circuit design with emphasis on IMC concept for AI applications.



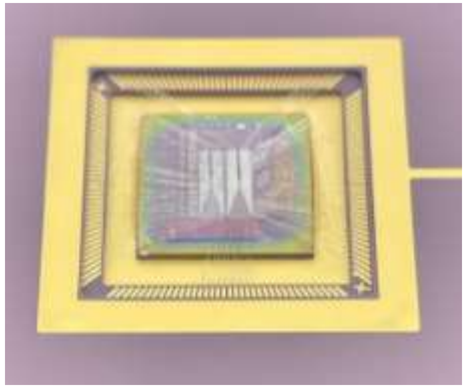


IBM Research | Zurich



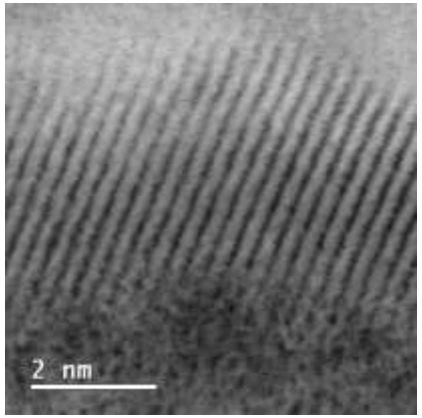


Established in 1956 in Switzerland, it represents the European branch of IBM Research (~300 employees)



Infrastructure: state-of-the-art cleanroom facilities (Binnig and Rohrer Nanotechnology Center) for processing nano-scale devices, advanced thin-film deposition techniques, materials characterization equipment. Characterization facilities for investigating memristive components and circuits.

Exploitation strategy (in addition to dissemination): Enabling technology for AI hardware: potential technology transfer to IBM's Albany Nanotechnology Center.



CEZAMAT

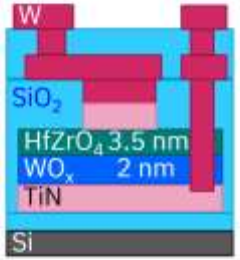
POLITECHNIKA WARSZAWSKA

JAGIELLONIAN UNIVERSITY IN KRAKÓW

CEZAMAT partner is leading the ToF sensor technology. They have strong expertise in ToF technology, design of Front End/Back End of sensors.
They are developping the software environment for the UNICO project.

Material/devices: HZO development

$HfZrO_4$ deposited by Atomic Layer Deposition then crystallised in the ferroelectric phase at **400˚C**
Remanent electrical polarisation: **Ferroelectric Field-Effects** in 2 or 3 terminals configuration.
Analog resistive switching obtained by the gradual switching of ferroelectric domains.



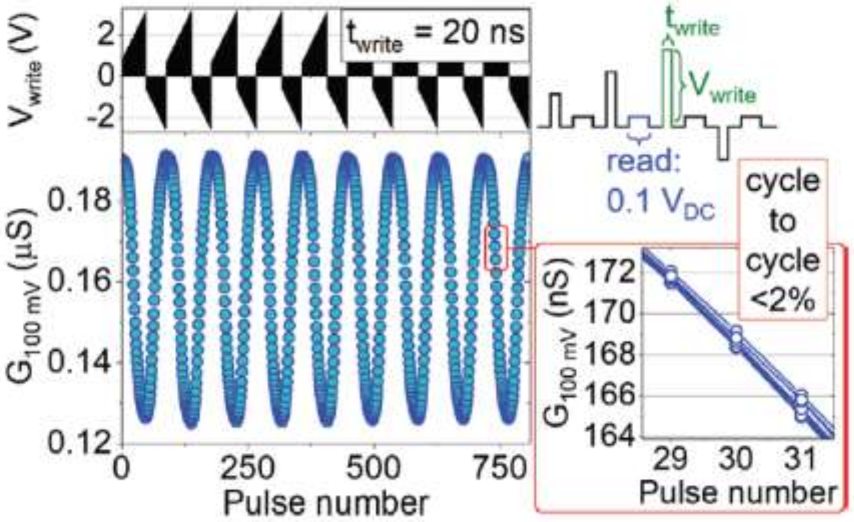


Figure 4. Synaptic potentiation and depression with pulses of increasing amplitude. The pulse amplitude V_{write} and the pulse duration t_{write} are defined as in the measurement scheme (top right). The inset shows the cycle-to-cycle variation.

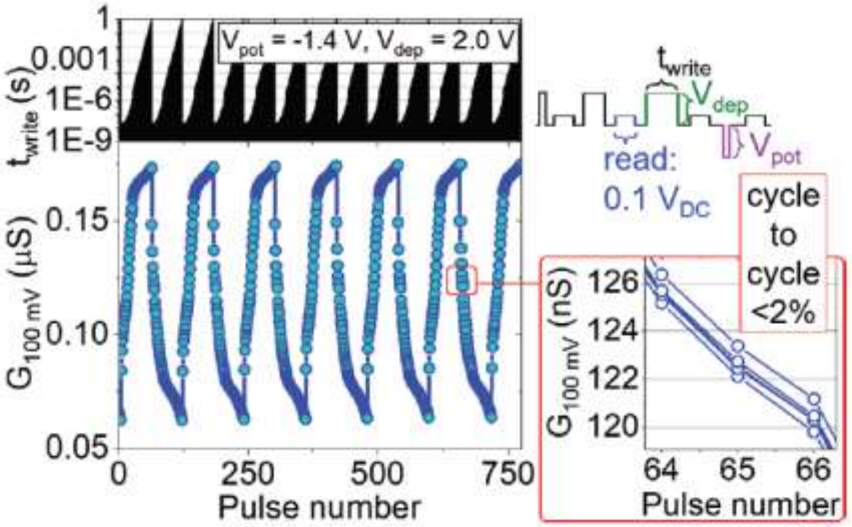
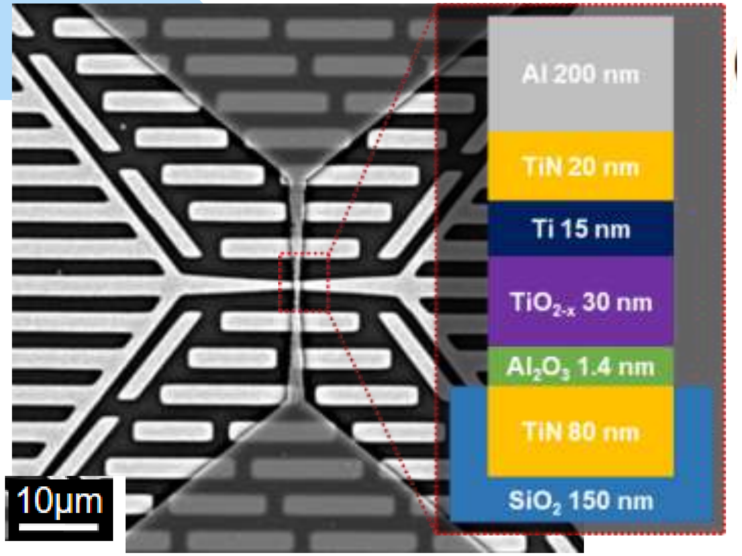


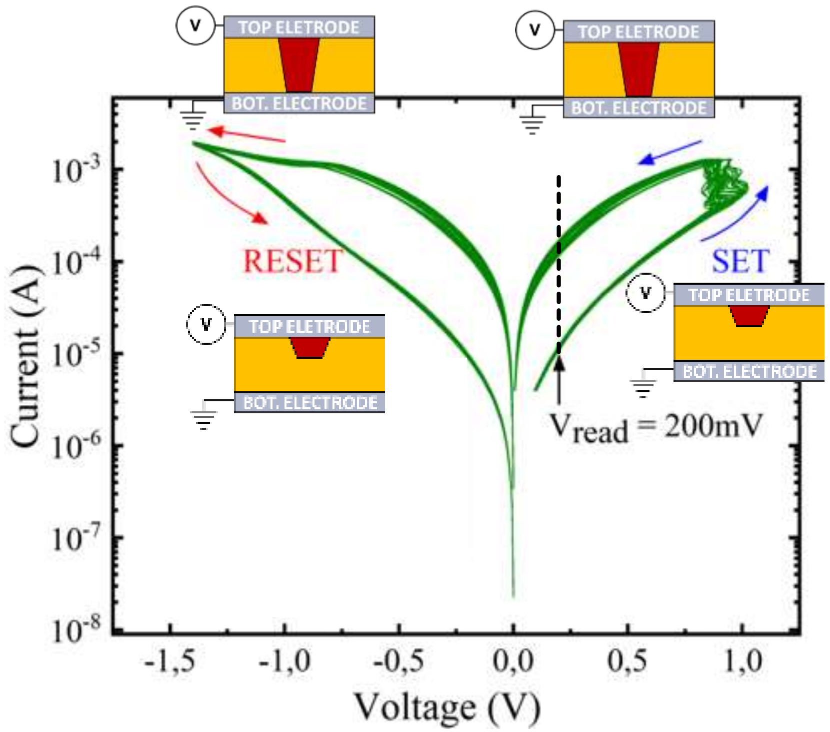
Figure 5. Long-term potentiation and depression for pulses with constant amplitudes V_{pot} and V_{dep} (as defined in the measurement scheme) and increasing duration t_{width}. The inset shows the cycle-to-cycle variation.

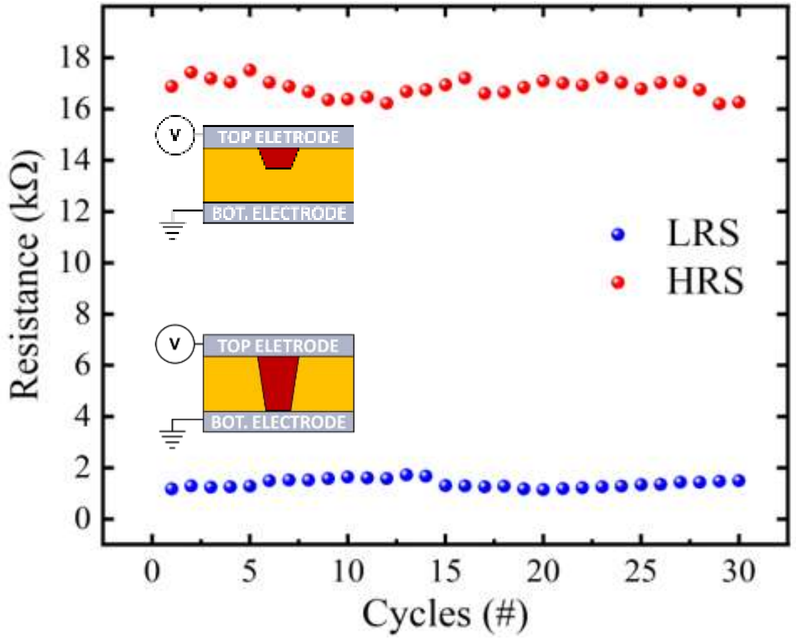
Bégon-Lours *et al.* Adv. Elec. Mat. (2022)
https://doi.org/10.1002/aelm.202101395.

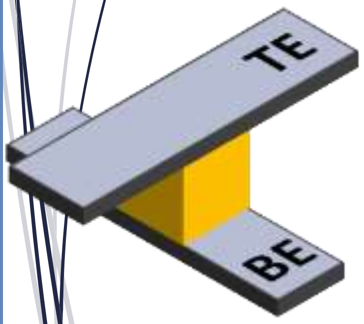
Material/devices: TiOx optimization

Top SEM view

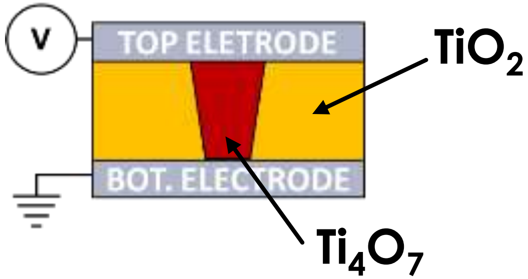




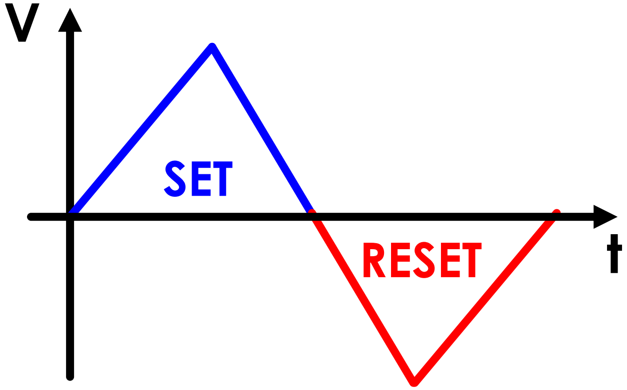




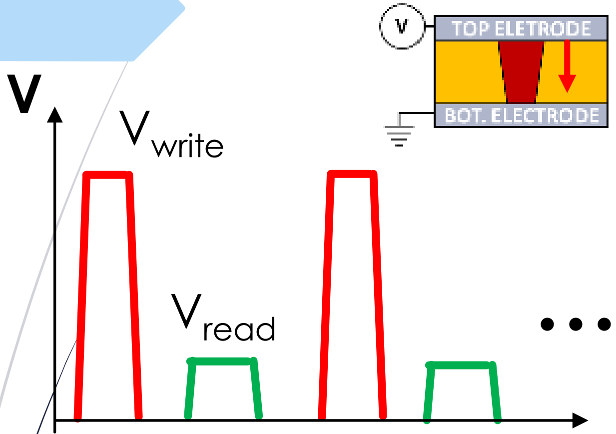
3D view

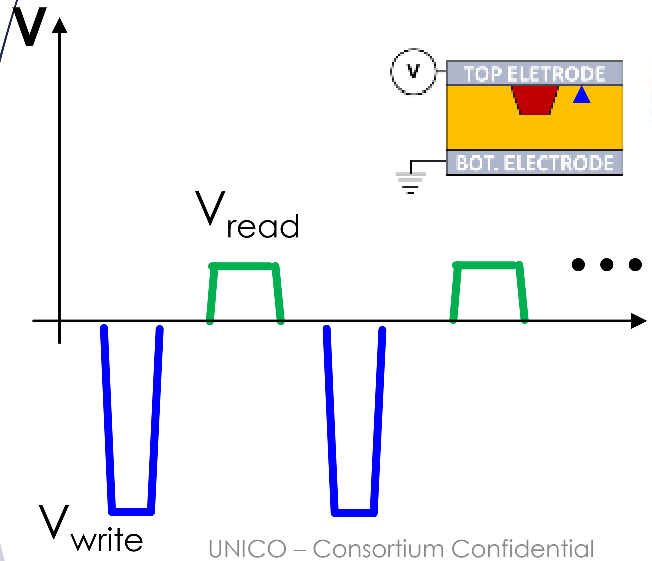


Side view

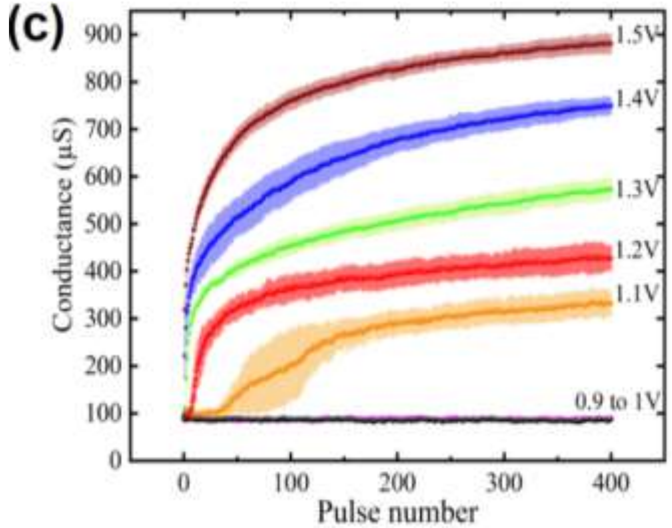


Material/devices: TiOx optimization

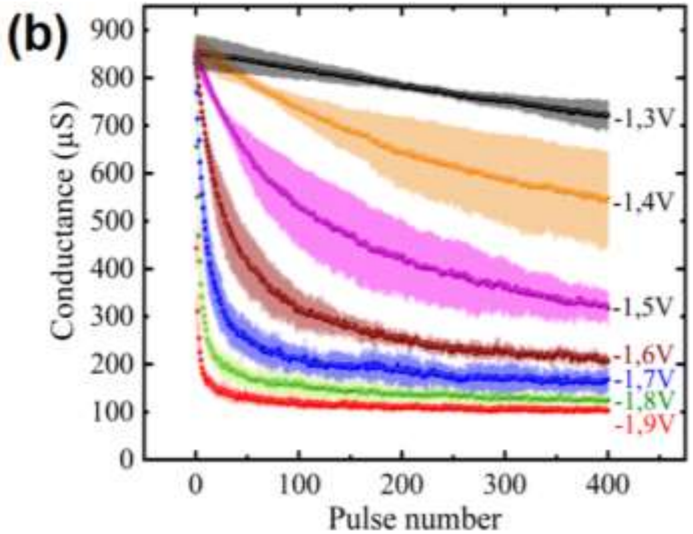




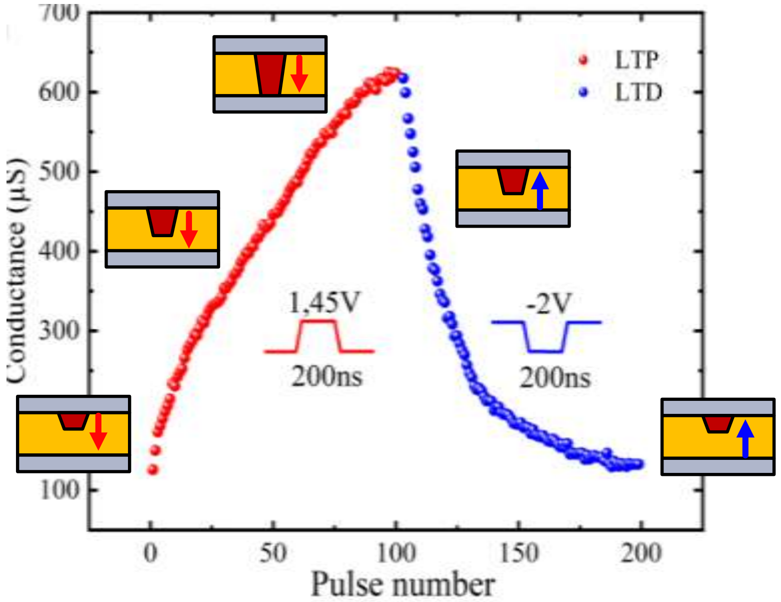
Long-term potentiation (LTP)



Long-term depression (LTD)



LTP followed by LTD: Synaptic plasticity



Conductance-to-weight:

$$W = \frac{g - g_{min}}{g_{max} - g_{min}}$$

chist-era

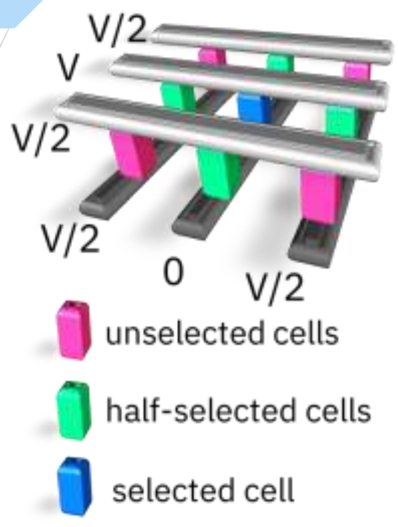
Devices: benchmarking

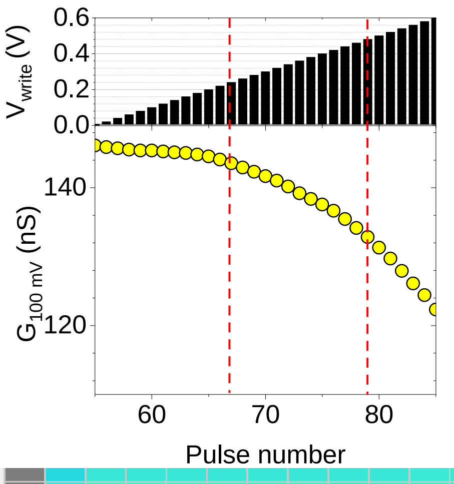
Comparison of key metrics of the devices developped in UNICO

	RTO FeFET	ALD FeFET	FinFeFET	FTJ 5 nm	FTJ 4 nm	FTJ 3 nm	IBM TiOx	3IT TiOx	RRAM MOx
	Halter ACS Appl. Mater. Interfaces (2020)	*Halter, under review*	*Falcone, Frontiers (2022)*	*Bégon-Lours JEDS (2021)*	*Bégon-Lours Adv. Elec. Mat. (2022)*	*Bégon-Lours Neur. Comp. & Eng. (2022)*			
Dynamic Range	1,9	**10 to 60**	1,4	7,2	7	1,3	2 to 10	2 to 20	3
Ron	>100 kOhms	10 Mohms to 500 Mohms	1 MOhms	1 GOhms	10 MOhms	10 MOhms	100 Ohms	1 kOhms	1-100 kOhms
N states	20	>100	>100	>100	>100	>100	10	>100	20
Prog time	50 ns to 5 us	10 ns to 1 s	50 ns to 5 us	10 ns to 1 s	10 ns to 1 s	10 ns to 1 s	800 ns	~100 ns	200 ns to 1 us
Prog Voltage	3 V	6 V	5 V	3 V	2 V	**1 V**	1.2 V	1,5 V	2 V
Write energy	0.5 fJ	1.2 pJ	1 fJ	1 pJ	1 pJ	1 pJ	1 nJ	10-100 pJ	200 pJ
Area	2E1 um^2	1E-1 um^2	**1E-3 um^2**	1E4 um^2	1E2 um^2	1E1 um^2	1E-2 um^2	1E-2 um^2	1E-3 um^2
D-to-D	100%	39%	50%	<2%	<2%	<2%			20%
C-to-C	1%	1,90%	5,00%	<2%	<2%	<2%	10%		5%
Endurance	8,00E+06	>1e10	NC	>1e10	>1e10	>1e10			>1e6

Circuit: HZO passive crossbar integration

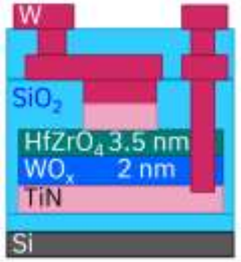
T1.3: Integration of HZO into crossbar arrays:

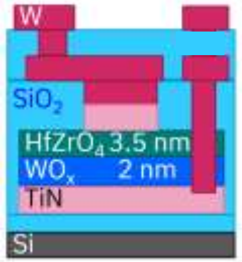


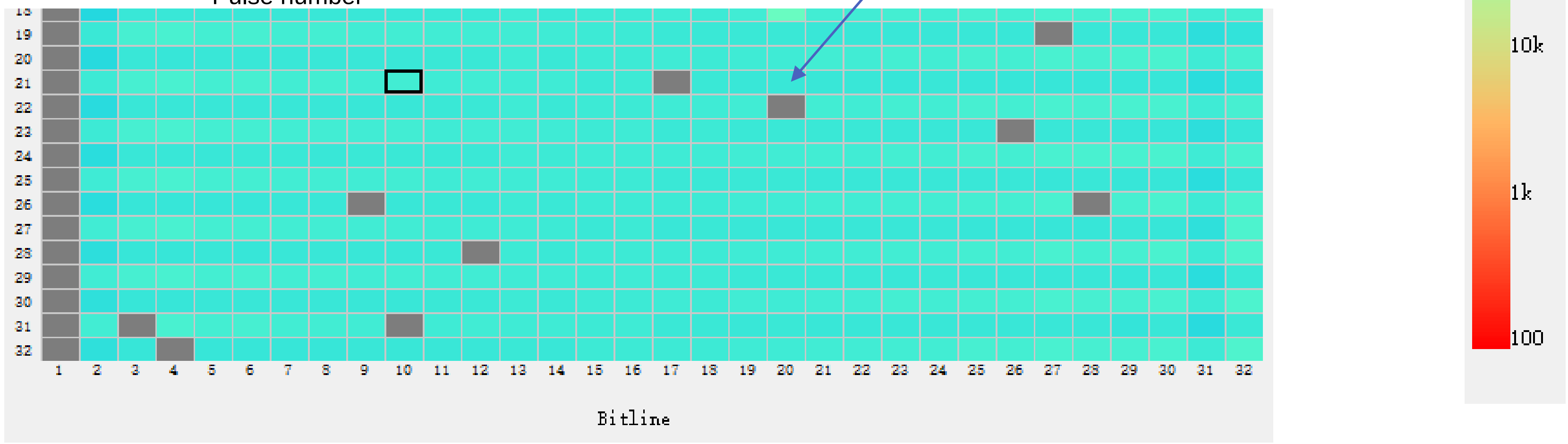


Half-selection scheme reduces the available dynamic range: half-selected cells can be disturbed if $V_{write}/2$ higher than E_C

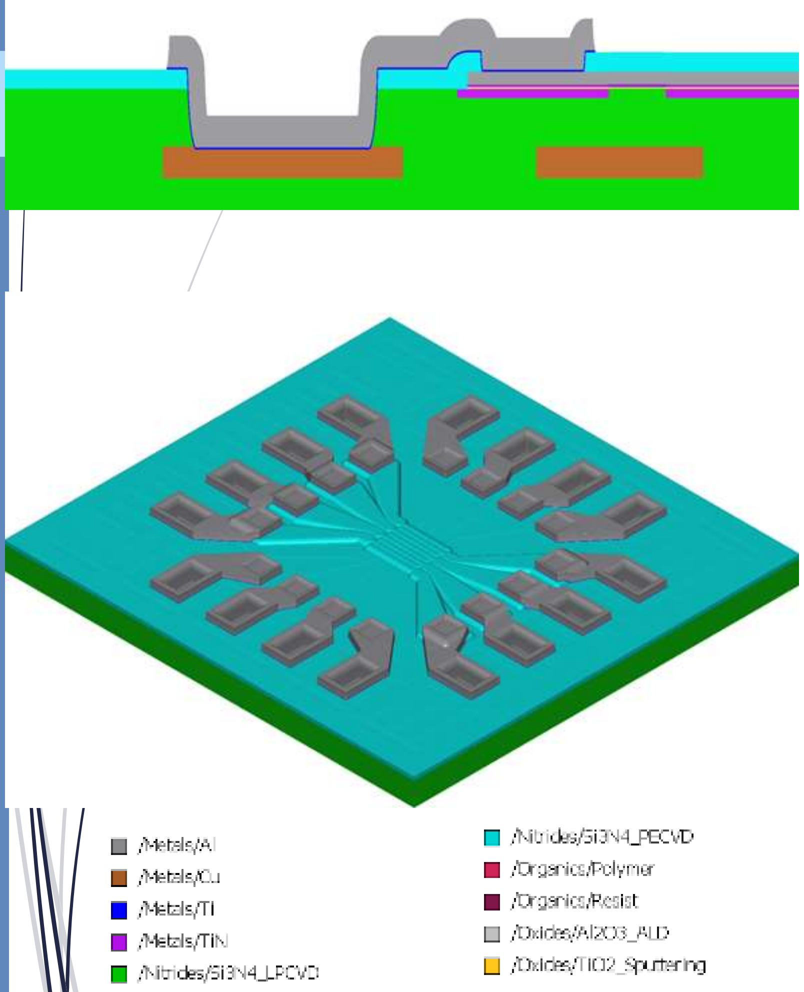
Array with access pads to disconnect shorted devices at M2



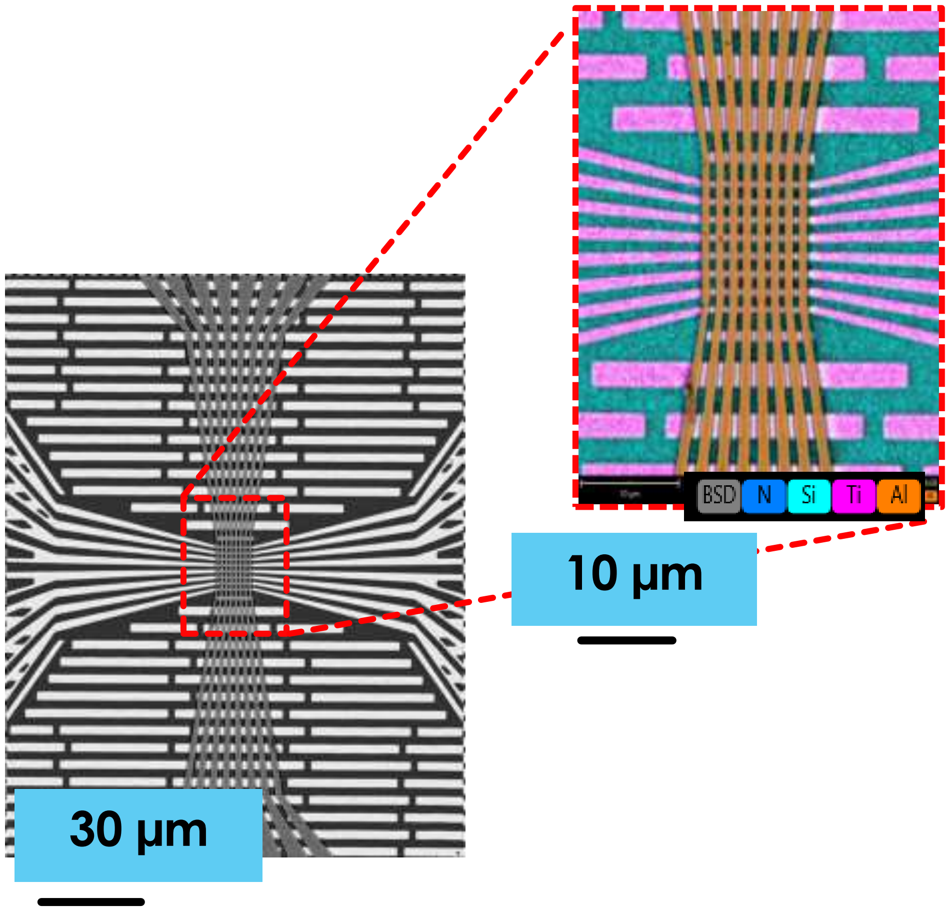


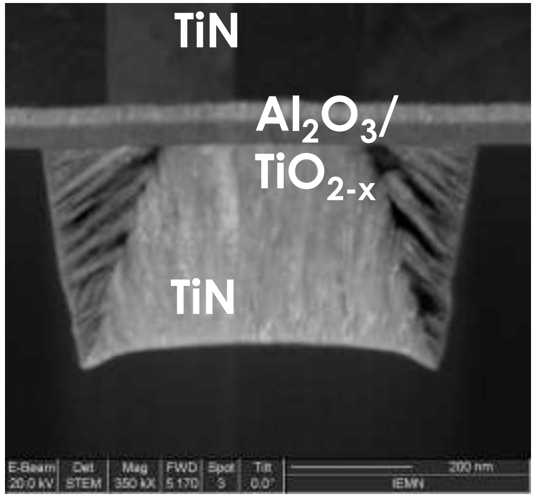


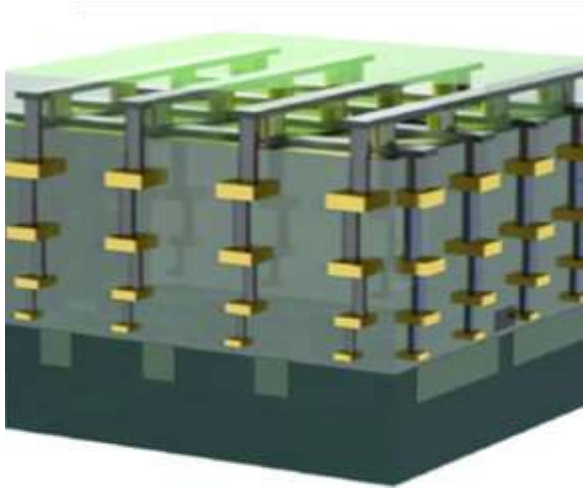
Circuit: BEOL integration



Development of a CMOS compatible fabrication process



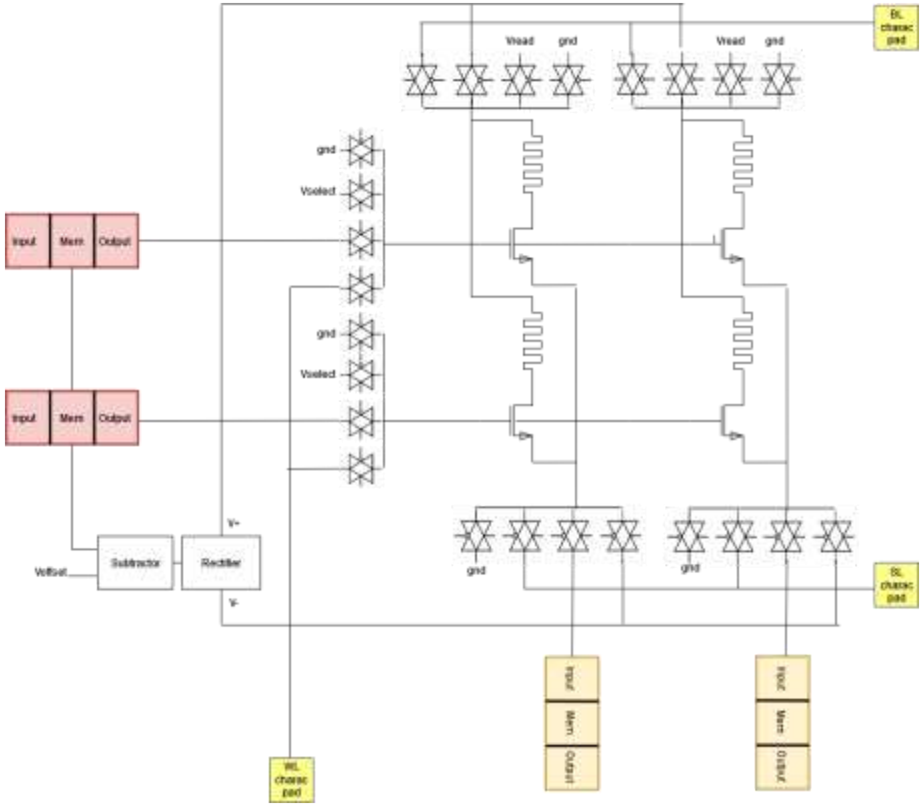


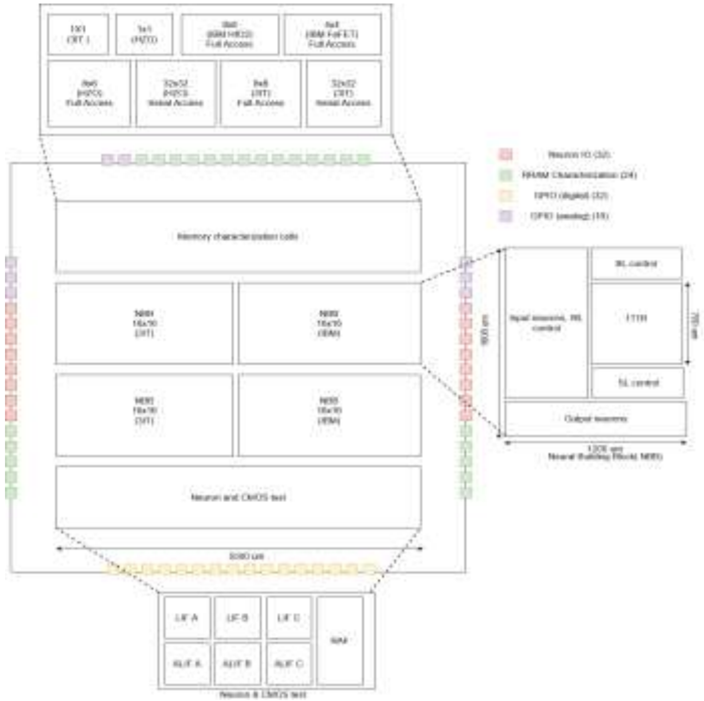


chist-era

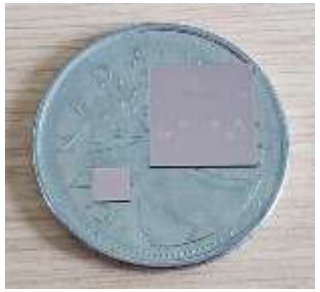
Circuit: CMOS design

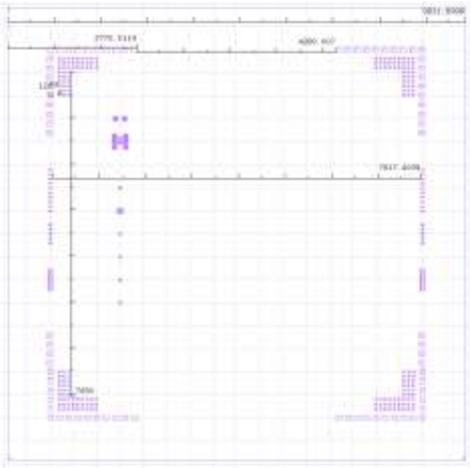
Design of LIF neurons and 1T1R arrays in 130 nm CMOS for BEOL of memristors





130 nm node

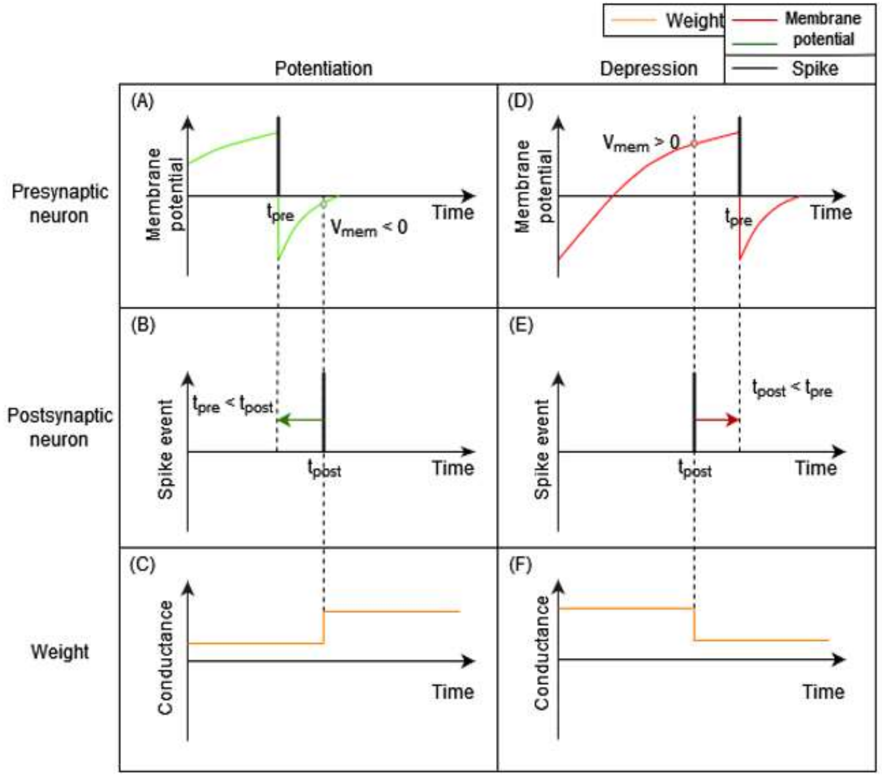


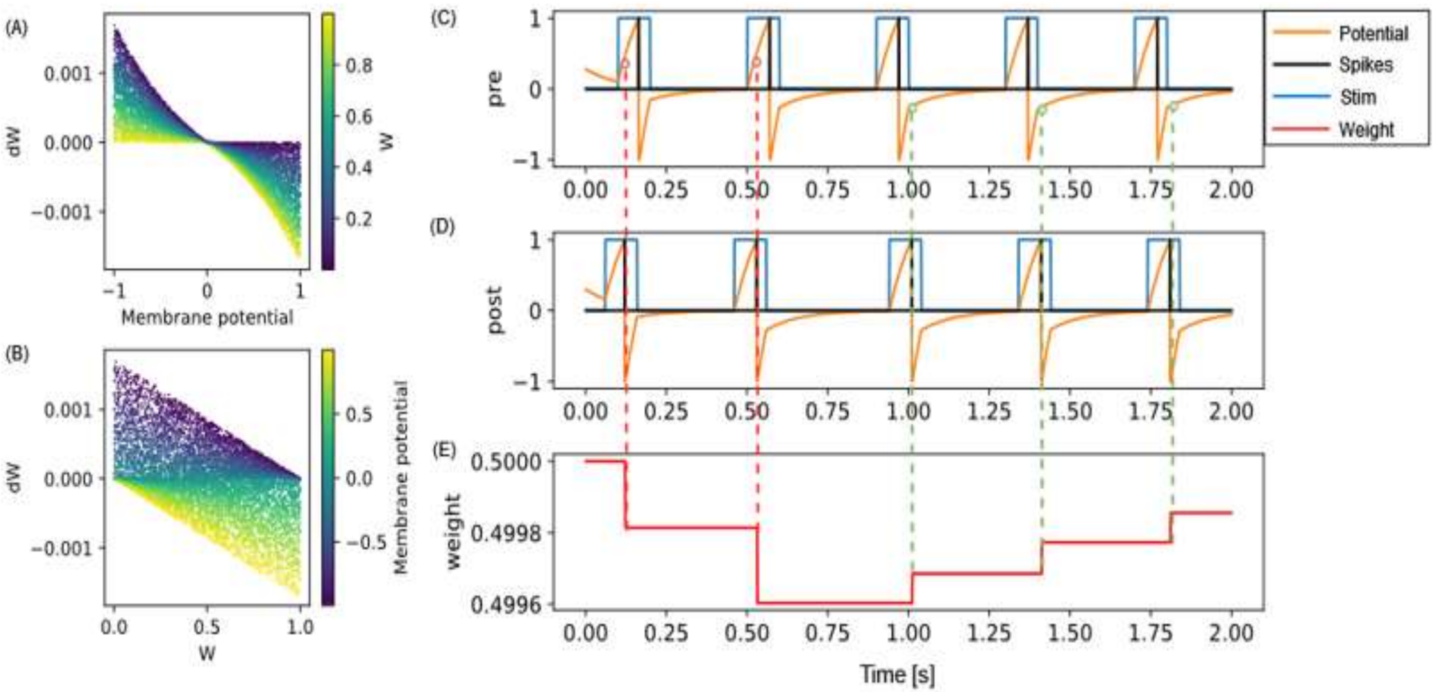


Plan:
2 chips of 6*6 mm² for SNN (04/22 & 04/23) for SNN demonstration

Hardware friendly learning algorithm: VDSP (1)

- We propose an adaptiion of STDP that solves major constraints of material implementation. VDSP relies on membrane potentiual instead of timing and reproduce key STDP features, while material implementation is largely simplified

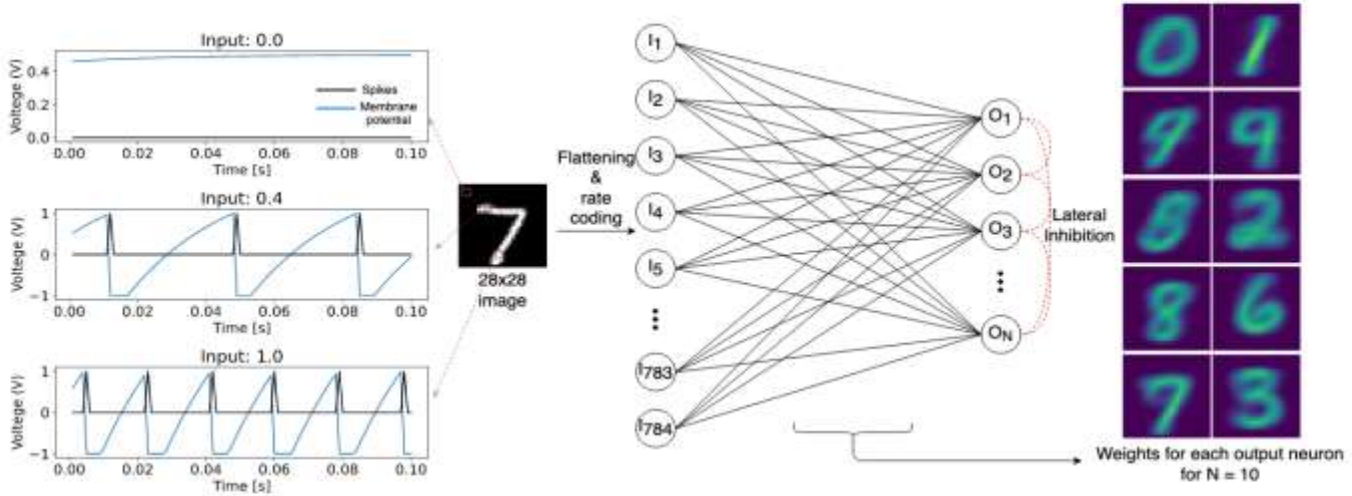


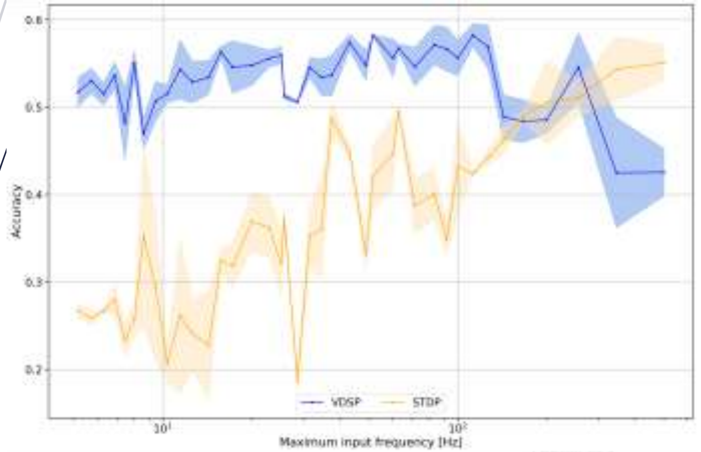


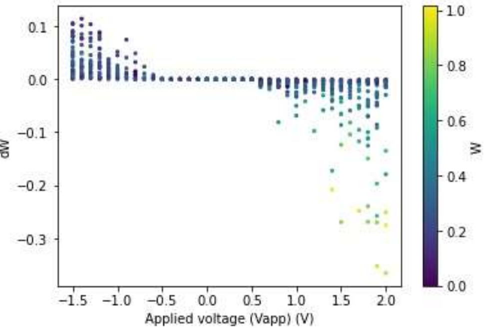
Hardware friendly learning algorithm: VDSP (2)

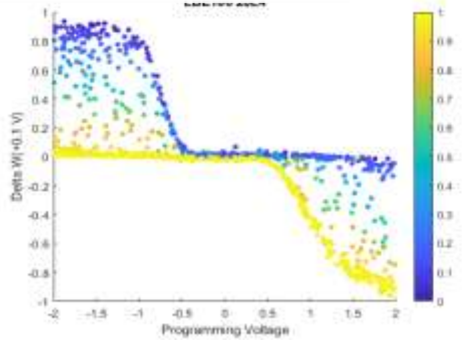
Evaluation of VDSP performances on MNIST classification:

- SotA performances
- Frequency adaptation
- Compatible with Memristor variability



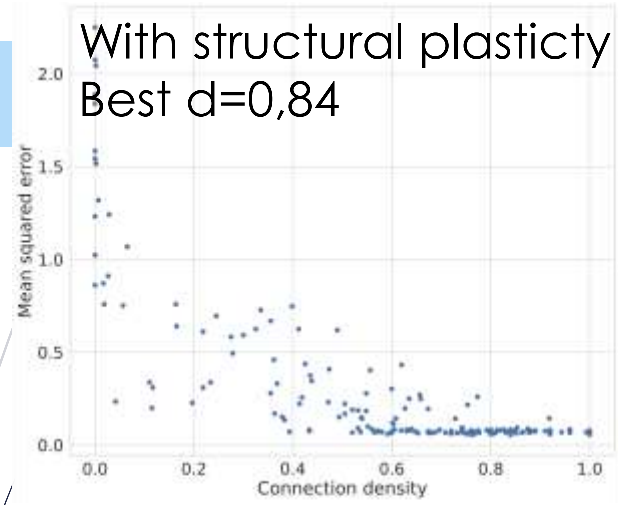


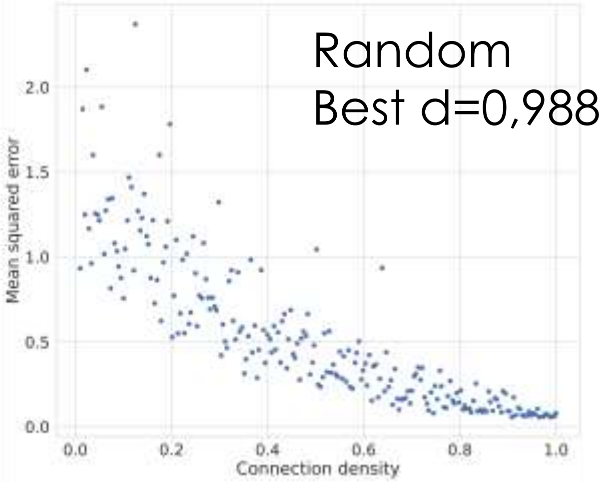




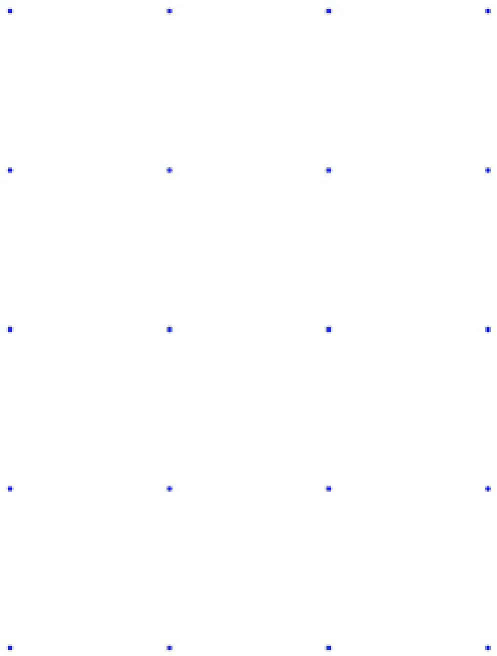
This work			Past studies			
Neurons	Epochs	Accuracy (%) **μ ± σ**	Neurons	Epochs	Accuracy(%)	Ref
10	1	61.4 ± 0.78	10	1	60	(Querlioz et al., 2013)
50	1	78.84 ± 1.28	50	1	76.8	(Guo et al. 2019)
50	3	81.3 ± 1.76	50	3	77.2	(Boybat et al. 2018)
100	3	84.74 ± 1.08	50	3	81	(Querlioz et al., 2013)
100	5	85.01 ± 0.76	100	3	82.9	(Diehl and Cook, 2015)
400	3	89.26 ± 0.54	400	3	87	(Diehl and Cook, 2015)
400	5	89.93 ± 0.41				

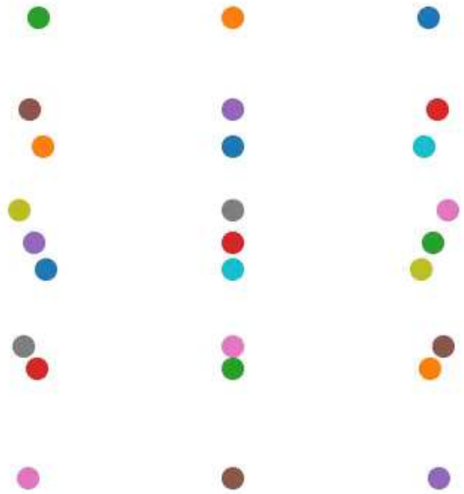
SNN topologies: structural plasticty



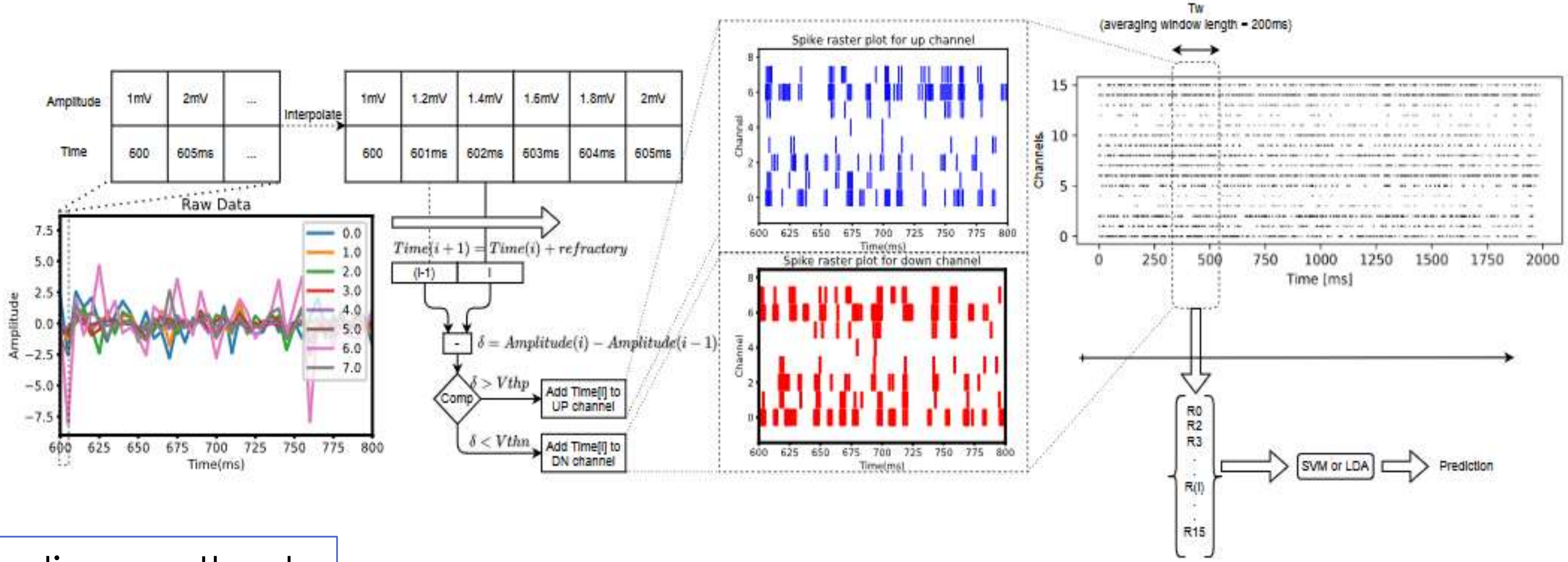


Defining the building block topologies: can we take inspiration of neuron genesis for optimal SNN definition?

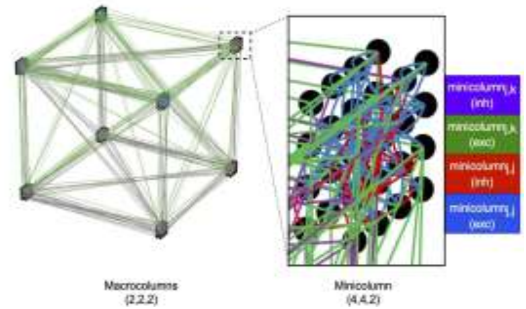


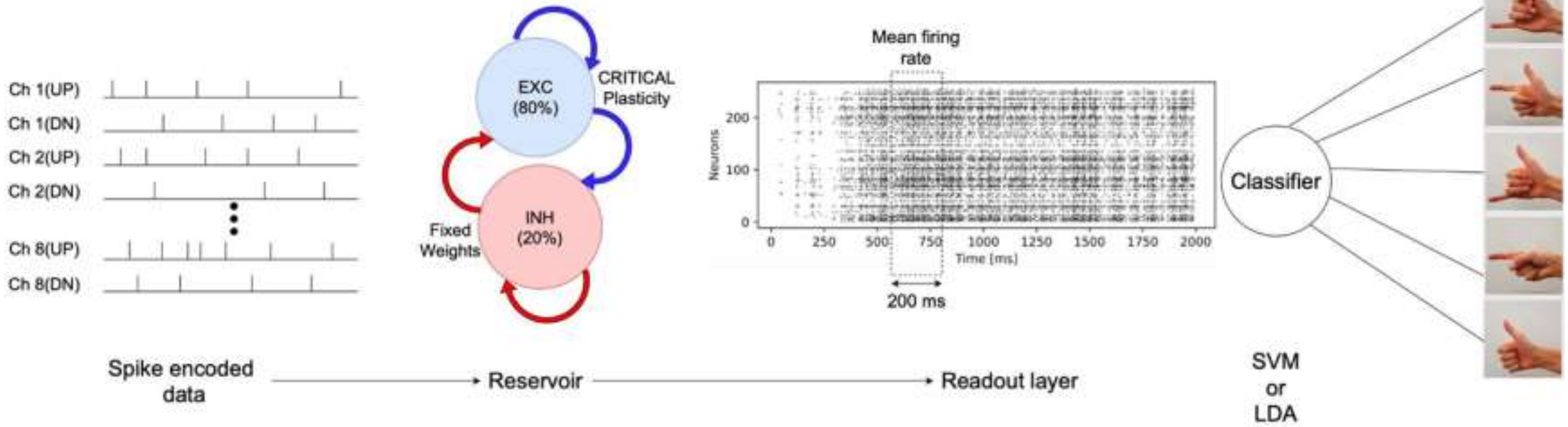


SNN applications : EMG

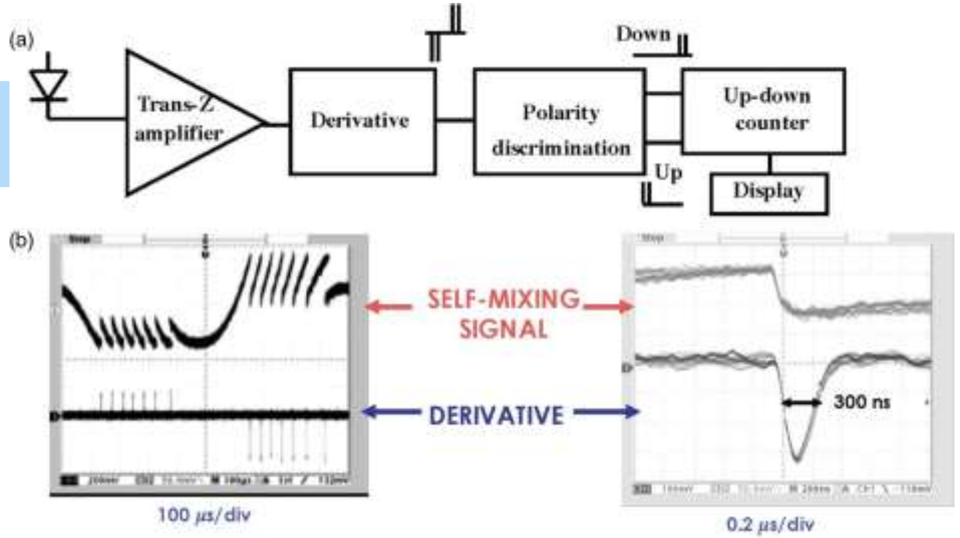


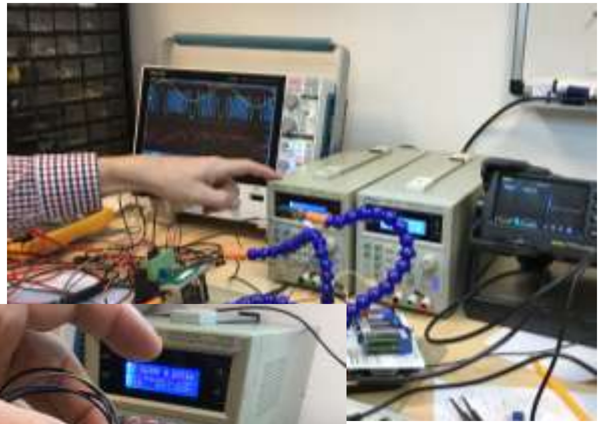
- Optimization of spike encoding methods
- SotA performances for EMG dataset
- Reservoir optimization

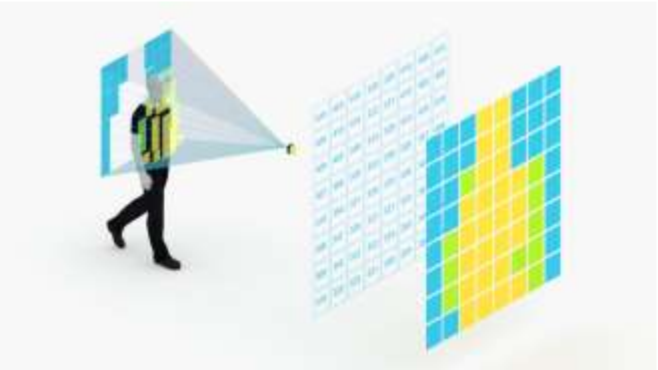




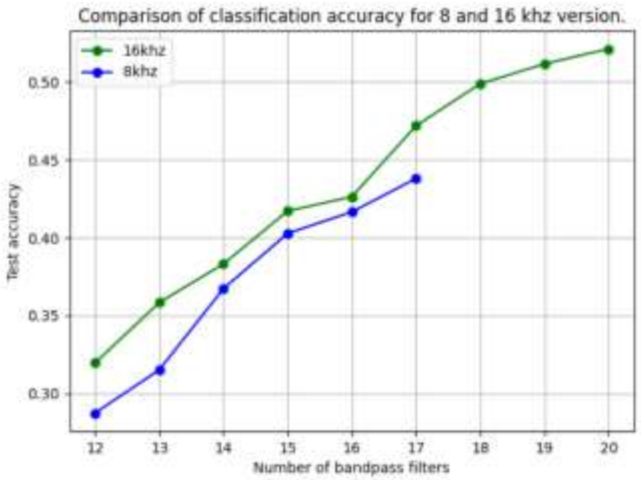
SNN applications : ToF sensors







- Spike encoding in-sensor
- Circuit overhead design and optimization
- Benchmarking on audio signal (through HP membrane mouvment)



"Audiomnist" dataset consists of 30000 audio samples of spoken digits (0-9) of 60 different speakers

Next steps

- Finalization of CMOS chip for neuron / synapses validation
- Fabrication of memristor (TiO2, HZO) in the BEOL
- Spiking ToF sensor in hardware demo
- Diversification of bio-signals (EEG, ECoG dataset under investigation) for benchmarking
- Chip 2 design for BB demonstration

Dissemination / production

Scientific production:

- HZO: Adv. Elect. Mat 2022
- TiO2: MNE 2022
- EMG: ICONS 2021
- VDSP: Frontiers (in prep 2022)

Data sharing:

- TiO2 modeling in open access through repository
- Datasets and benchmarking in open access

Conferences:

- ICONS
- D. Querlioz, "Equilibrium Propagation: a Road for Physics-Based Learning", Keynote presentation, Third International Workshop on Theoretical and Experimental Material Computing, Helsinki, Finland, October 2021
- D. Querlioz, "Equilibrium Propagation: a Road for Physics-Based Learning", virtual 84. Annual Meeting of DPG and DPG-Tagung (DPG Meeting) of the Condensed Matter Section (SKM), September 2021.

Thank you for your attention!