De l’IA à la conception d'ASIC

Pietro M. FERREIRA
maris@ieee.org

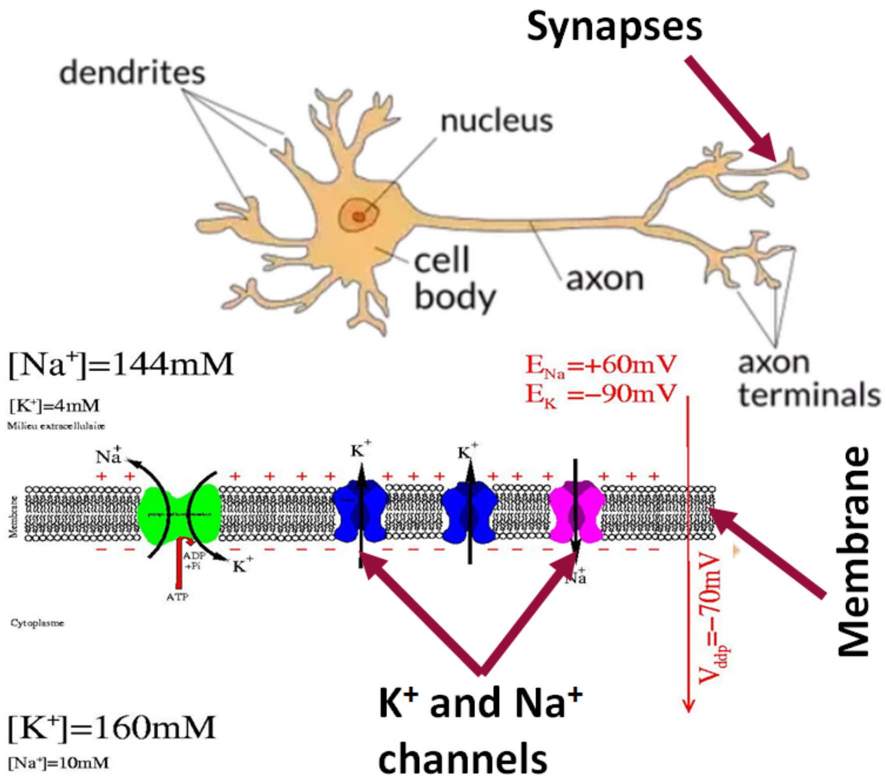
03/03/2026

Agenda

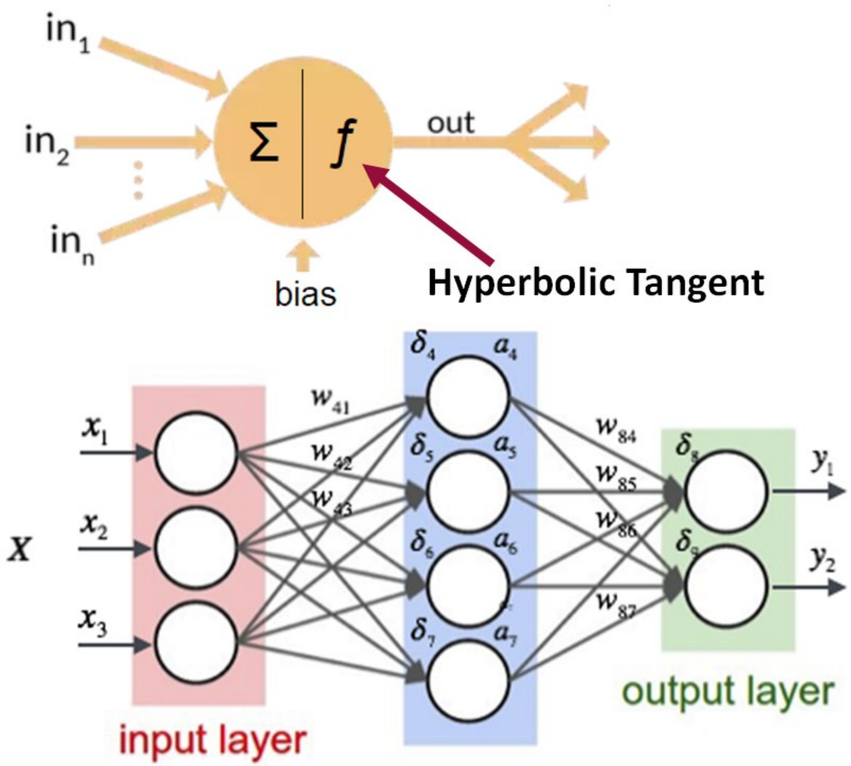
- Research Context and Challenges
- How to design Neuromorphic Circuits
- Fundamentals in Neuromorphic Circuits
- Physical Implementation
- Neuromorphic Circuit and Systems Illustrations

Natural & Artificial Intelligence

Biological solutions: Ion conduction, Living cells



Software solutions: Binary based, Von Neumann Computer



Electronics & Artificial Intelligence

Neuromorphic Circuits

- Achievements
  - ❑Edge AI
  - ❑Real Time
  - ❑Digital Circuits + Memory
- Perspectives
  - ❑Less Silicon surface
  - ❑Low Power
  - ❑Energy Efficiency

Cloud Computing

- Achievements
  - ❑Deep-Neural Networks
  - ❑High processing capabilities
  - ❑Easy to use tools
- Perspectives
  - ❑Novel architectures
  - ❑Bigger and deeper models
  - ❑NN properties

Cloud Artificial Intelligence Sustainability

- Data center energy consumption: 240-340 TWh in 2022
  - ❑1-1.3% of total electricity demand (IEA, 2025)
  - ❑Google queries: 0.3 Wh per request
- France total generation: 475 TWh in 2022
  - ❑nuclear 295 TWh (62%) (World Nuclear Association, 2025)
- Energy Consumption of Cloud AI
  - ❑GPT4.0: 2.9 Wh per request
  - ❑100M queries per day (EpochAI, 2025)
  - ❑Total: 109.5 TWh in a year for ChatGPT (only)

Research Challenges

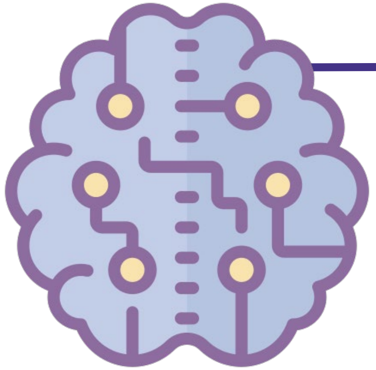
Goal: Propose a Low-Power edge-AI solution in IoTs

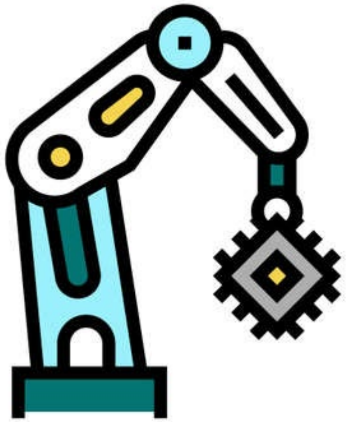
- Hardware-based Spiking Neural Networks
- From human (vision and hearing) to IoT (electromagnetic field, temperature gradient, pressure/acceleration) cognition
- Information to Digital Converters
- Minimize communication data to AI cloud

Agenda

- Research Context and Challenges
- How to design Neuromorphic Circuits
- Fundamentals in Neuromorphic Circuits
- Physical Implementation
- Neuromorphic Circuit and Systems Illustrations

Hierarchical Design Flow





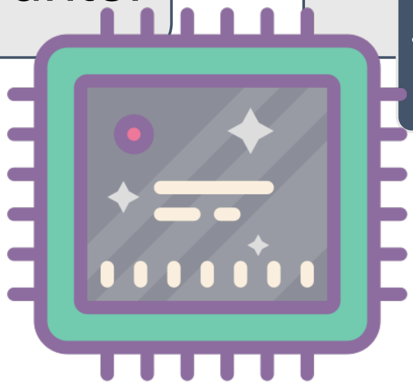
Scientific Methodology in Electronics Research

System-level

DNN, CNN,
FNN, SNN

Circuit-level

eNeuron,
Synapses,
Spike counter



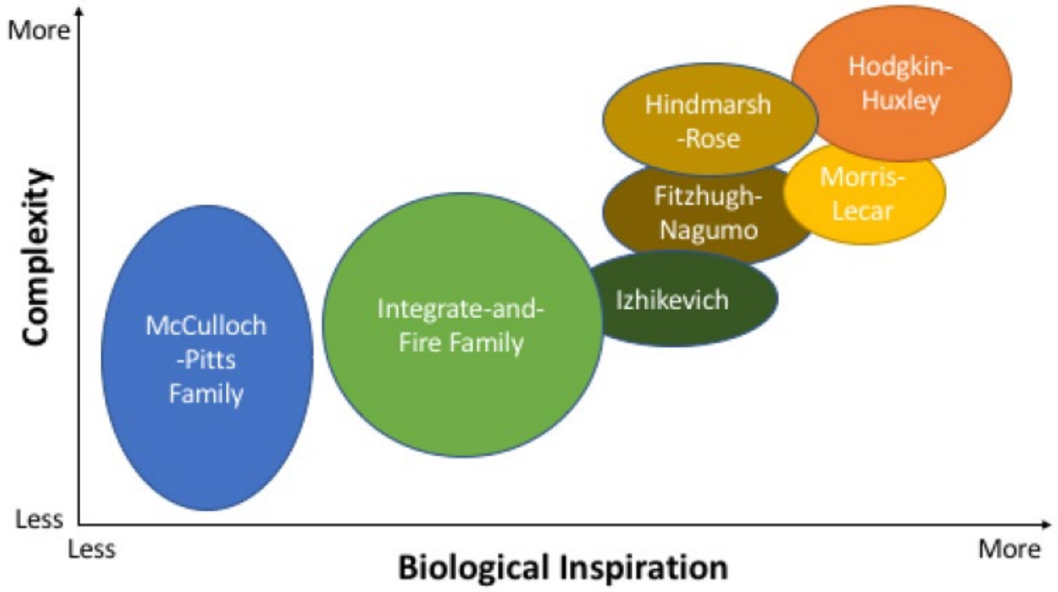
Transistor-level

ML_bio
ML_simple
LIF

Process-level

65 nm or 28 nm
f_T > 320 GHz
$I_{on}/I_{off} > 10^8$
1 fF @1.4 µm x 55 nm

Literature Review: eNeuron



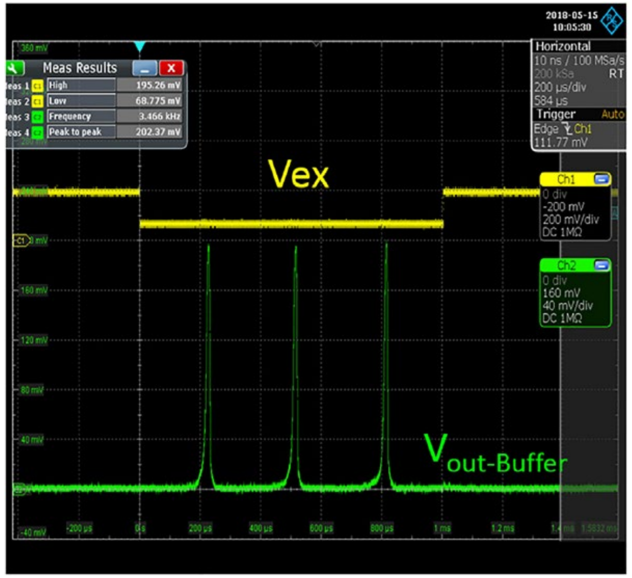
http://arxiv.org/abs/1705.06963

- ~~**McCulloch** – Digital (software or FPGA)~~ ✗
- **Leak-Integrate and Fire (LIF)** – Analog and Digital (first order diff eq.)
  - ❑Simple MatLab Model ✓
  - ❑Charge/discharge of a capacitor ✓
  - ❑Less biomimetic ✗
- **Izhikevich** – Analog or Digital
  - ❑Biomimetic ✓
  - ❑Simple MatLab Model ✓
  - ❑Discontinuous model, and hard to be implemented in low-power ✗
- **Morris-Lecar (ML) or Hodgkin-Huxley** – Analog only
  - ❑Biomimetic ✓
  - ❑Complex MatLab Model (tanh) ✗
  - ❑Low-power implementation ✓

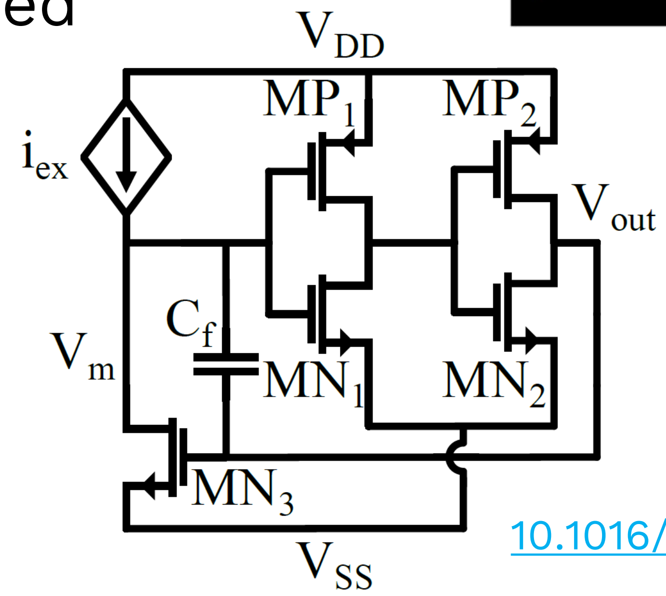
Which one is the best option?

10.3389/fnins.2011.00073
10.1109/NEWCAS52662.2022.9841988

Leak-Integrate and Fire eNeuron



- How does it work?
  - ❑Two-inverter cascade (INV1, INV2)
  - ❑Current i_{ex} is integrated in C_f
  - ❑Transistor MN3 discharge V_m for a V_{out} threshold
  - ❑Leaky parasitic devices are considered



- Variant with smaller leakage and lower dynamic range

$$\tau_m \frac{dv(t)}{dt} = -v(t) + s(t) + n(t)$$

$$s(t) = \begin{cases} 1\,, & if\ v(t) > 1 \\ 0, & else \end{cases}$$

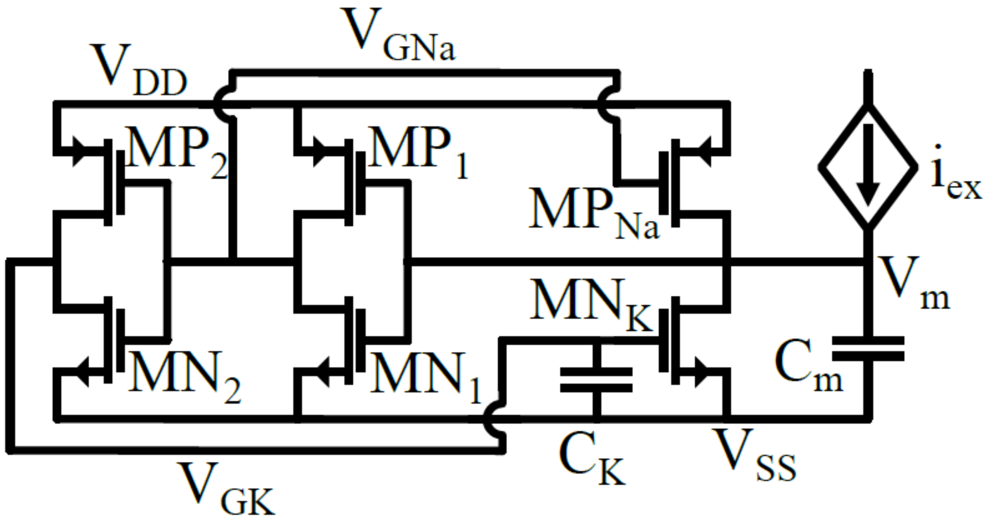
10.1016/j.sse.2019.01.002

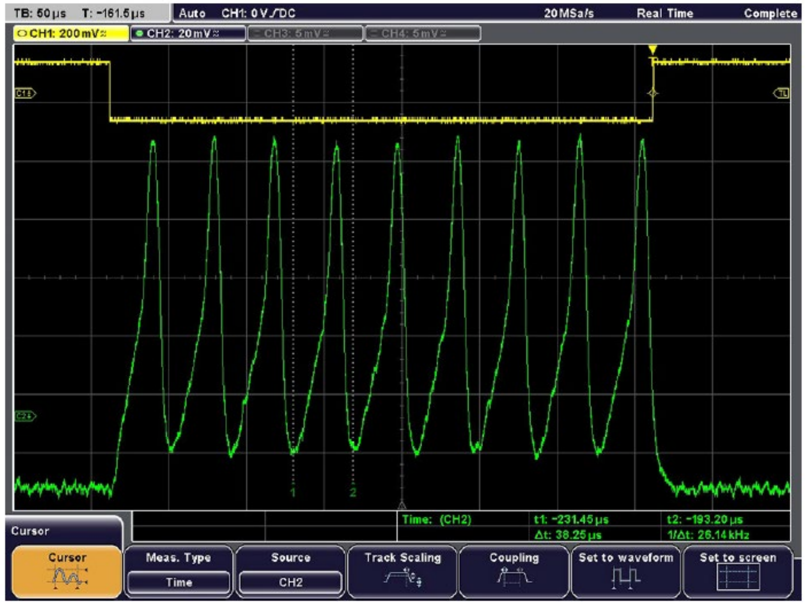
Morris-Lecar eNeuron

- How does it work?
  - ❑Double INV1/INV2 time constant control both
  - ❑Na charges the V_m
  - ❑K discharges the V_m
  - ❑Topology is like LIF simplified (not biomimetic)

- Variant reducing area and power consumption
  - ❑Increase f_{spike} strategy
  - ❑Do even better in 28 nm

10.1109/NEWCAS52662.2022.9842088

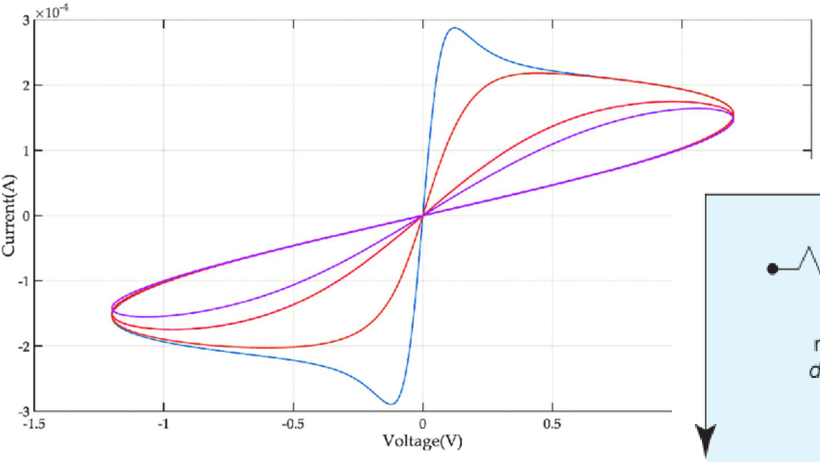


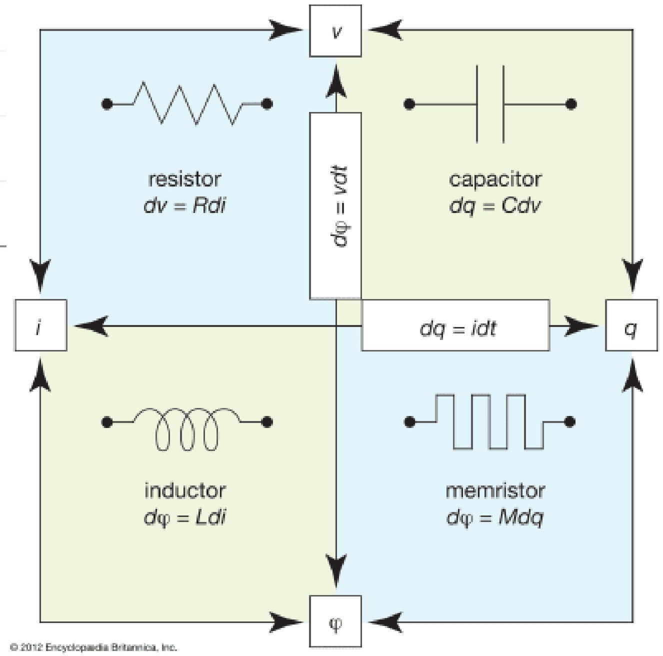


10.3389/fnins.2017.00123

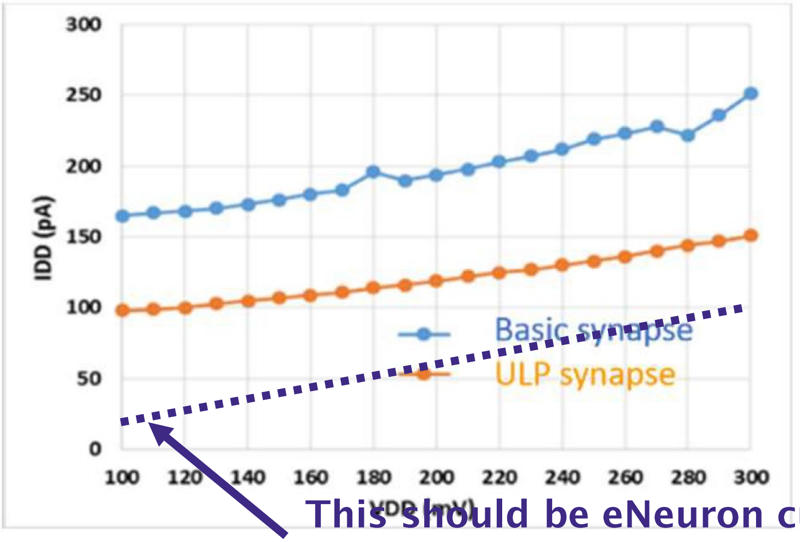
Literature review: eSynapse

- Memristors – not CMOS enable technology (ie. high





- CMOS alternative
  - ❑Simple current mirrors [10.3389/fnins.2011.00073](https://doi.org/10.3389/fnins.2011.00073)
  - ❑Application specific – offline learning [10.1109/CBMI50038.2021.9461899](https://doi.org/10.1109/CBMI50038.2021.9461899)



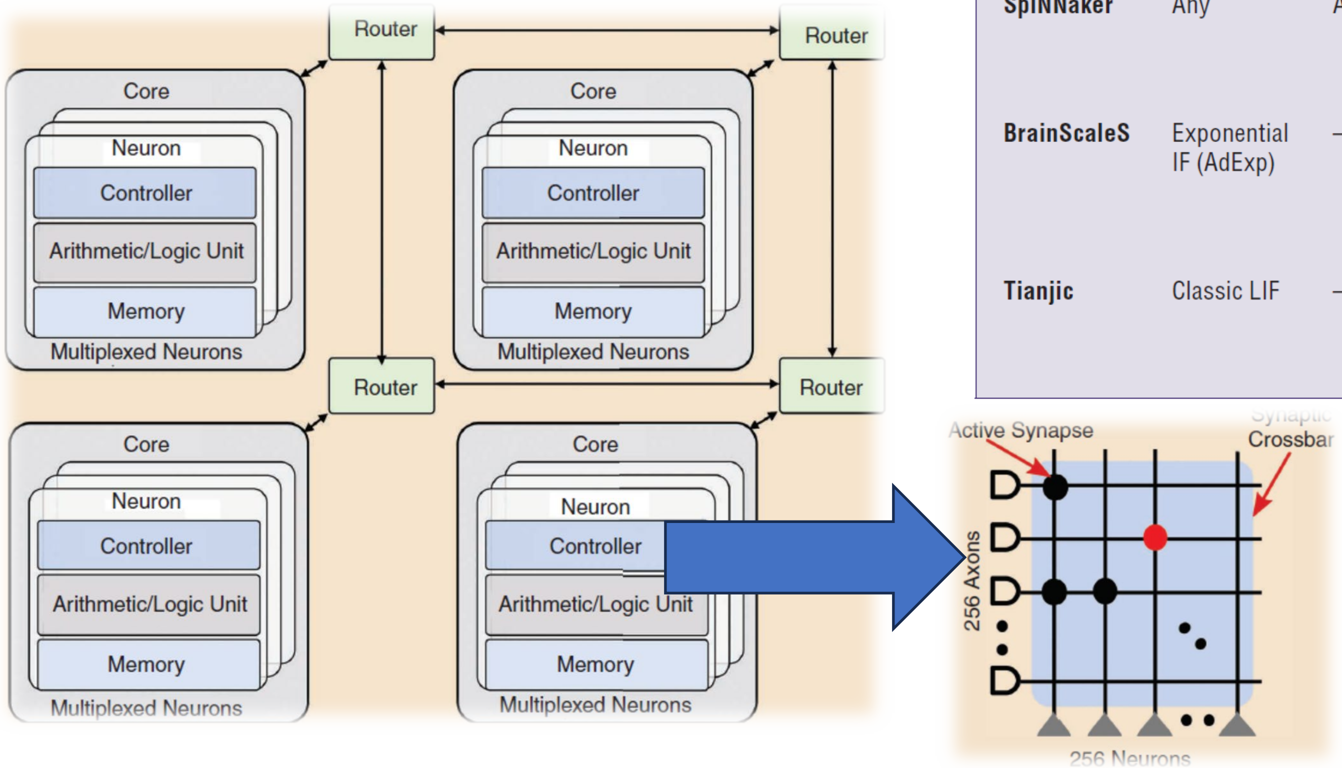
This should be eNeuron current consumption of same authors

eNN Architectures

Digital implementation

Table II.
Design Choices in the Large-Scale Neuromorphic Systems.

	Neuron	Synapse	Implementation Choice	Architecture	Software Support
TrueNorth	Classic LIF	Binary with a choice of four 8-bit weights	Digital	256x256 crossbar per core, 64x64 core array	Matlab-based object-oriented Corelet language
Loihi	CUBA LIF	Variable precision weight, allows PSP with exponential kernel filter	Digital	No crossbar, 1048 neurons per core, 128 cores	Python-based API NxSDK, also supported by Nengo
SpiNNaker	Any	Any	Digital, Multiprocessor SOC	—	SpiNNaker API, PyNN, Nengo, NEST, Brian, sPyNNaker
BrainScaleS	Exponential IF (AdExp)	—	Analog/Mixed signal	A wafer with 56x8 ANC containing ANNCORE, each of which has 128k synapse and 512 membrane circuits	PyNN
Tianjic	Classic LIF	—	Digital	2D mesh many-core with 156 Fcores, each with 32 weight index and 256 fan-ins/fan-outs	—



10.1109/MCAS.2022.3166331

State of the Art Comparison

https://doi.org/10.1109/MCAS.2022.3166331

Table III.
Comparison of the Large-Scale Neuromorphic Systems.

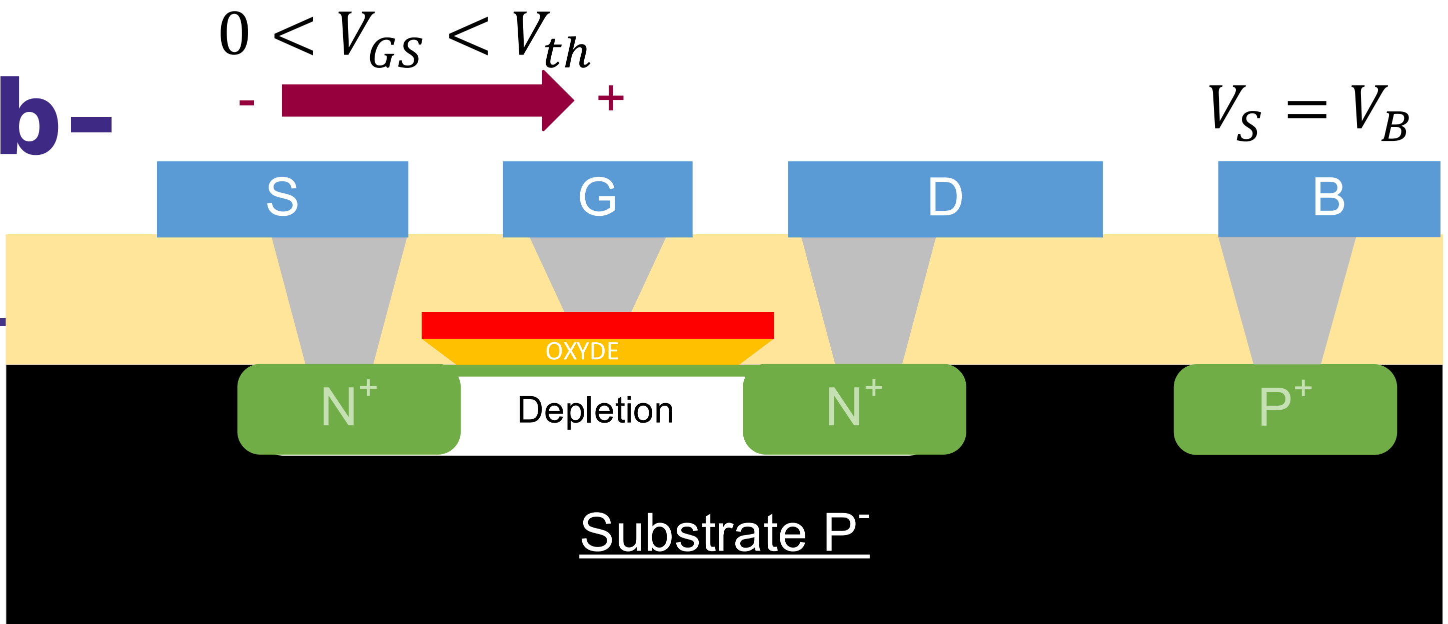
Neuromorphic Chip	TrueNorth	SpiNNaker	Loihi	BrainScaleS	Neurogrid	Braindrop	Dynap-SEL	Tianjic
Implementation	Digital	Digital	Digital	Analog	Analog	Analog	Mixed-signal	Digital
Technology	28 nm CMOS	ARM968 130 nm CMOS	14 nm CMOS	180 nm CMOS	180 nm CMOS	28 nm CMOS	180 nm CMOS	28 nm CMOS
# transistors	5.4 B	100 M	2.07 B	15 M	23 M			
Neurons per Core	256	~1k	max 1024	8 to 512	65k	4096	1024	16
Synapses per Core	256x256	~1M	~16k	~130k	100M	64k	64k	22k
Cores per Chip	4 096	16	128	352	16	16	4	156
Chip Area (mm^2)	430	102	60	50	168		43.79	14.44
Energy/SOP (pJ)	26	10000	23.6	100	100	0.38	17	0.95
NoC	2D mesh unicast	2D mesh multicast	2D mesh unicast	Hierarchical	Tree multicast	Tree Multicast	Hierarchical 2D mesh multicast	Hierarchical 2D mesh multicast
Packet Size (bits)	32	40 + optional 32		30	12		20	
Time	Discretized	Discretized	Discretized	Discretized	Real time	Real time	Real time	Real time
Neuron Update	Time Multiplexed	Time Multiplexed	Time Multiplexed	Real time	Real time	Real time	Real time	Real time
Bio-Plausibility	Low	Medium	Medium	High	High	High	High	Low
Simulation Time	1x to 21x real-time	Real-time	>Real-time but variable	104x to 105x	Real-time	70 MHz clock	Real-time	300 MHz
On-Chip Learning	No	Yes	Yes	Yes	No	Yes	Yes	No

Agenda

- Research Context and Challenges
- How to design Neuromorphic Circuits
- Fundamentals in Neuromorphic Circuits
- Physical Implementation
- Neuromorphic Circuit and Systems Illustrations

cro
ma

NMOS Physics : Sub-threshold Region



S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. New York: Jonh Wiley 'Sons, 1981.

Low Power Consumption

- Electric Field : $G \xrightarrow{\vec{E}} S$ pull electrons from the source
- Begin of a charge inversion, channel is formed

- Charge conduction :

$$\Delta Q = q\Delta n_p e^{(V_{GS}-V_{th})/\phi_t}$$

Δn_p – Minority carriers are injected into the channel

$\Delta n_p \approx 0$ if the channel is long ($L > 300\ nm$), so $\Delta Q \approx 0$

All-region “interpolation” model of the MOSFET

Unified current control model

$$I_{WI} = \eta g_m \phi_t \xrightarrow{\text{interpolation}} I_{SI} = \frac{n g_m^2}{2\mu_n C_{ox}(W/L)}$$

For $g_m \to \infty$, I_{WI} is negligible; while for $g_m \to 0$, I_{SI} is negligible

10.1109/NEWCAS.2018.8585657

10.1109/ACCESS.2022.3198644

$$\boxed{I_D = I_{WI} + I_{SI} = \eta g_m \phi_t \left[1 + \frac{g_m}{2\mu_n C_{ox} \phi_t (W/L)}\right]}$$

$$I_D = I_{WI}\left[1 + \frac{(W/L)_{th}}{(W/L)}\right] \leftrightarrow g_m = 2\mu_n C_{ox} \phi_t (W/L)_{th}$$

where $(W/L)_{th}$ is the normalized aspect ratio

How to design an eNeuron?

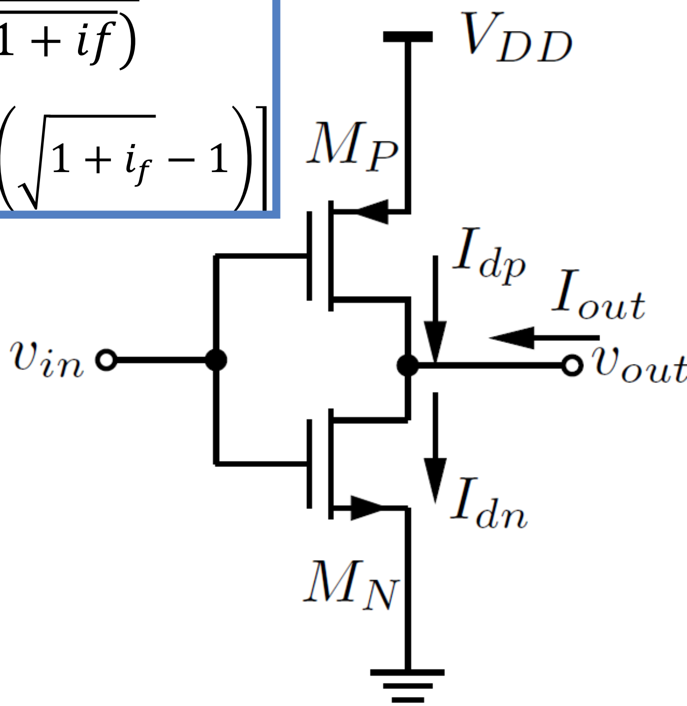
- It is **not** an Inverter
- Learn how to use:
  - ❑Weak inversion bias
  - ❑Unified current control model
  - ❑gm/Id Methodology
- Do not believe 100% in your simulator

https://doi.org/10.1109/NEWCAS.2018.8585657

$$I_d = \eta g_m \phi_T \left[1 + \frac{g_m}{2\mu C_{ox}\phi_T (W/L)}\right]$$

$$\frac{g_m}{I_D} = \frac{2}{\eta\phi_t\left(1+\sqrt{1+if}\right)}$$

$$V_{DSsat} = \phi_t\left[\ln\left(\frac{1}{\xi}\right) + (1-\xi)\left(\sqrt{1+i_f}-1\right)\right]$$



$$I_d = g_m V_{DS} e^{\frac{V_{GS}-V_{th}}{\eta\phi_T}}$$

$$v_{out} = -V_{DD}\cdot\tanh\left[\frac{v_{in}}{\eta\phi_T} + \frac{1}{2}\ln\left(\frac{g_{m,N}}{g_{m,P}}\right)\right]$$

$$Av = \left.\frac{dv_{out}}{dv_{in}}\right|_{v_{in}=v_{out}} = \frac{V_{DD}}{\eta\phi_T}$$

https://doi.org/10.3389/fnins.2017.00123

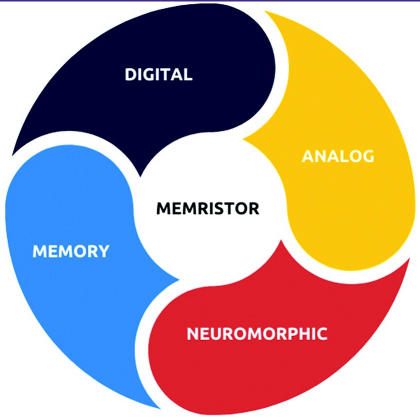
https://doi.org/10.1109/TCSI.2009.2034233

$$I_d = I_s e^{\frac{V_{GS}-V_{th}}{\eta\phi_T}}\left(1 - e^{-\frac{V_{DS}}{\eta\phi_T}}\right)$$

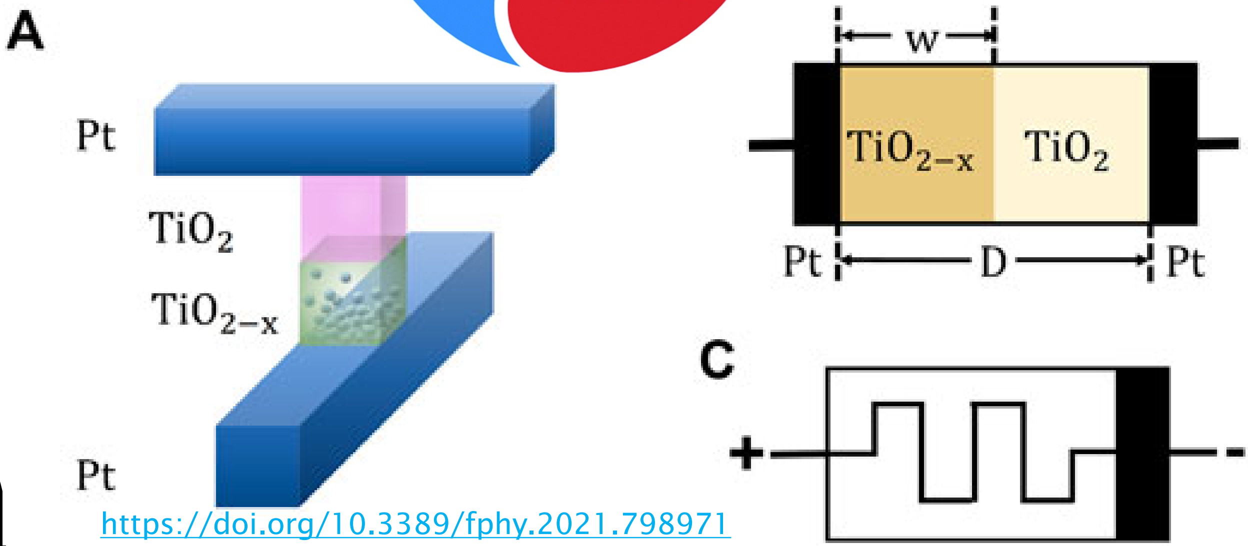
$$I_s = \mu C_{ox}\frac{W}{L}\cdot\frac{\phi_t^2}{2}$$

Best Trade off is $if \approx 1.0$

What is a memristor?



- Device capable to memorize a charge/flux
  - ❑Integrate Current and Voltage
  - ❑Resistance Inertia



https://doi.org/10.3389/fphy.2021.798971

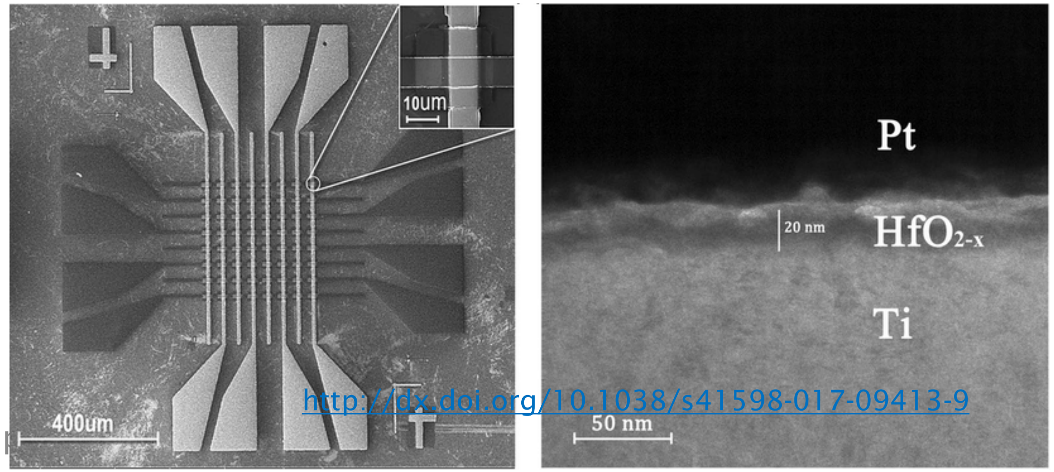
- Property: Memresistance

$$M(Q) = \frac{R_{on} w(t)}{D} + R_{off}\left(1 - \frac{w(t)}{D}\right)$$

where μ is material permeability

- Ohm's law

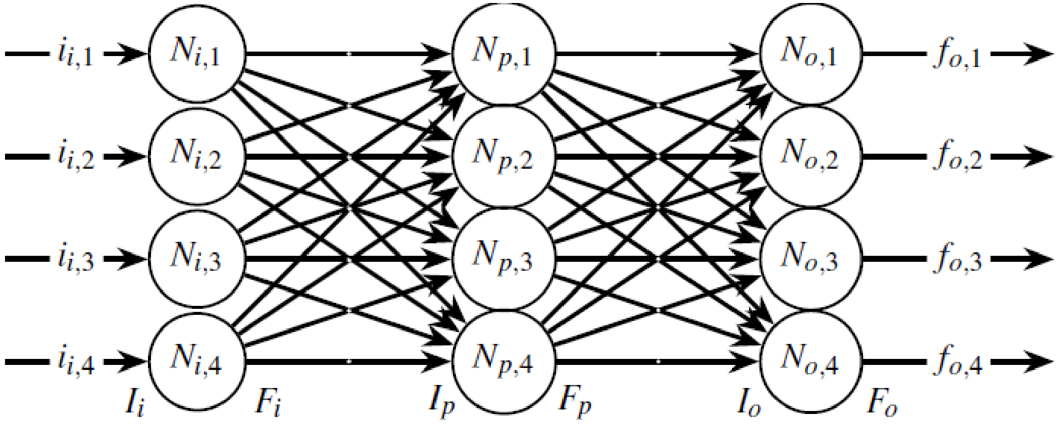
$$\phi = M(Q) \cdot Q$$

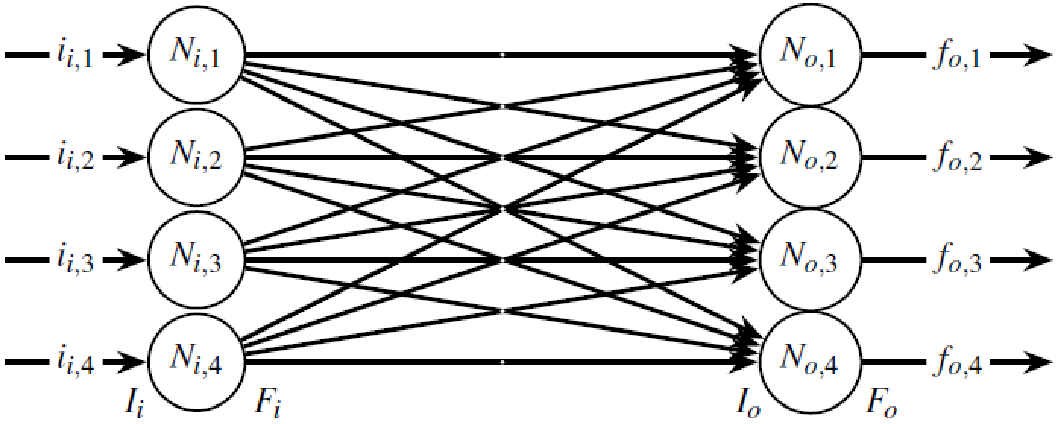


http://dx.doi.org/10.1038/s41598-017-09413-9

Neural Network Property

Nonlinearity is mandatory





$$f_{spike_i} = h(i_{syn_i}) = h(\sum_{k=1}^{n} g(f_{spike_k})),$$

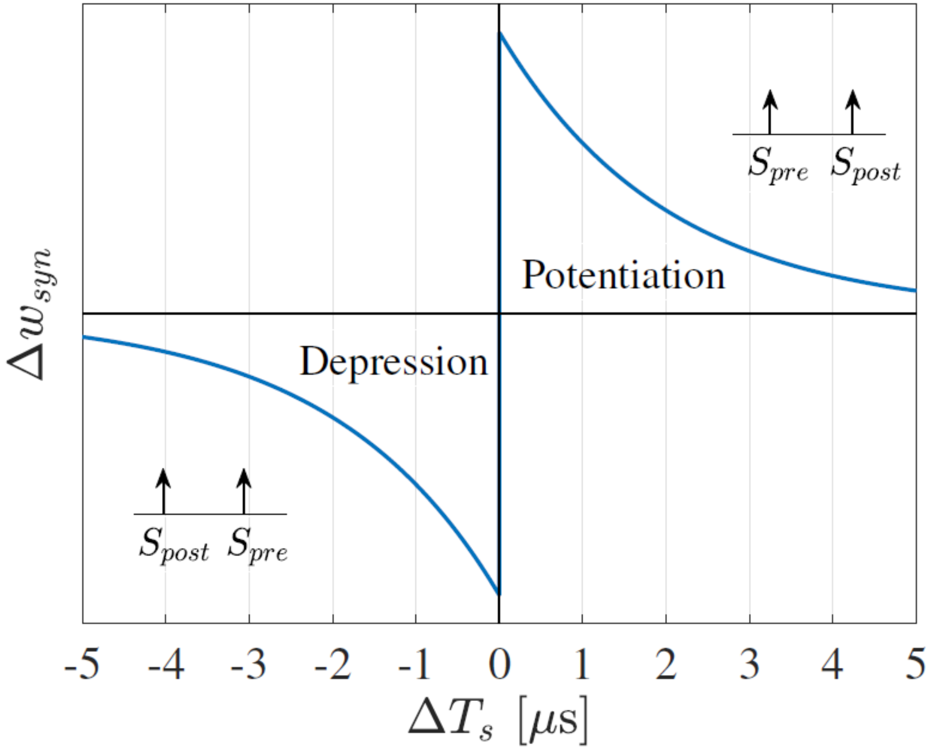
$$= h(\sum_{k=1}^{n} w_{syn_{ki}} \cdot f_{spike_k} + b_1),$$

$$F_N = \prod_{p=1}^{N} H_p C_p \cdot I_1$$

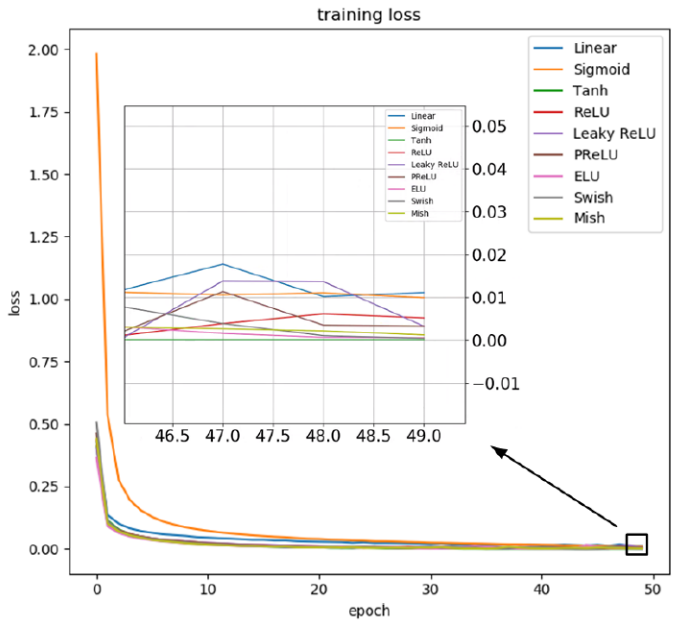
https://doi.org/10.1109/SBCCI55532.2022.9893216

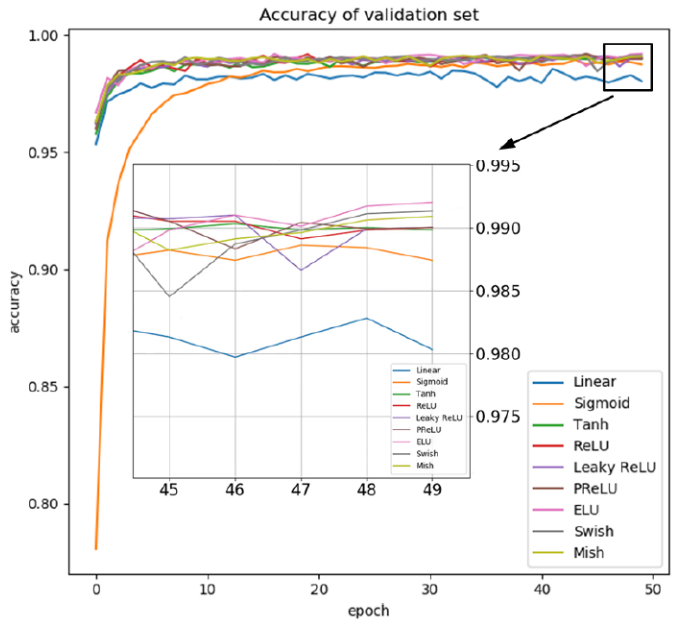
Neural Network Training

STDP Learning



SGD Optimizer





https://doi.org/10.1109/TNNLS.2021.3084827

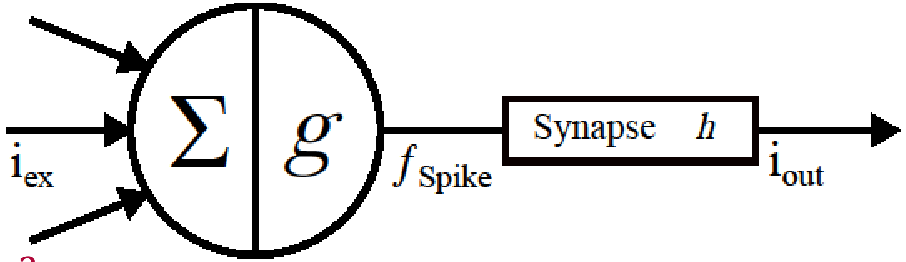
https://doi.org/10.1109/TCSI.2022.3178989
https://doi.org/10.1109/TCSI.2022.3204645

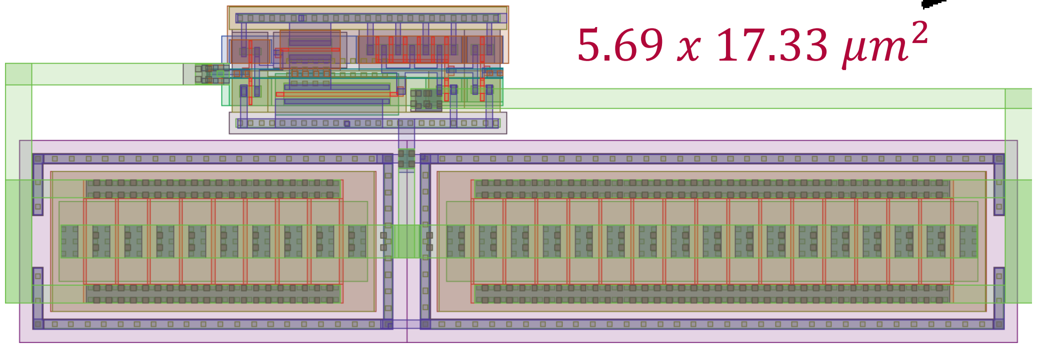
Agenda

- Research Context and Challenges
- How to design Neuromorphic Circuits
- Fundamentals in Neuromorphic Circuits
- Physical Implementation
- Neuromorphic Circuit and Systems Illustrations

eNeuron Physical Design

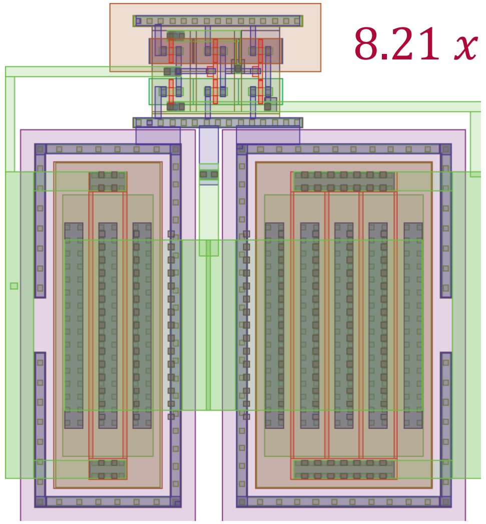
Library of dedicated devices in 55 nm





$5.69\ x\ 17.33\ \mu m^2$

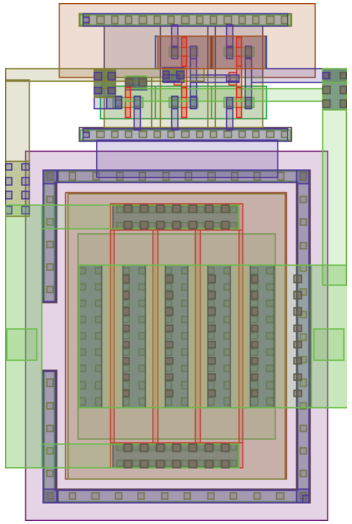
Morris-Lecar Biomimetic



$8.21\ x\ 7.52\ \mu m^2$

Morris-Lecar Simplified

$6.56\ x\ 4.33\ \mu m^2$



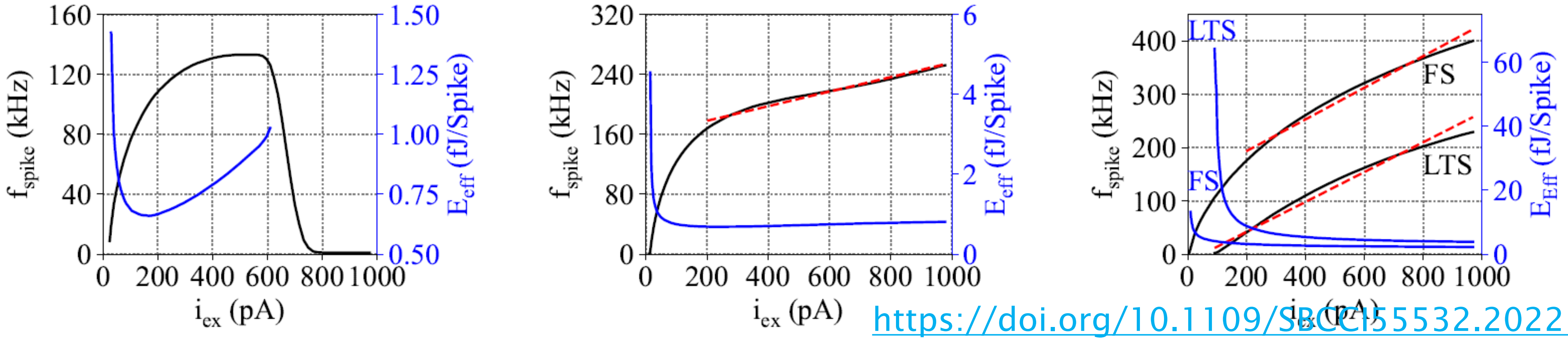
Leak Integrate and Fire

https://doi.org/10.1109/SBCCI55532.2022.9893216

eNeuron Post-Layout Simulations

- In house implementation achieves similar PLS results using 55 nm
  - ❑ E_{eff} as low as 2 fJ/spike
  - ❑ f_{spike} higher than 100 kHz
- Low leakage is the key to this design and most of the challenges are in the layout

https://doi.org/10.1007/s10470-020-01729-3



https://doi.org/10.1109/SBCCI55532.2022.9893216

eNeuron Performance Comparison

Ref.	Model	Techn. (nm)	Area (μm^2)	f_{spike} (kHz)	E_{eff} (fJ/spike)
https://doi.org/10.1007/s10470-020-01729-3 (Our version)	ML bio.	55	98.6	400	1.95
https://doi.org/10.1016/j.sse.2019.01.002	LIF	65	31	15.6	2
https://doi.org/10.1109/SBCCI55532.2022.9893216 (Our version)	LIF	55	28.4	130	1.2
https://doi.org/10.1109/NEWCAS52662.2022.9842088	LIF	28	34	343	1.2
https://doi.org/10.3389/fnins.2017.00123	ML simp.	65	35	25	4
https://doi.org/10.1109/SBCCI55532.2022.9893216 (Our version)	ML simp.	55	61.7	243	1.8

ML : Morris-Lecar

LIF : Leak Integrate and Fire

$$E_{eff} = P_{rms}/f_{spike}$$

Synapses Physical Design

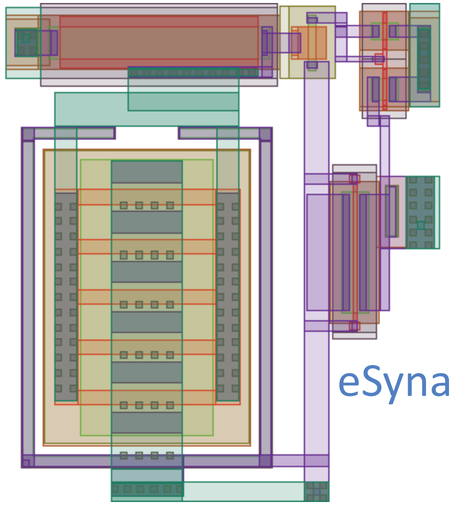
<u>No Memristor, no online learning</u>

- T. Soupizet et al. proposal

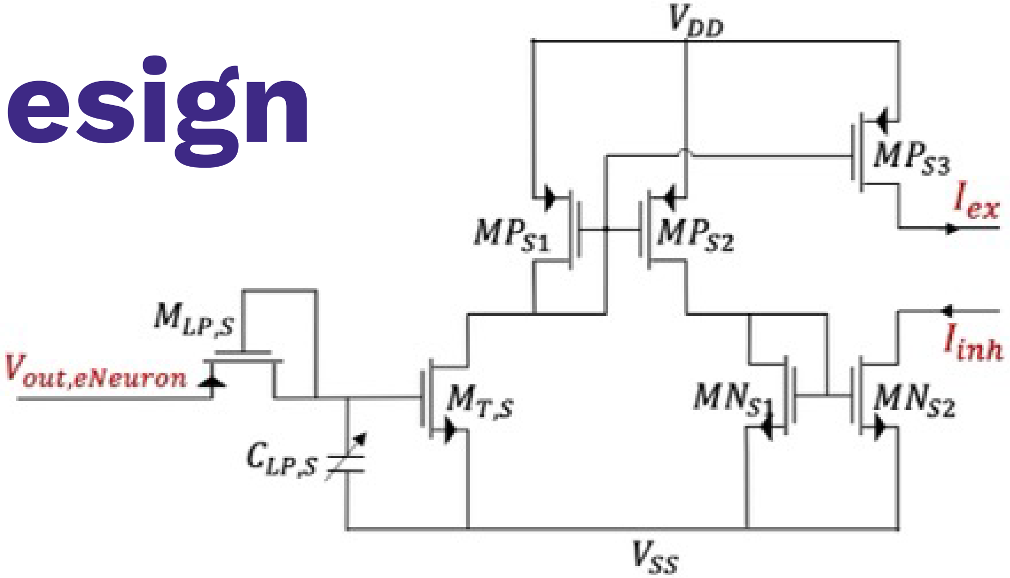
  https://doi.org/10.1109/SBCCI55532.2022.9893216

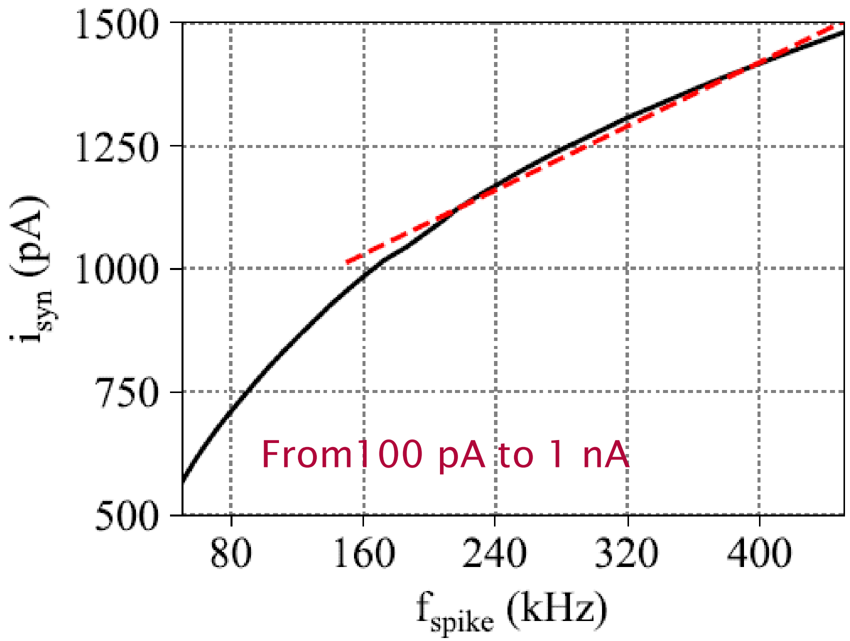
  - ❑Diode-MOS and LPF
  - ❑Simple Current mirrors I_{ex}
  - ❑Dynamic range comparable to eNeurons

$6.60 \ x \ 7.55 \ \mu m^2$

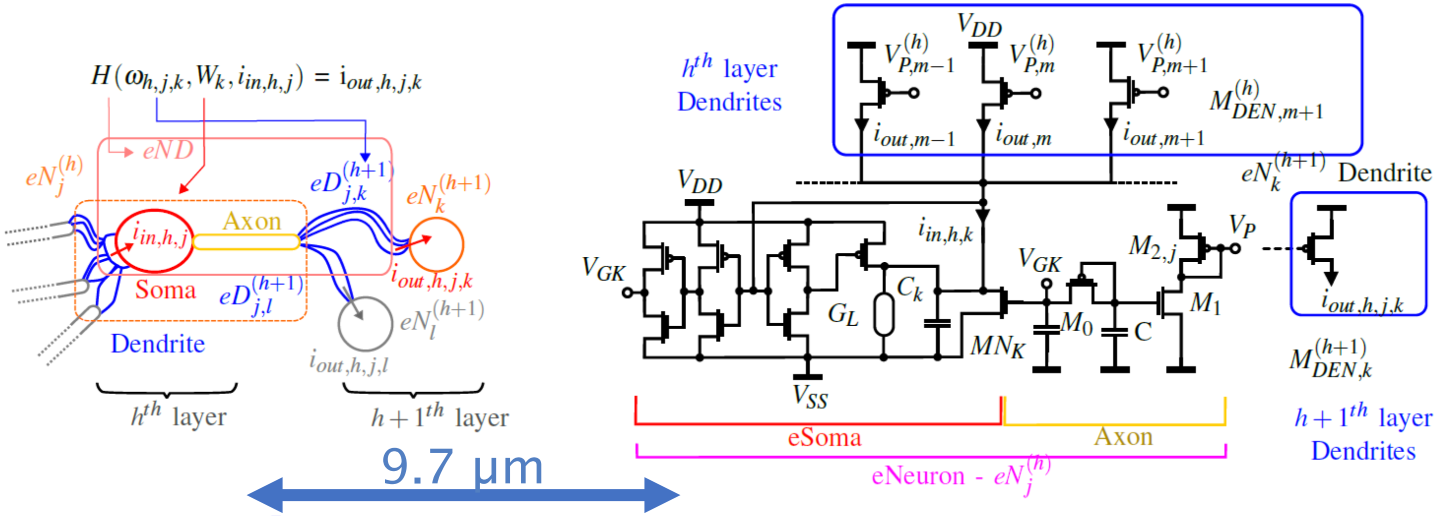


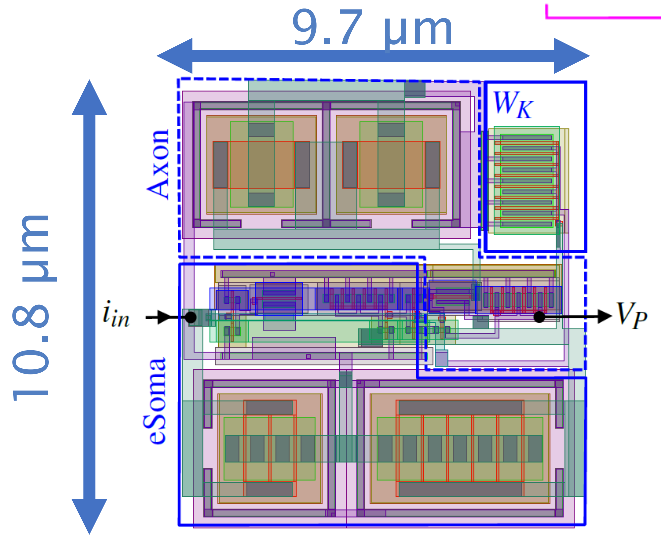
eSynapse



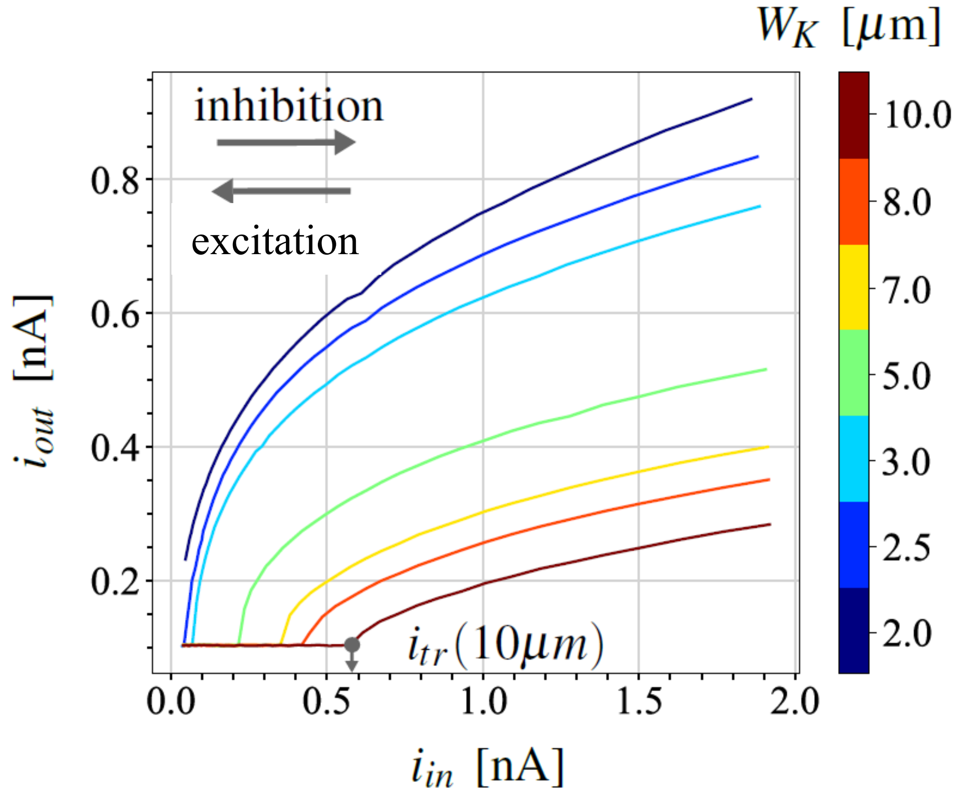


eNeuron with Faithful Biomimetic Behavior





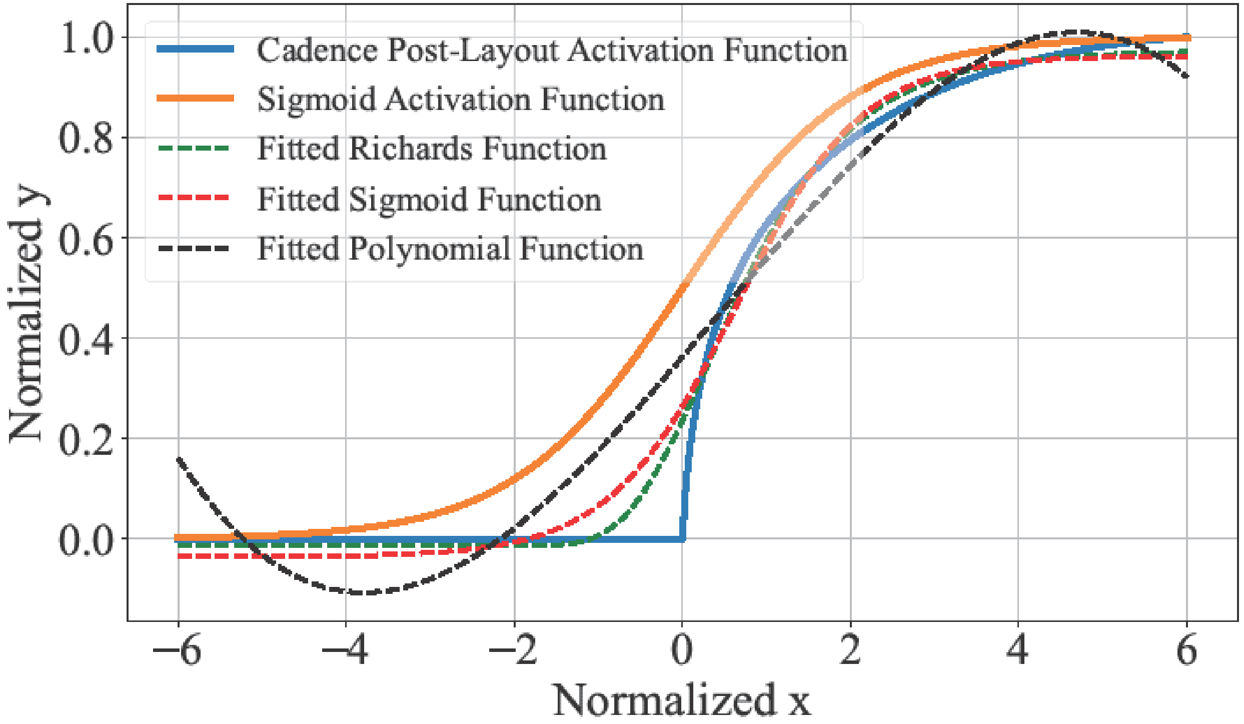
$$E_{eff} = 10\ fJ/spike$$
$$\sigma = 1.8\ \%$$
$$Mismatch \leq 4\ pA$$



https://doi.org/10.1109/SBCCI60457.2023.10261961

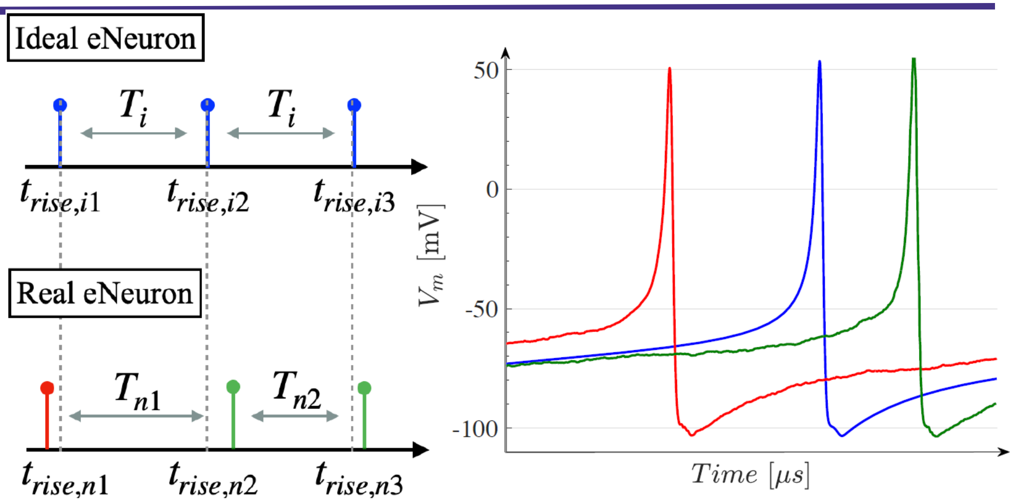
From eNeurons to Spiking Neural Networks

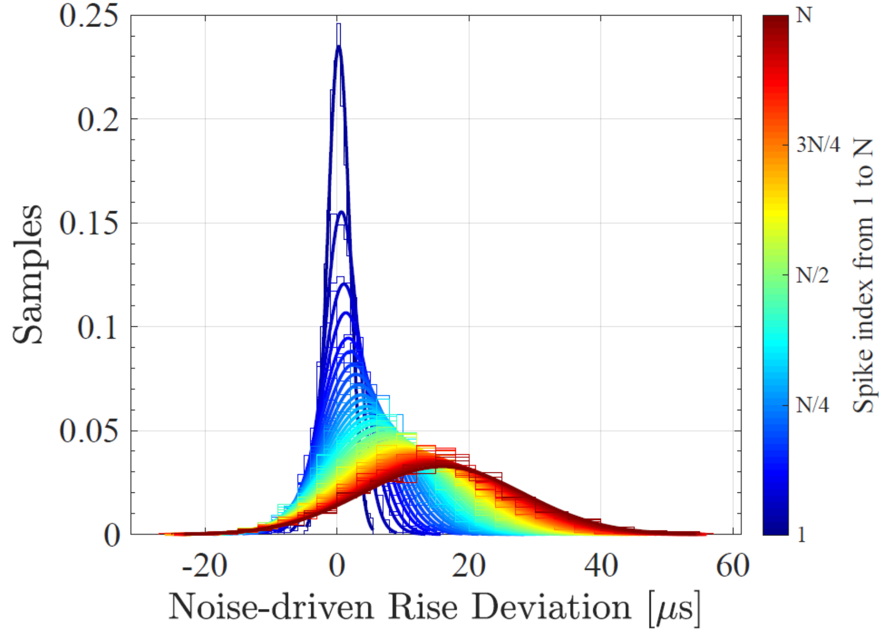
- Activation Function: Physical informed models
- Fitting is not always the best solution
- SNN needs training and inference
  - ❑ Physical AF?
  - ❑ Mathematical AF?

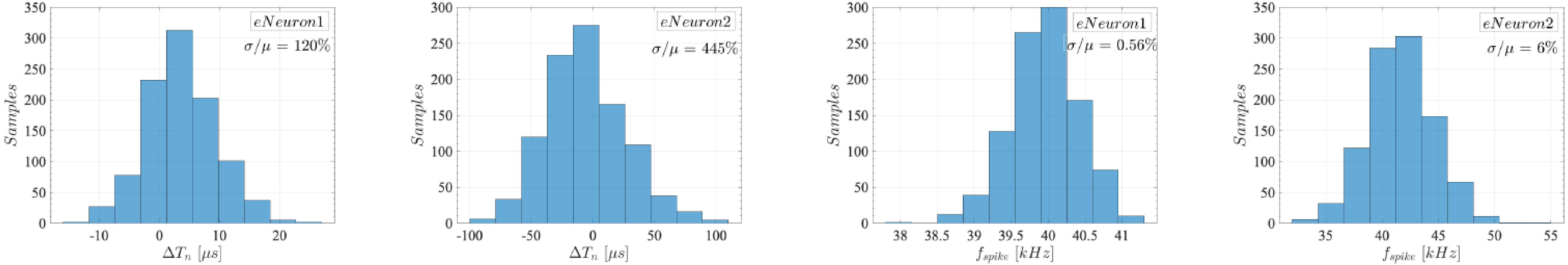


https://doi.org/10.29292/jics.v19i3.889

SNN Training and Inference under Jitter Noise







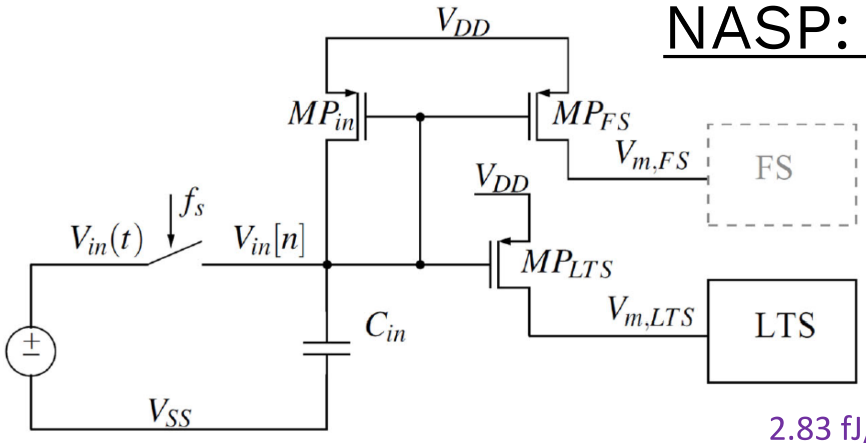
https://doi.org/10.1109/SBCCI60457.2023.10261661

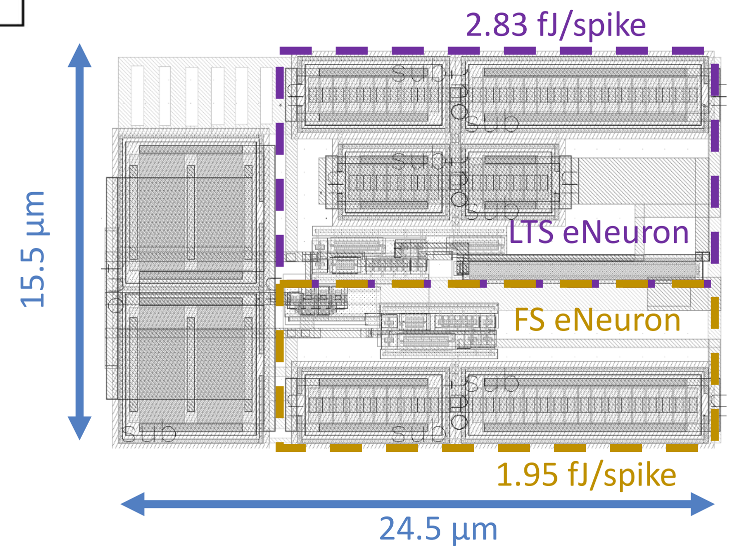
Agenda

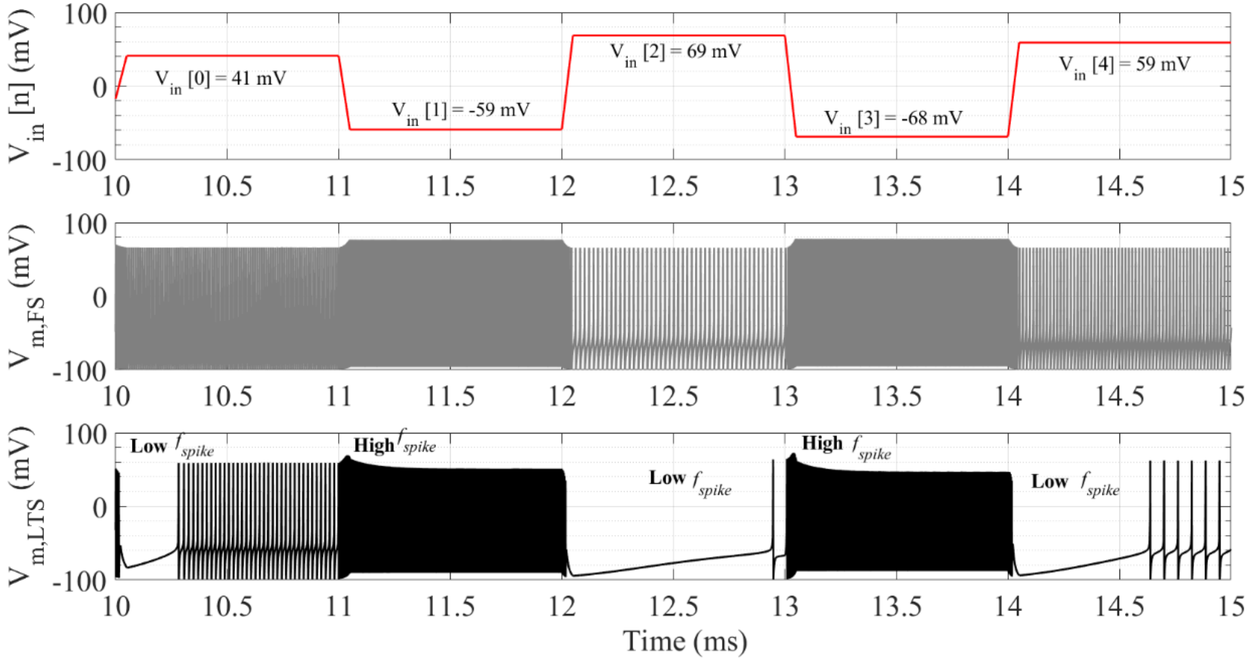
- Research Context and Challenges
- How to design Neuromorphic Circuits
- Fundamentals in Neuromorphic Circuits
- Physical Implementation
- Neuromorphic Circuit and Systems Illustrations

Neuromorphic Analog Spiking-Modulator (NASP)

NASP: 8 fJ/conv, 9 bits resolution 55 nm Integrated SNN







10.1007/s10470-020-01729-3

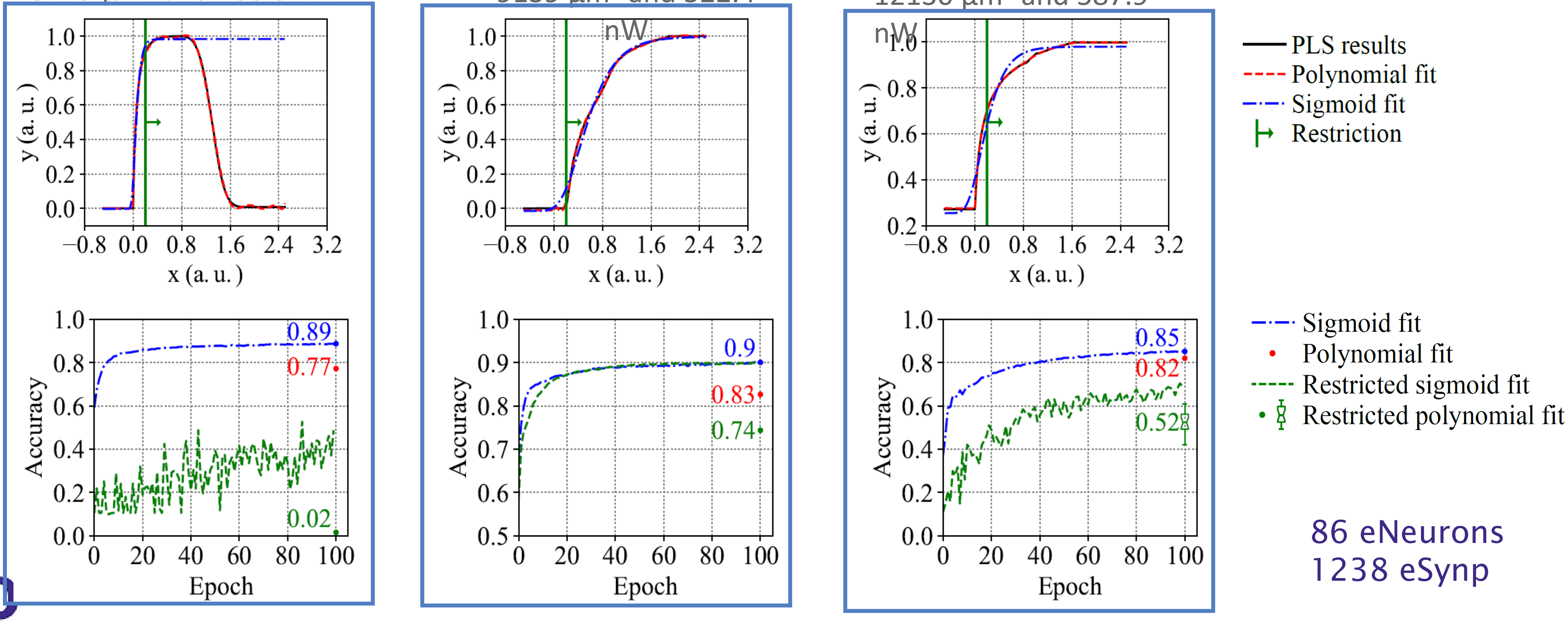
Analog Spiking Neural Network Synthesis for the MNIST

10.29292/jics.v18i1.663

LIF
5749 µm² and 255.9 nW

ML Simplified
9139 µm² and 522.4 nW

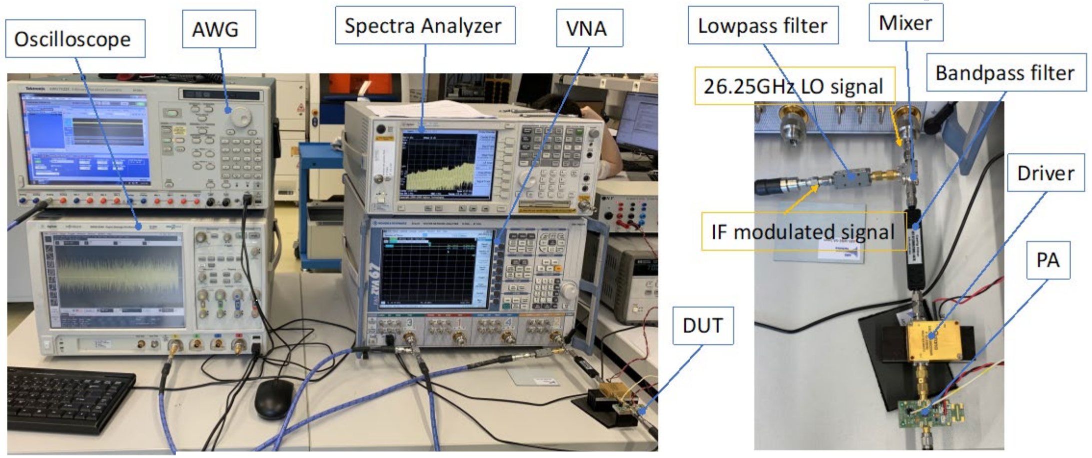
ML Biomimetic
12130 µm² and 587.9 nW

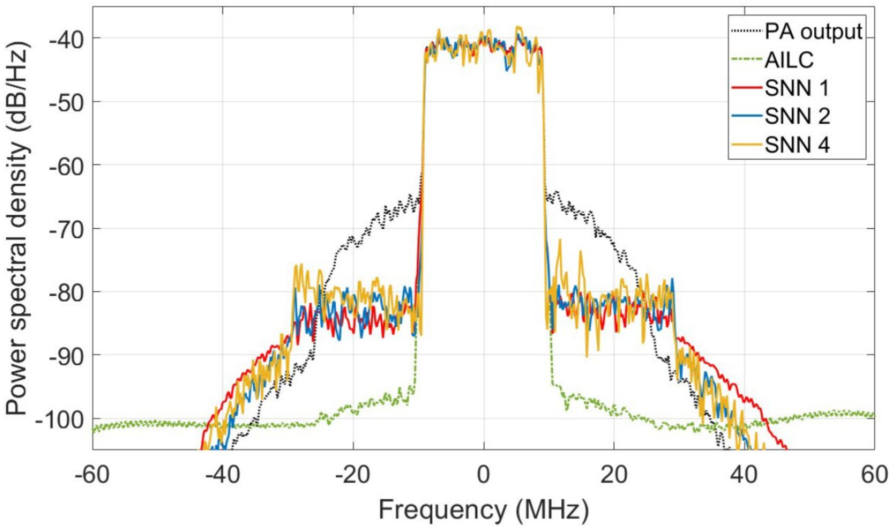


86 eNeurons
1238 eSynp

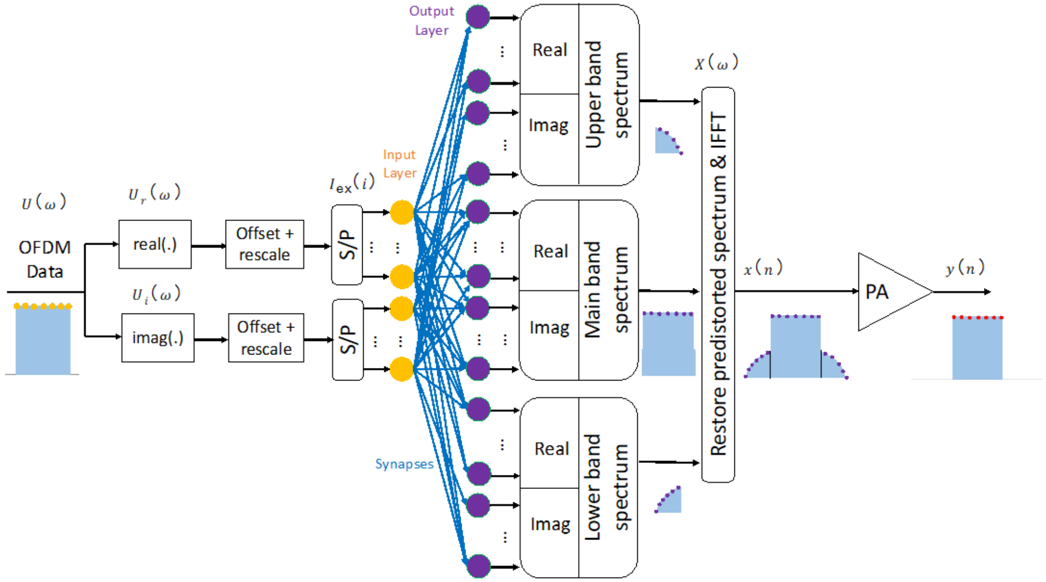
Physics Informed Spiking Neural Network DPD for PA linearization

Measurement Setup





SNN Architecture



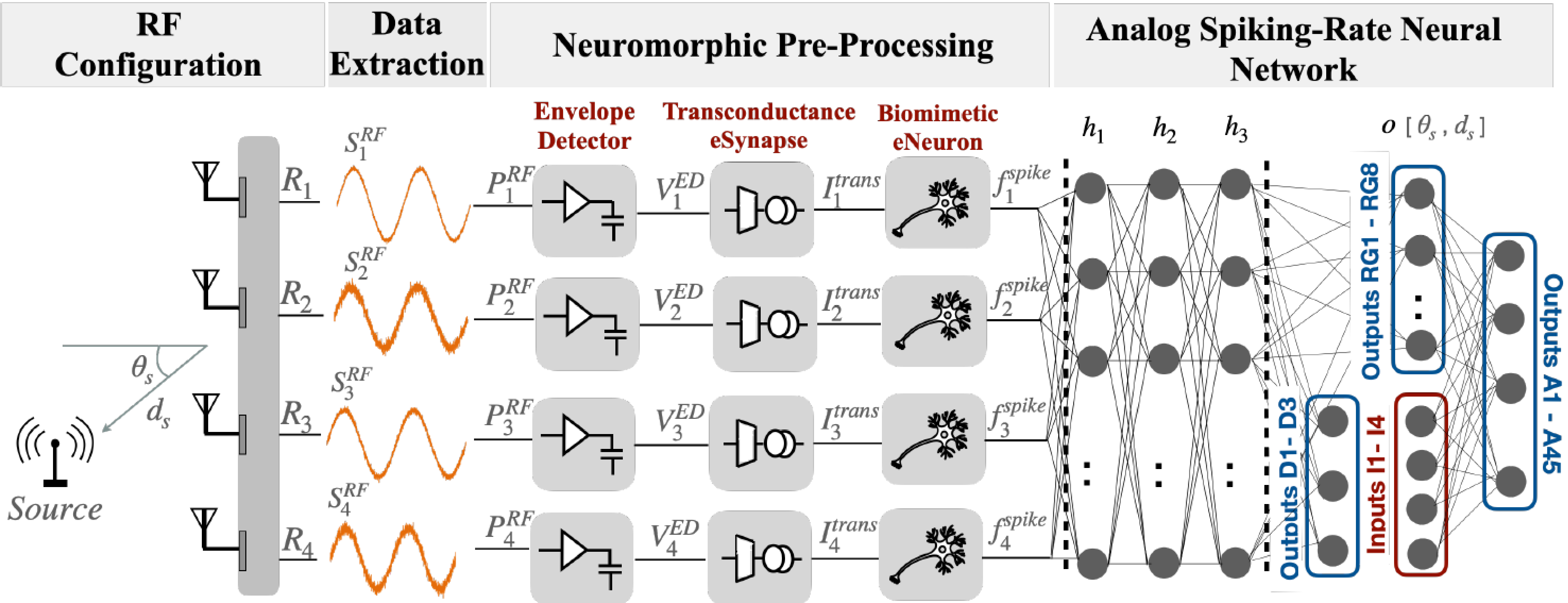
80 eNeurons, 54 eSynps

[10.1109/ACCESS.2023.3275434](https://doi.org/10.1109/ACCESS.2023.3275434)

[10.1109/NEWCAS58973.2024.10666321](https://doi.org/10.1109/NEWCAS58973.2024.10666321)

Spiking-Rate Neuromorphic System for Efficient RF Source Localization

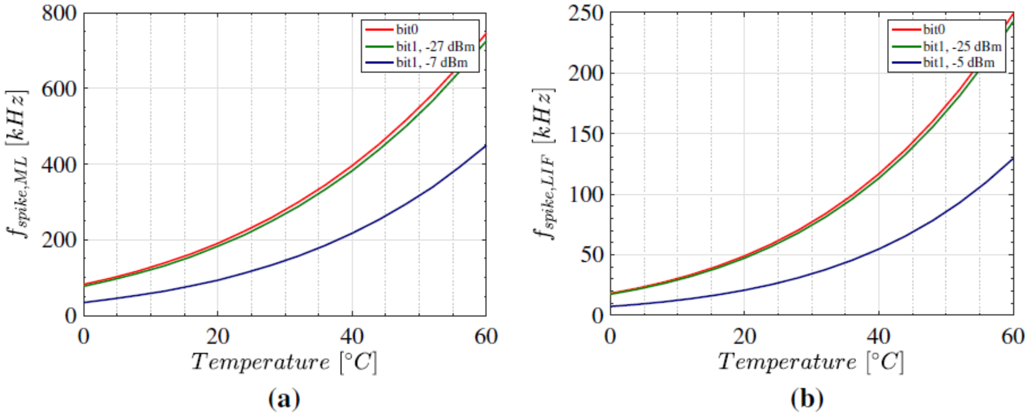
Zalfa JOUNI, PhD cand.

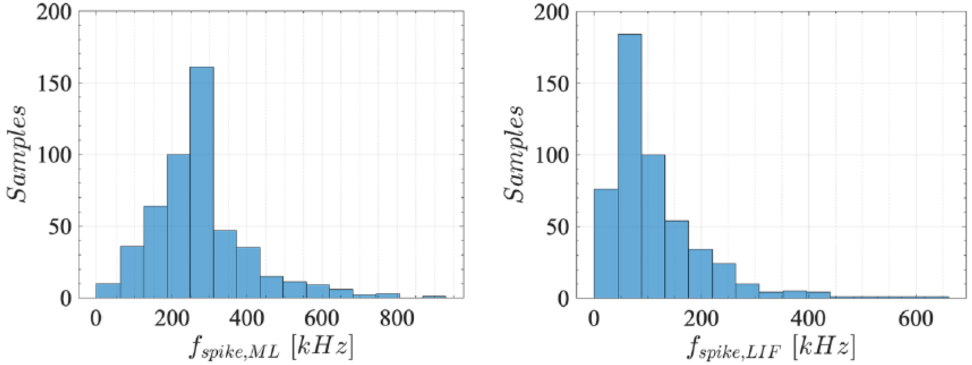


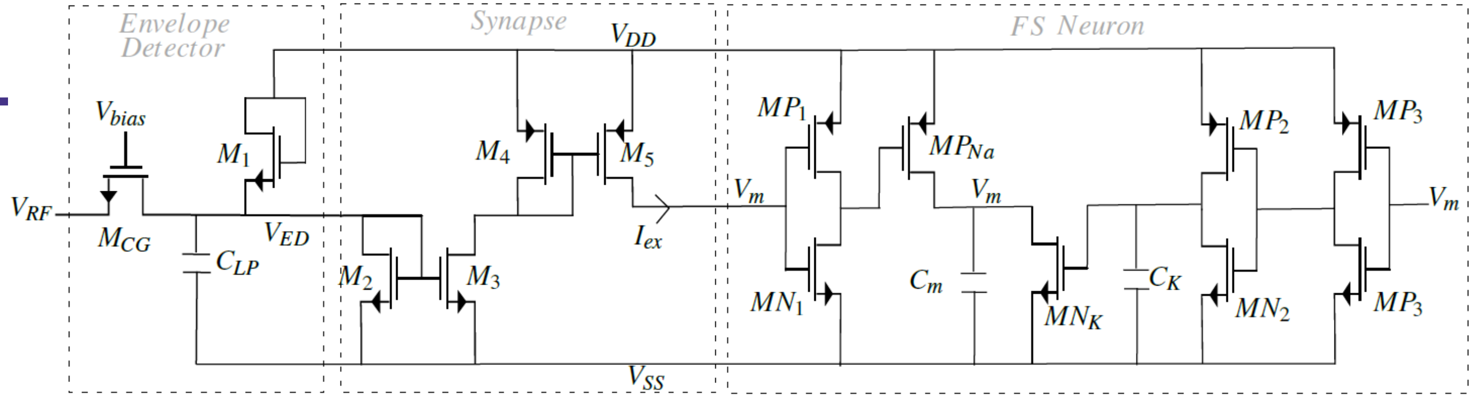
10.1109/ICECS61496.2024.10849126

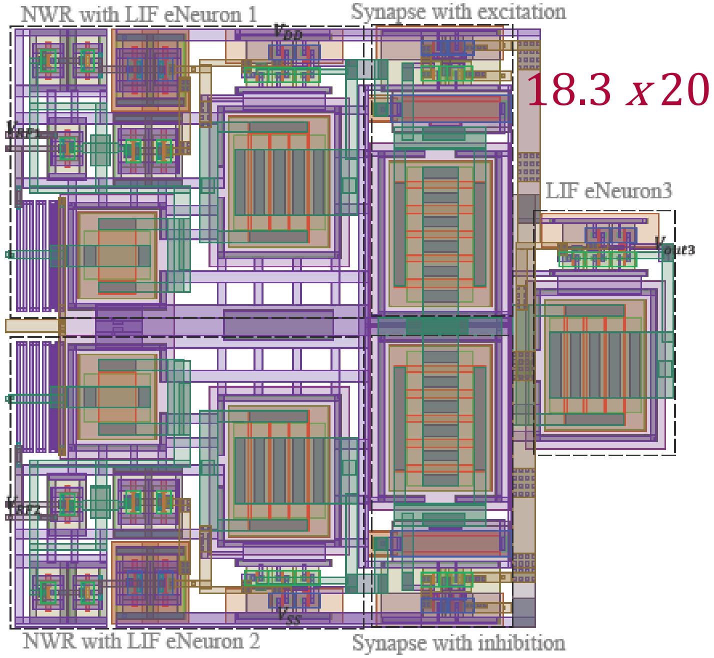
10.1109/SBCCI55532.2022.9893247
10.1007/s10470-023-02164-w

Neuromorphic Pre-Processing







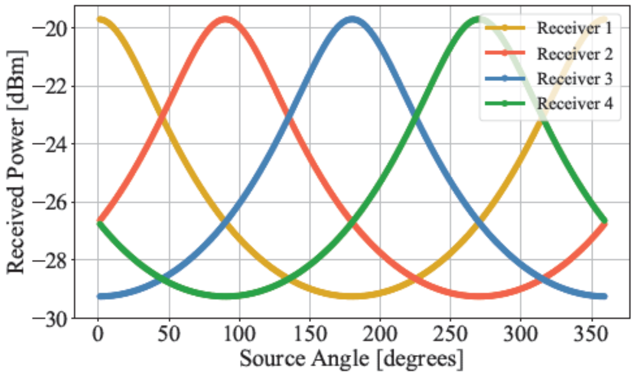


$18.3 \ x \, 20.24 \ \mu m^2$

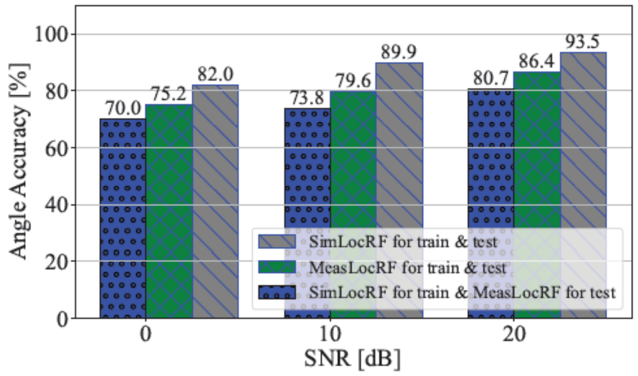
From data generation to NN training

- RSS detection
- 1-degree angular
- resolution

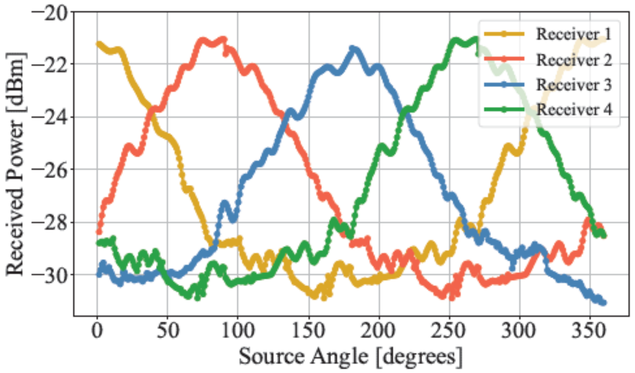
- Power 403 nW
- SNR 20 dB
- high accuracies: 93.5% (simulated) and 86.4% (experimental)



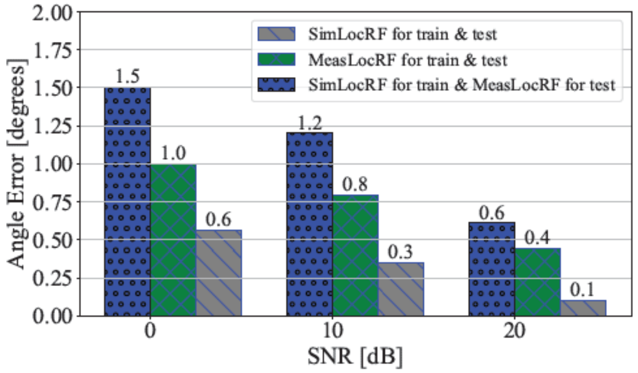
(a)



(a)



(b)



(b)

croma

Agenda

- Research Context and Challenges
- How to design Neuromorphic Circuits
- Fundamentals in Neuromorphic Circuits
- Physical Implementation
- Neuromorphic Circuit and Systems Illustrations

Conclusions

- Use less and design better electronics
  - ❑Cloud AI and pure software-based AI is not all Green
- AI-edge is the IoT 3.0
  - ❑Need Codesign: Computer + Electronic
  - ❑Energy Harvesting for battery less devices

- Mimicking Humans in Semiconductors is a nice idea ?
  - ❑Machine senses are not often Human senses
  - ❑Machine speed is higher than Human's one (~10 Hz)
  - ❑Machine power consumption is greater than Human's one (20 W)

Thank you for your attention

Questions? maris@ieee.org

croma Centre de Radiofréquences, Optique et Micro-nanoélectronique des Alpes

GRENOBLE INP UGA

UNIVERSITÉ SAVOIE MONT BLANC

UGA Université Grenoble Alpes

cnrs