Bio-inspired sensing and computation for a future energy efficient Edge AI

Adrian M. Ionescu, EPFL, Switzerland

Ecole Polytechnique Fédérale de Lausanne

EPFL = a nice place to live and work, near Lehman lake

~10,000 students, ~6,000 staff, ~350 professors and labs, annual budget ~1BCHF

QS World University Rankings by Subject 2023: Electrical and Electronic Engineering

Rank	University	Overall Score	Academic Reputation	Employer Reputation	
1	Massachusetts Institute of Technology (MIT) Cambridge,United States	97.5	100	100	95.3
2	Stanford University Stanford,United States	94.4	96.2	97.5	97.9
3	University of California, Berkeley (UCB) Berkeley,United States	91.4	93.3	92.8	95.4
4	ETH Zurich Zürich,Switzerland	90.7	91.2	94.6	93.9
5	University of Cambridge Cambridge,United Kingdom	90.2	89.2	95.6	93.3
6	EPFL Lausanne,Switzerland	89.9	92.5	90.3	92.9
7	Harvard University Cambridge,United States	89.5	85.7	98.6	98

Outline

- Moore’s law and the quest for energy efficiency
- Why bioinspired spiking neuromorphic hardware @ the Edge?
- A ferroelectric junctionless synapse
- Ferroelectric 2D FETs, NCFETs, Tunnel FETs and NC Tunnel FETs
- Memristive phase change / Mretal-Insulator-Transition materials and devices
- Conclusion

Moore’s law: *in memoriam*

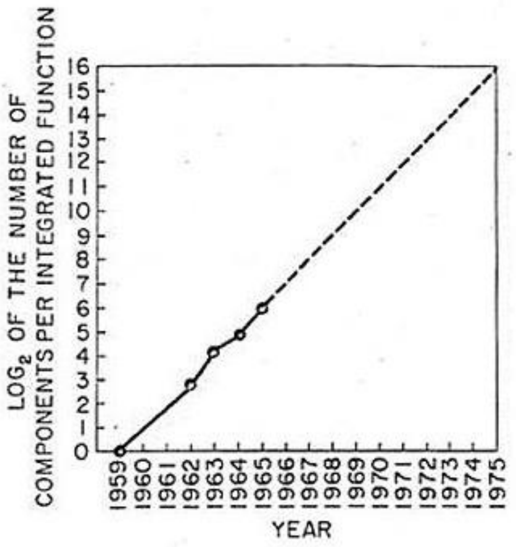
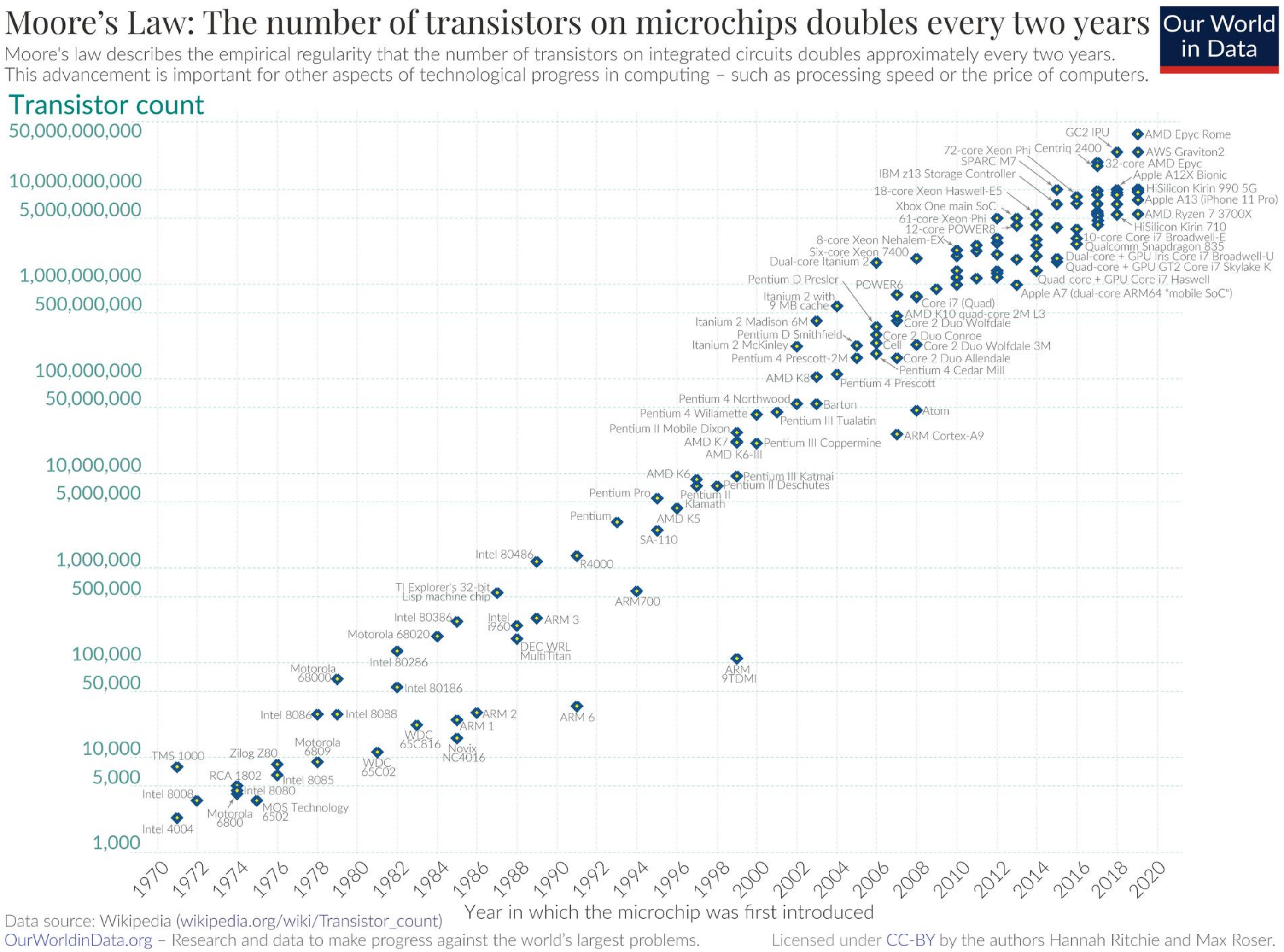
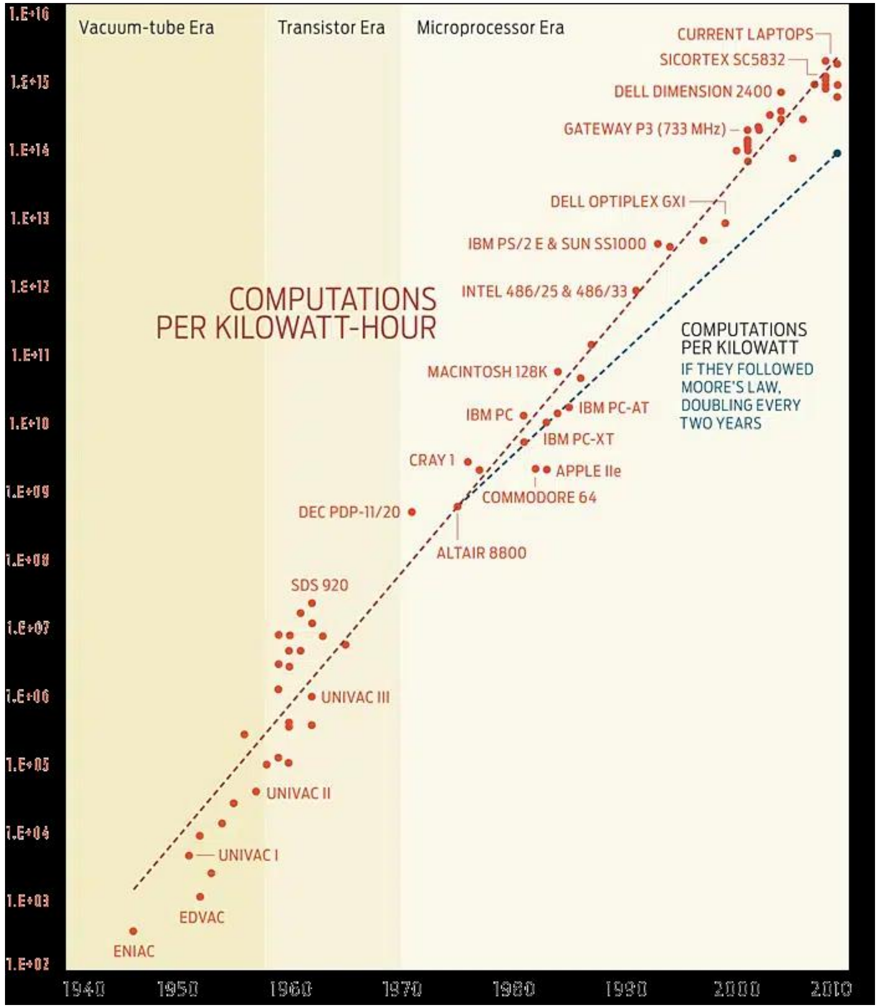


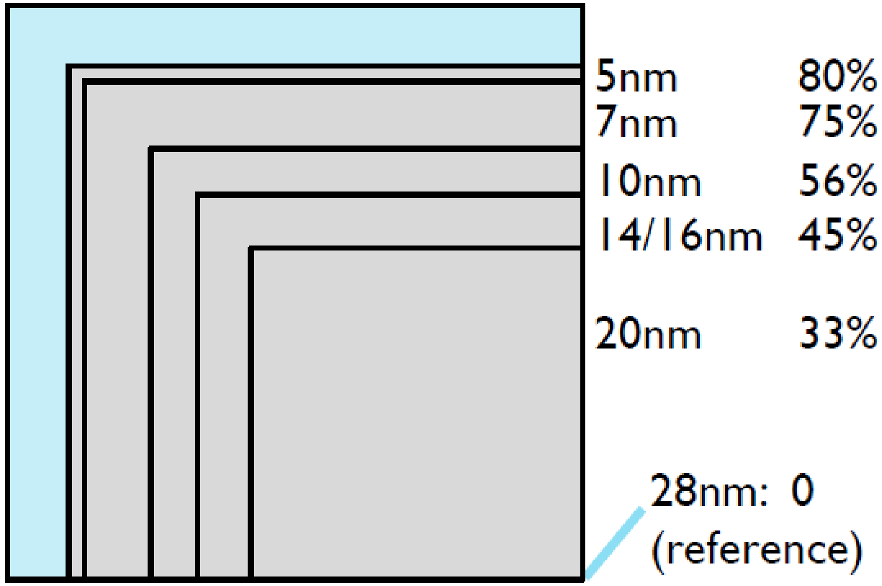
Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

From Moore’s law to Koomey’s law





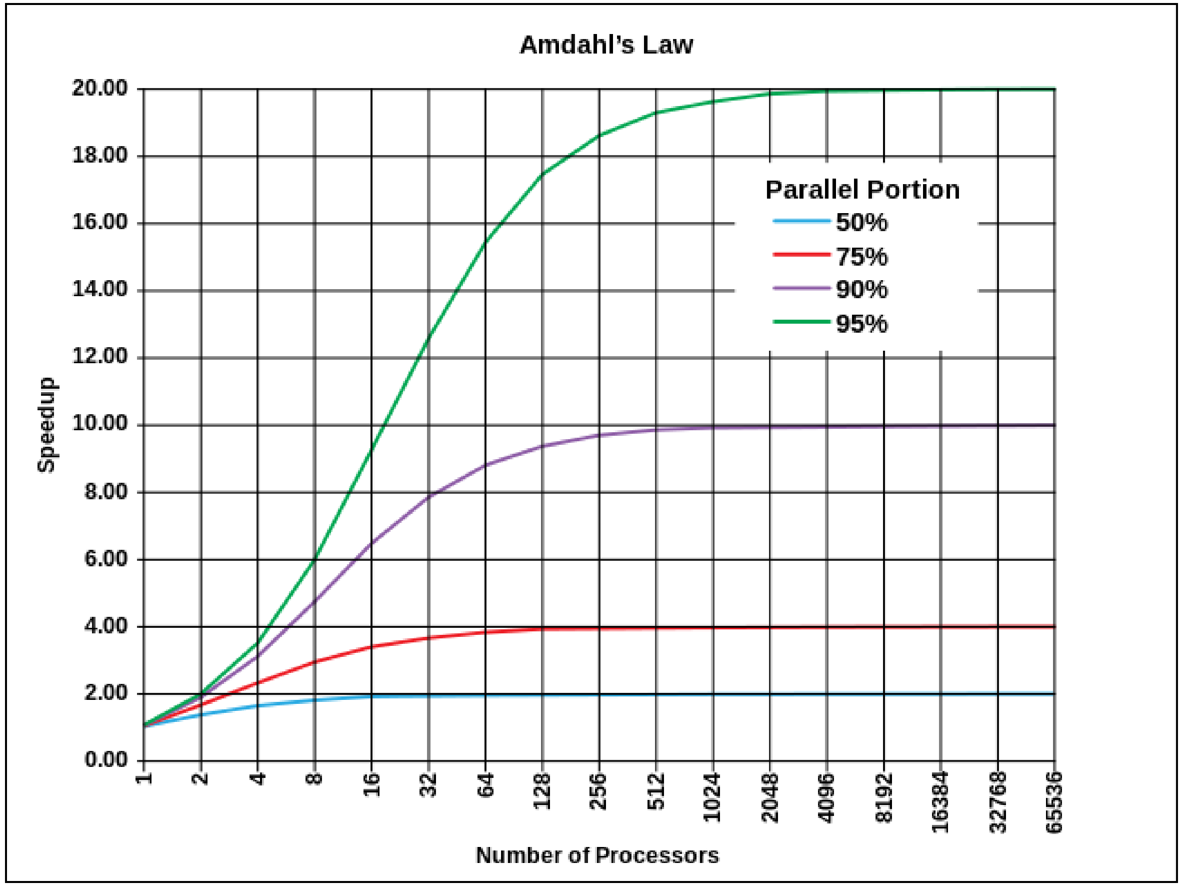
Power density challenge and dark silicon



We get more transistors, we just can't afford to turn them all!

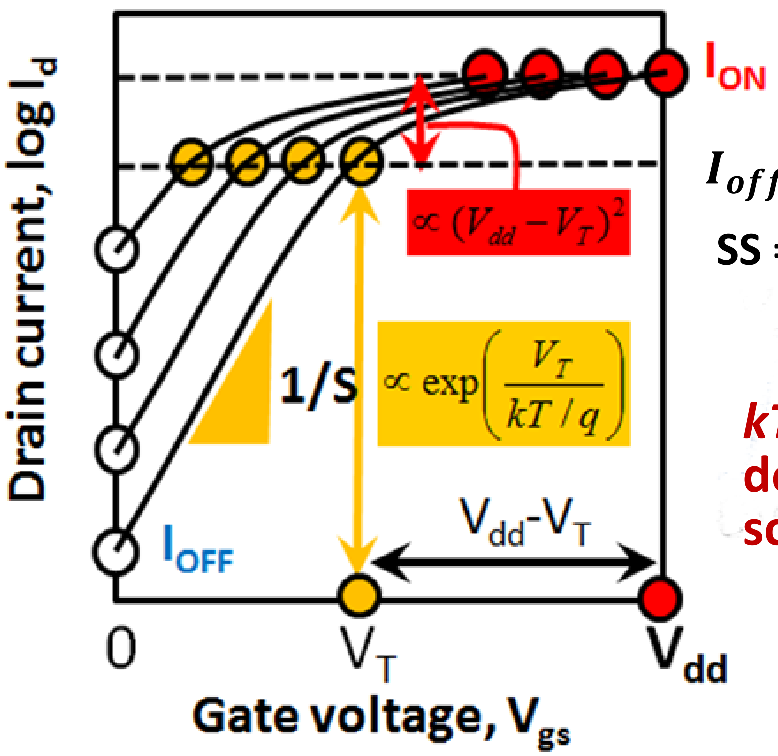
Greg Yeric, ARM @ IEDM 2015

One or two walls?



Leakage power and steep slope switches

- Leakage power: incompressible subthreshold swing of MOSFET: 60mV/dec @ RT
- Vdd scaling saturated @ ~0.7-0.8V → **scaling V_{dd} and V_T through steep slope switches**

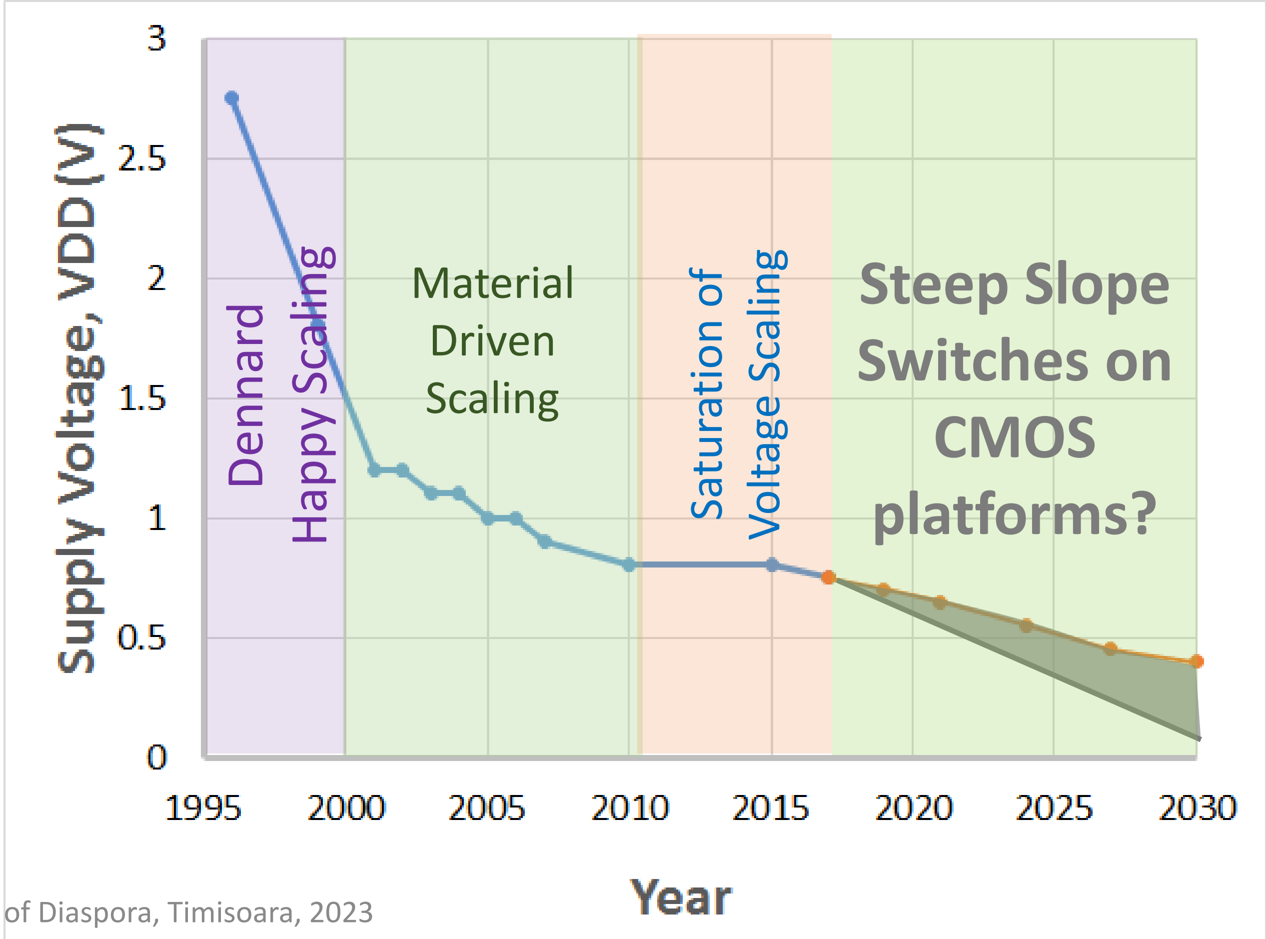


Ionescu & Riel, Nature, 2011.

$$I_{off} = I_0 10^{-V_{th}/SS}$$

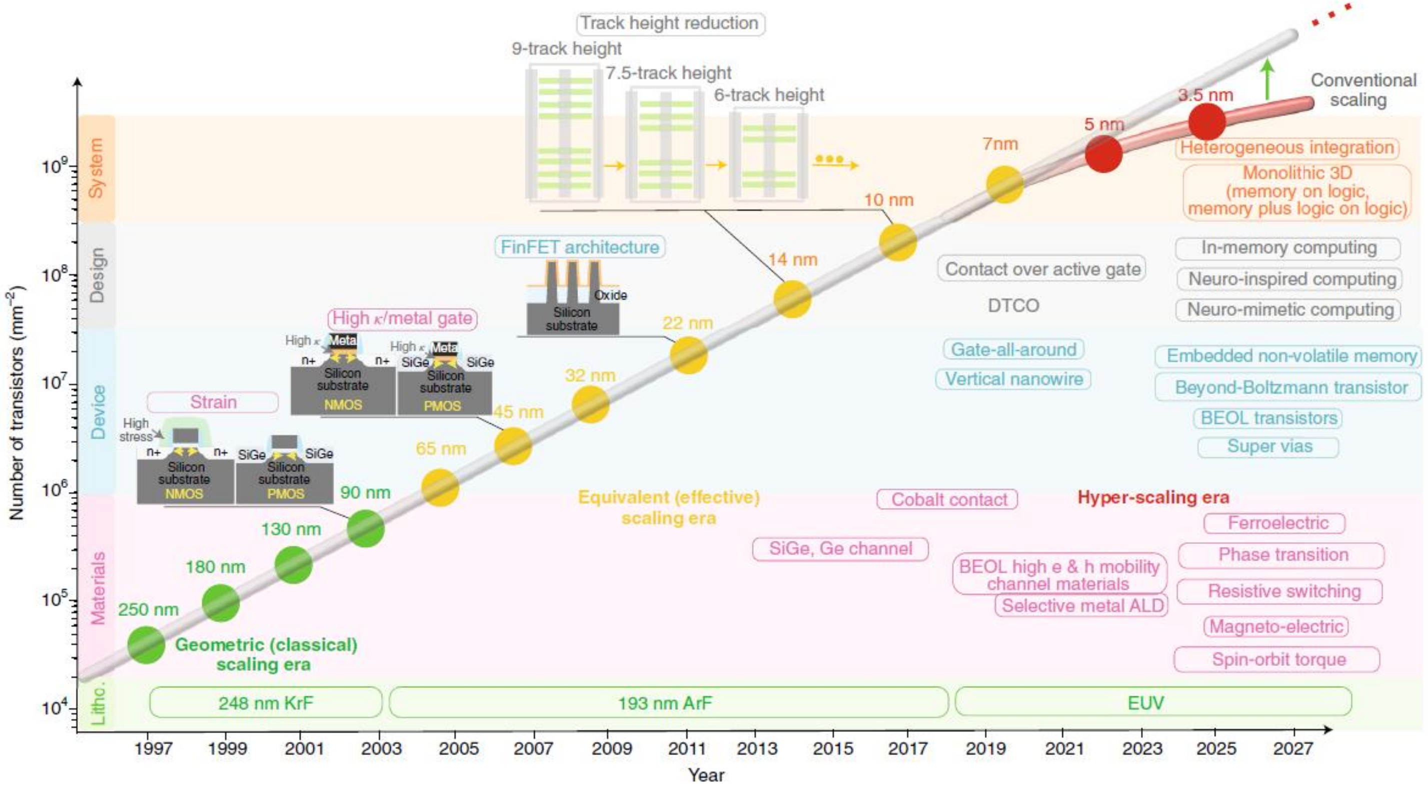
$$SS = \frac{dV_g}{dlog(I_d)} \text{ @ RT}$$

kT/q
does not scale!



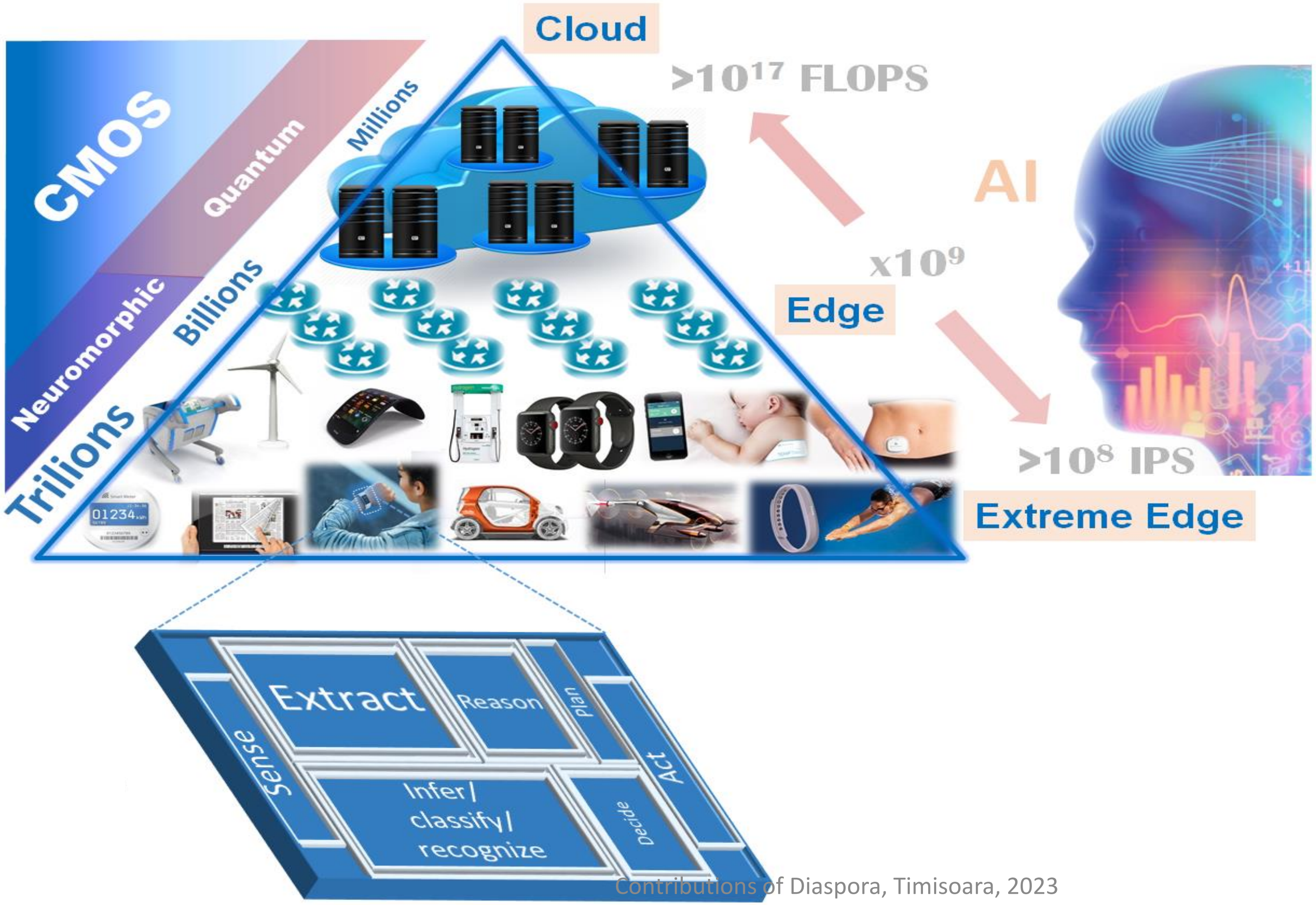
What is next: hyperscaling concepts…

- **Hyper-scaling** — approach needed to meet the demands of data abundant workloads.
- Driven by:
  - ❑ monolithic 3D integration and heterogeneous integration
  - ❑ embedded non-volatile memories
  - ❑ Beyond-Boltzmann transistors
  - ❑ …

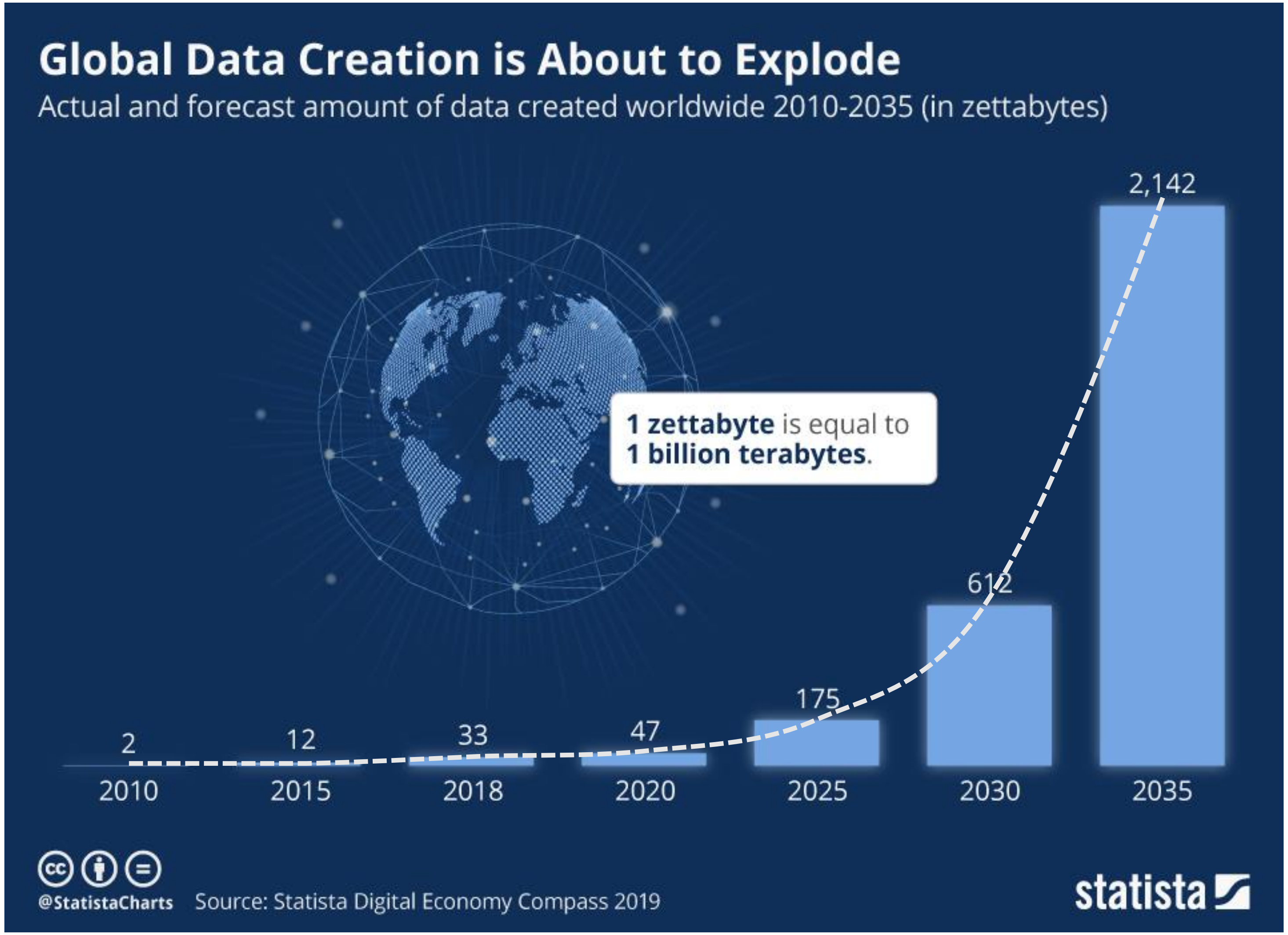


Salahuddin, S., Ni, K. & Datta, S. The era of hyper-scaling in electronics. Nat Electron 1, 442–450 (2018). https://doi.org/10.1038/s41928-018-0117-x

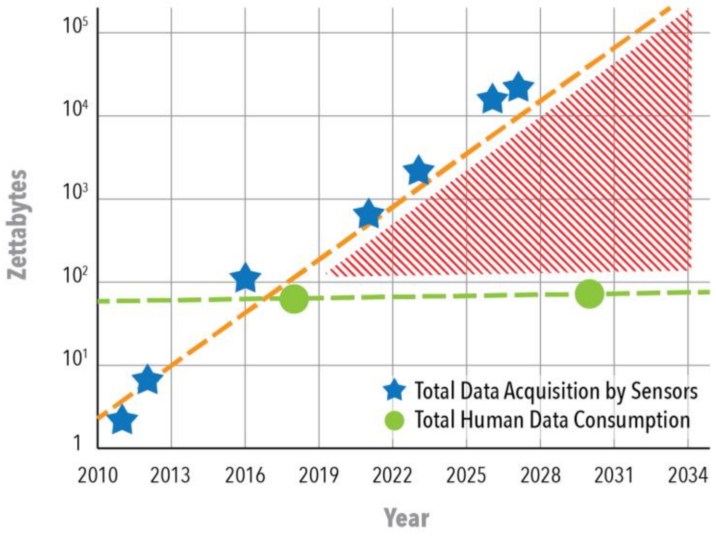
Edge to Cloud computation



Energy crisis in the Zettbyte era…



Global Data: IoT Sensors



+ 1 trillion IoT devices by 2035 with annual growth >20% (© ARM)

Two major issues:

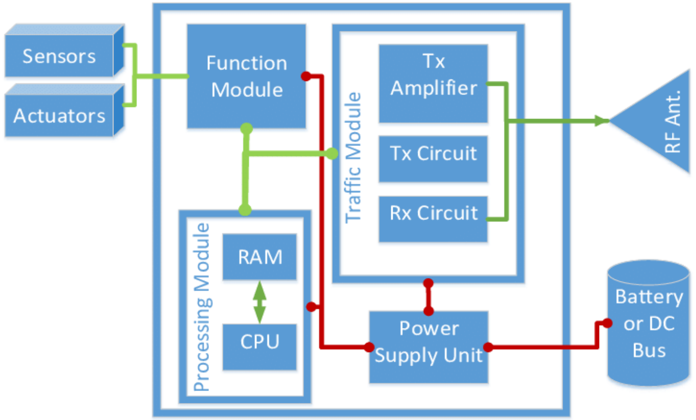
- **Energy efficiency**
- **Data proliferation**

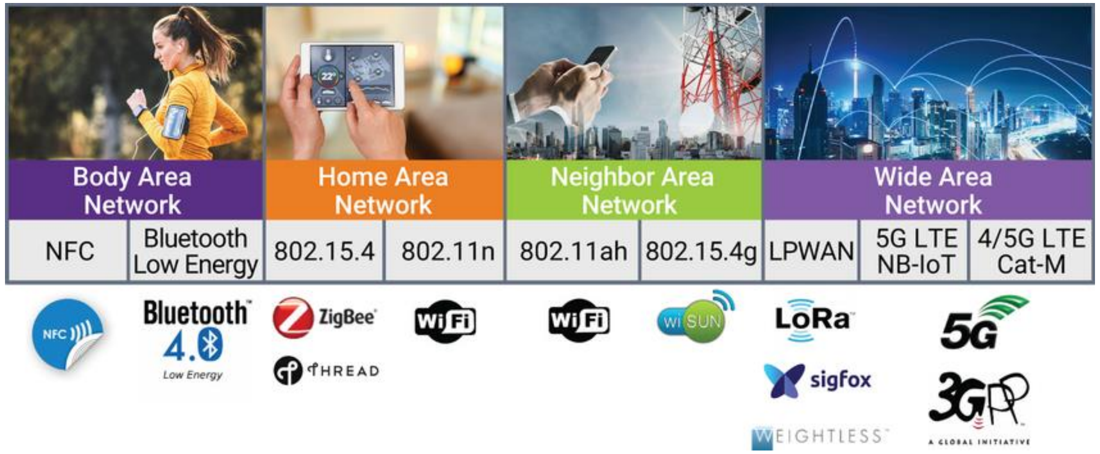
IoT nodes @ the Edge: sustainable for future deployment in trillions?

Industrial IoT node size and power consumption: mm^3 to cm^3 with 100's uW to 10's mW.

Silicon = only solution for all IoT Node Devices?

- Sensing
- Processing
- Communications



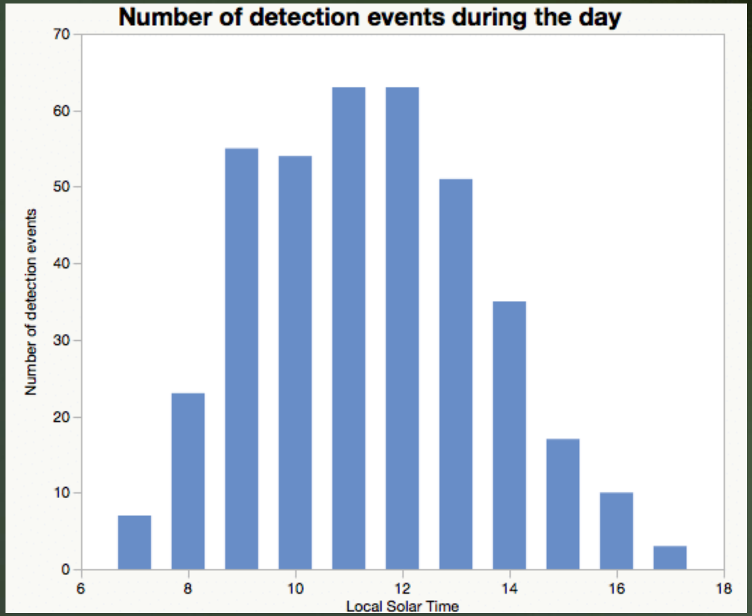


Energy problems @ node level:

- **No digital data reduction**
- **Expensive ADC and digital processing**
- **Expensive data communication**

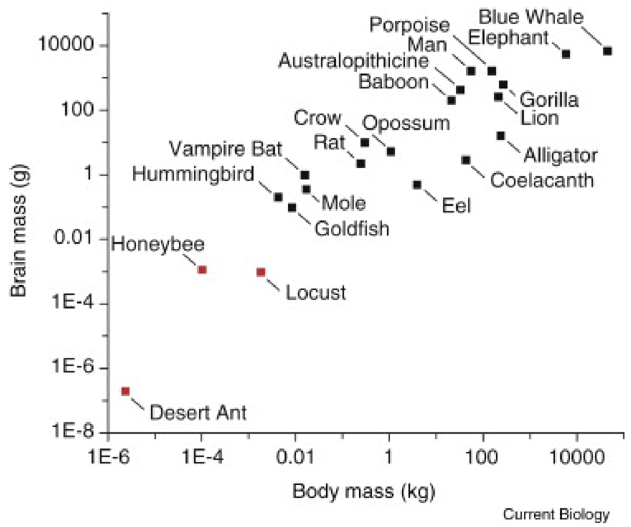
Massive reduction in IoT data proliferation

- bio-inspiration needed
- no digital, no ADC → time-domain spikes
- no sensed bits transmitted → event/tasks



The most energy efficient biological… node

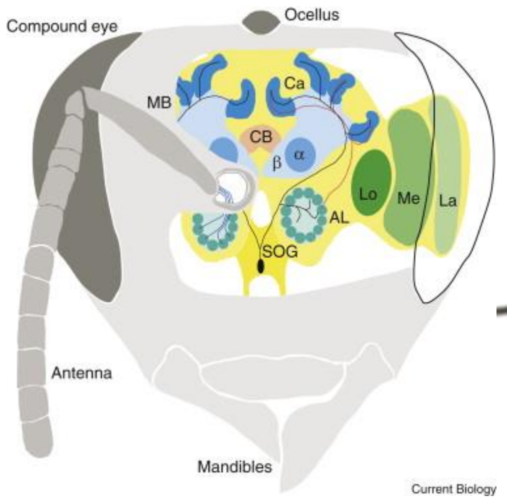
Are Bigger Brains Better?



Lars Chittka and Jeremy Niven, Current biology, 2009.

The honeybee

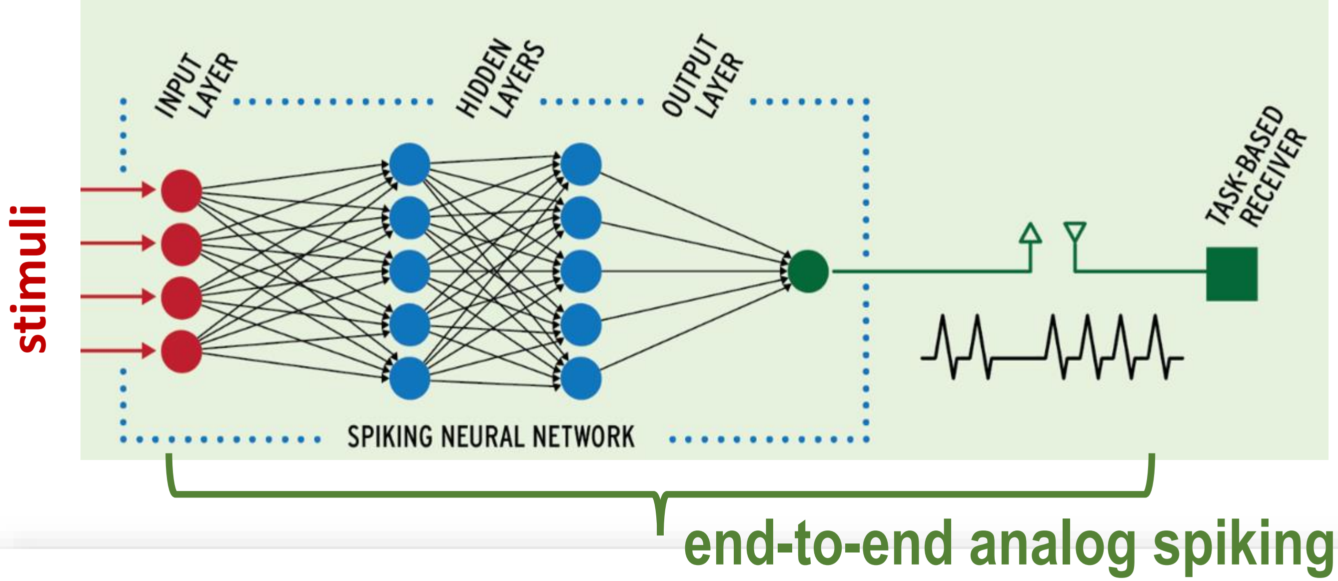
The brain of a honey bee has 960 000 neurons and is 1 mm^3 in size.

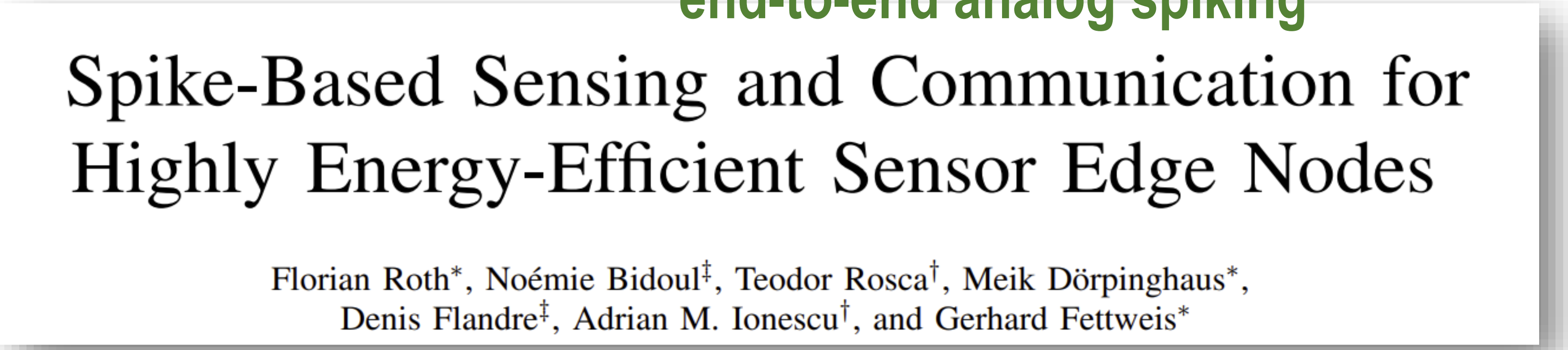


Neural network analyses show that **cognitive features found in insects, such as numerosity, attention and categorisation-like processes, may require only very limited neuron numbers.**

How to achieve BOTH energy efficiency and massive reduction data proliferation in IoT

Idea: a new type of revolutionary bio-inspired **IoT node = SWIMS ©** (Fettweis, Flandre & Ionescu, 2019).




Spike-Based Sensing and Communication for Highly Energy-Efficient Sensor Edge Nodes

Florian Roth*, Noémie Bidoul‡, Teodor Rosca†, Meik Dörpinghaus*, Denis Flandre‡, Adrian M. Ionescu†, and Gerhard Fettweis*

2022 IEEE International Symposium on Joint Communications & Sensing, 2022.

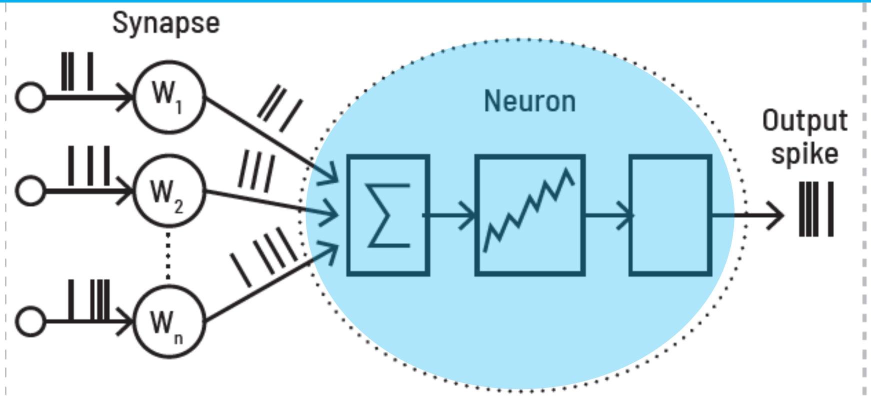
Bio-inspired all-spike analog signal processing

HUMAN BRAIN NEURONS:

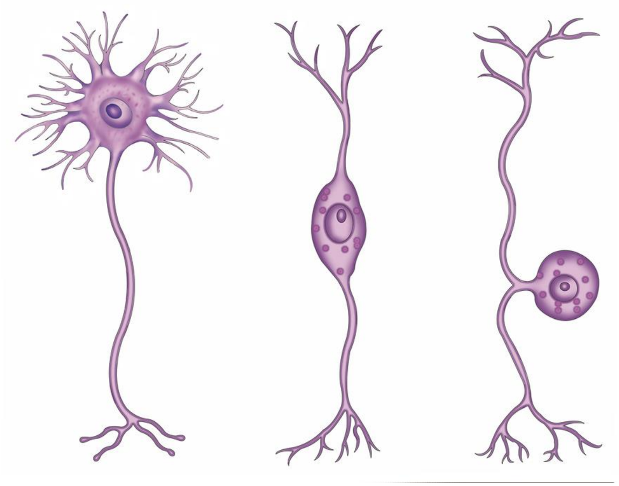
- 40 regions in the brain, each with different shaped neurons, perhaps a billion of each type.
- Different electrical properties
- Different neurons respond to transmission in different ways
- *Signaling of healthy and sick neurons can be differentiable.*

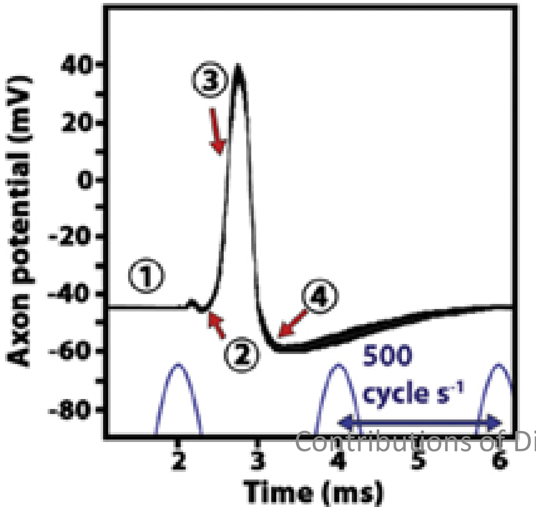
❑ **All-spike information processing @ ~110mV**
❑ **1.8 x 10^{14} spikes/Joule @ 36-37°C**
❑ **Neuron footprint: few to tens of um^2**

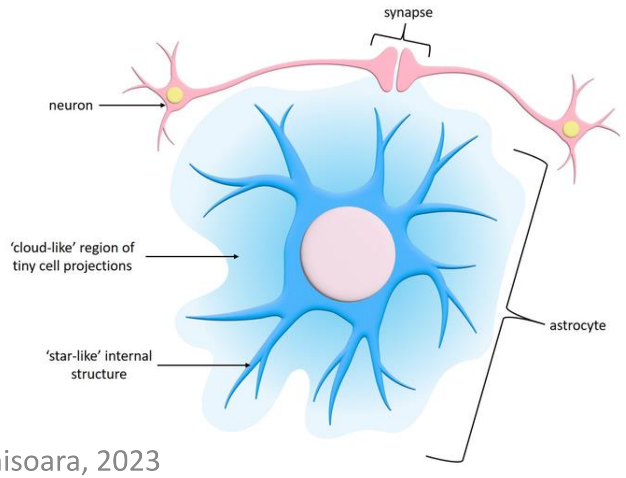
All-spike analog processing



Need: neuron, synapse, astrocyte, …







Astrocites (=a type of glial cell found of central nervous system) can modulate synaptic activity by releasing signaling molecules that can enhance or inhibit synaptic transmission.

Artificial neurons: functionality and state of the art

The Leaky Integrate-and-Fire (LIF) Neuron Mode

Memristive LIF (MLIF) Spiking Neuron Model

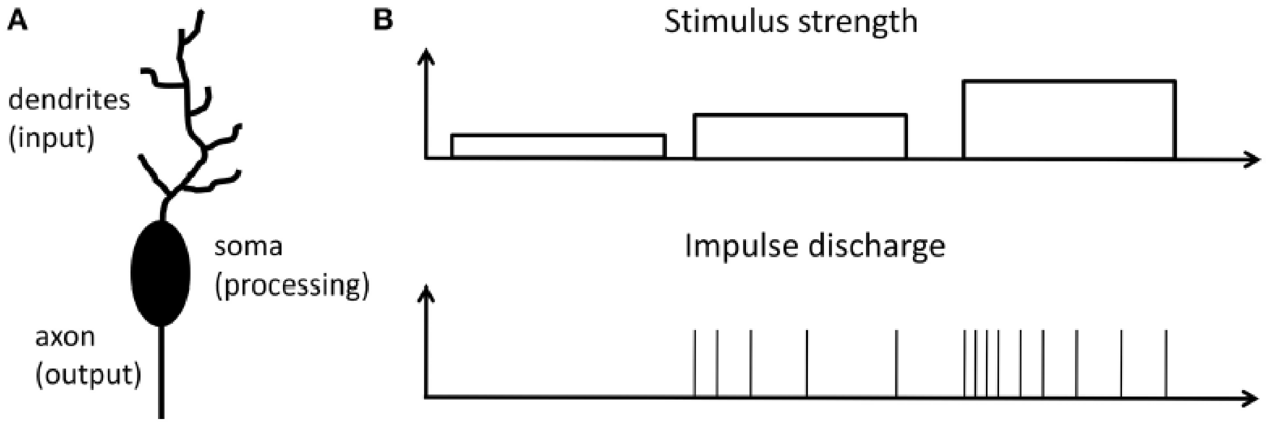
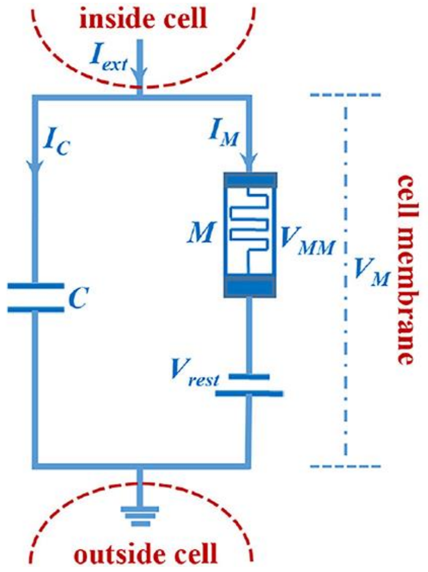


FIGURE 1 | Response to a stimulation principle: (A) Schematic of a single neuron, which can be divided into three functional parts: Dendrites, collect signals from other neurons; cell body (soma), the central processing unit of a neuron; axon, neuronal output stage. (B) Relationship between firing rate of a neuron and the strength of input stimulation reflecting the response to a stimulation principle as proposed by E. D. Adrian in 1926 (Adrian, 1926, 1928; Maass and Bishop, 2001).

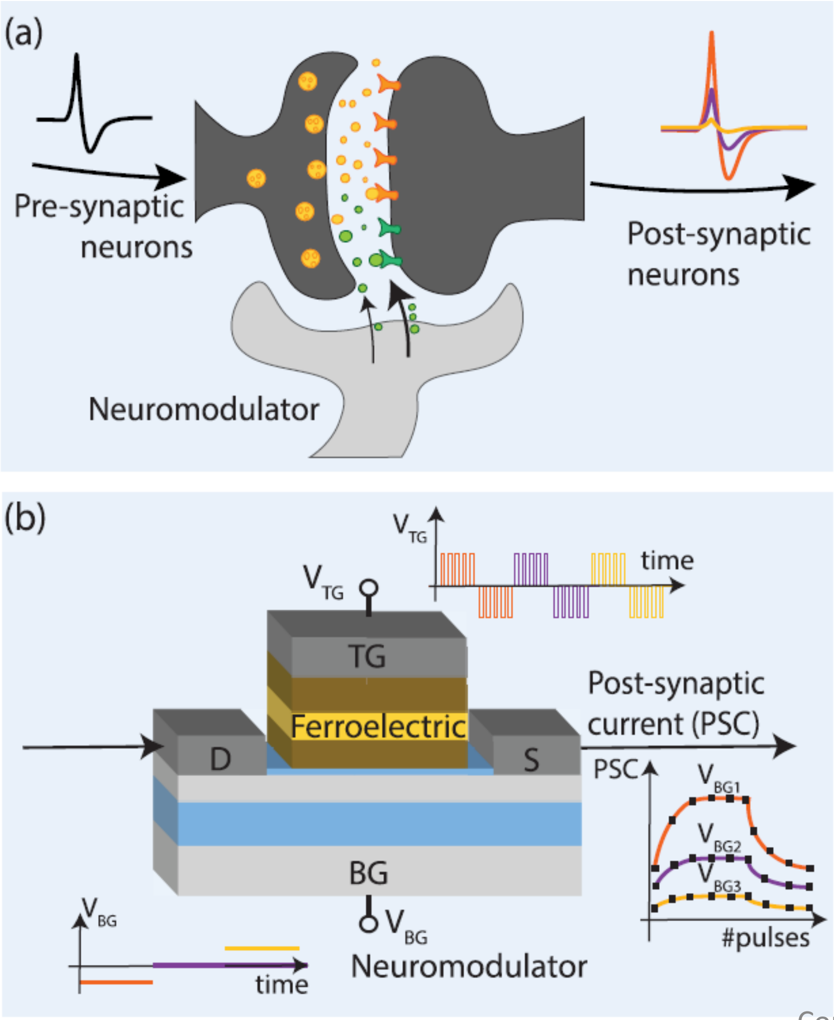


Artificial neurons: functionality and state of the art

	Indiveri et al., 2006	Joubert et al., 2012		Tuma et al., 2016	Sengupta et al., 2016	Jerry et al., 2017	S. Dutta et al., 2020,
Neuron type	LIF	Analog LIF	Digital LIF	LIF	LIF	Piecewise linear FHN	LIF
Material	CMOS	CMOS	CMOS	Phase change (PCM)	Magnetic tunnel junction (MTJ)	Vanadium dioxide (VO_2)	Ferroelectric HZO
Technology	800 nm	65 nm	65 nm	14 nm	–	–	45 nm
Integration mechanism	Capacitor charging	Capacitor charging	–	Joule heating	Magnetization dynamics	Capacitor charging	Polarization accumulative
Circuit elements	22 Transistor + one capacitor	33 Transistor + one capacitor	Pulse generator, counter, and comparator	One PCM + digital circuit	Two MTJs + four transistors	One VO_2 + one transistor + one capacitor	One FeFET + six transistors
Stochasticity	Yes	No	No	Yes	Yes	Yes	Yes
Power or energy/spike	900 pJ	2 pJ	41.3 pJ	120 μW	–	11.9 μW	1–10 pJ
Firing rate	200 Hz	2 MHz	2 MHz	35–40 kHz	–	30 kHz	50 kHz
Area	2573 μm^2	120 μm^2	538 μm^2	0.5–1 μm^2	–	–	2.05 μm^2

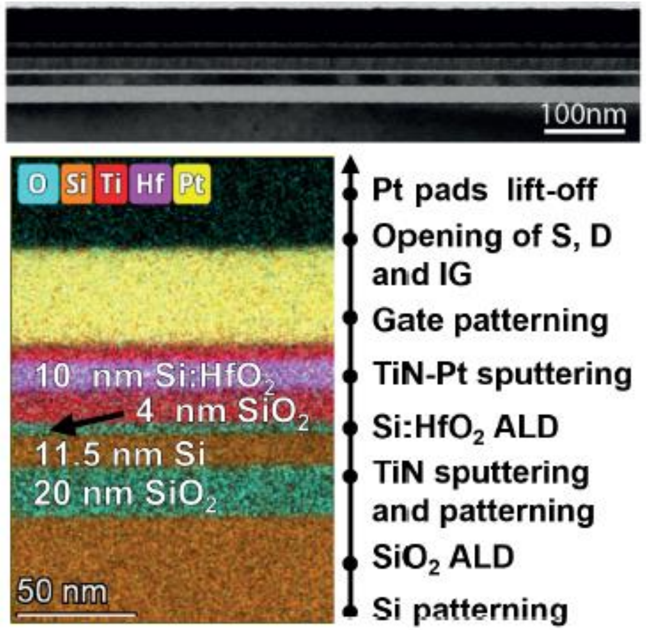
S. Dutta et al.,, "Supervised Learning in All FeFET Based Spiking Neural Network: Opportunities and Challenges," Front. Neurosci., 2020.

Ferroelectric Junctionless Double-Gate Silicon-On-Insulator FET as a Tripartite Synapse

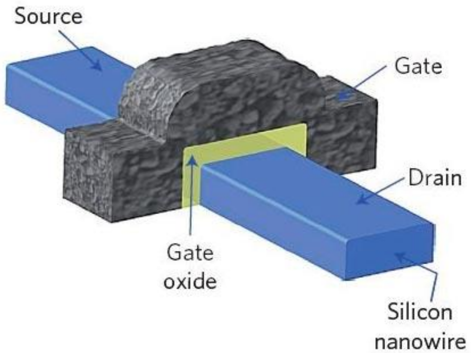


Junctionless DG SOI FET

- 11nm-thin film Fe-JLFET
- Top-gate ferroelectric (10nm Si-doped HfO2)
- Bottom-gate (20nm SiO2)

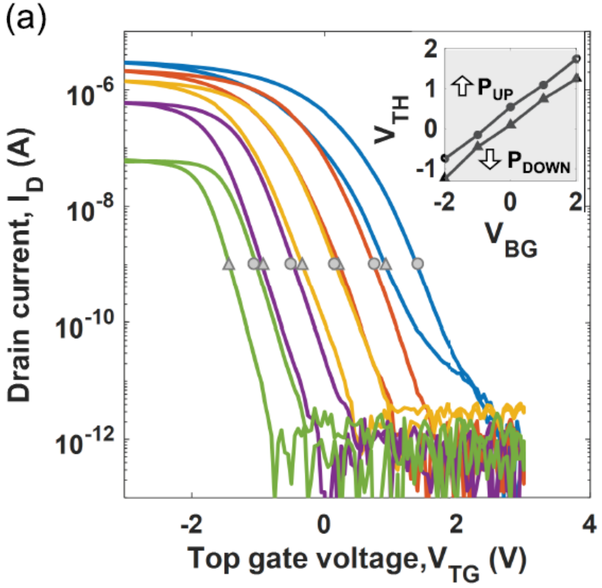


C. Gastaldi *et al.*, *IEEE Electron Device Letters*, April 2023

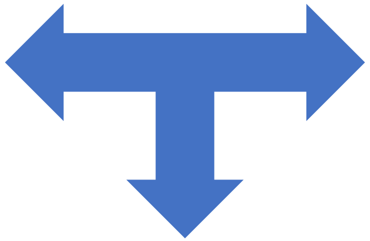


Colinge, JP., Lee, CW., Afzalian, *Nature Nanotech*, 2010.

Tripartite synapse electrical operation

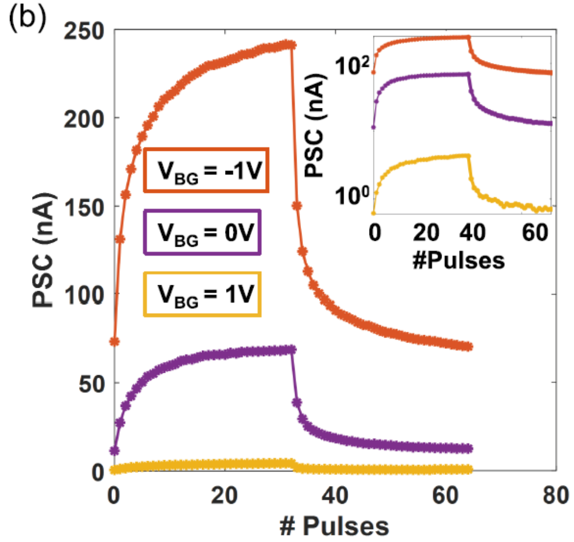


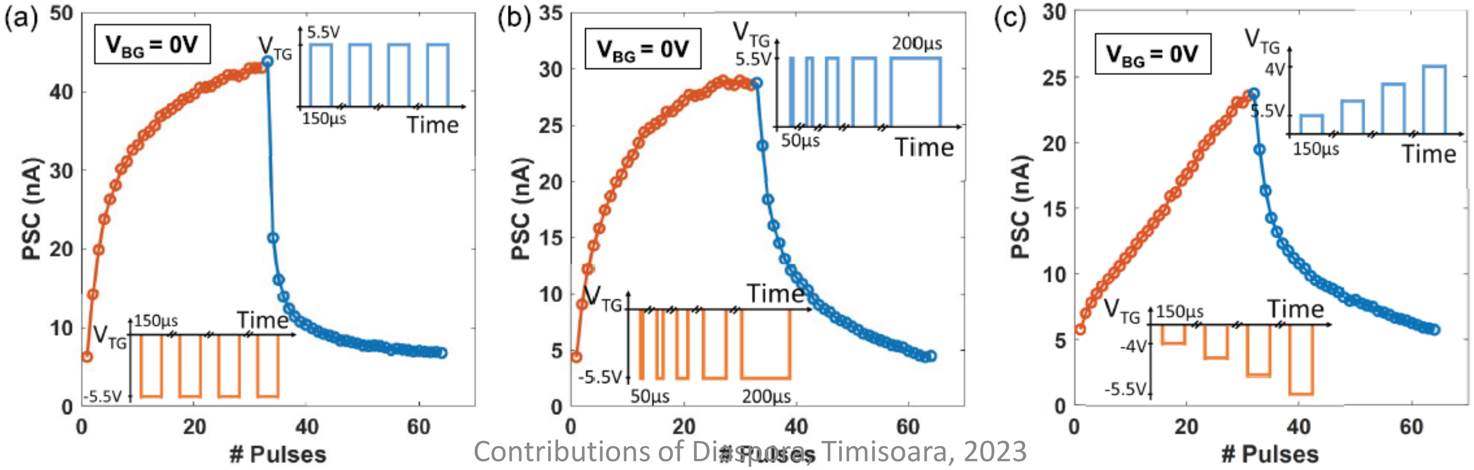
DG FET



Astrocyte modulating function

Post synaptic currents (PSC) programming

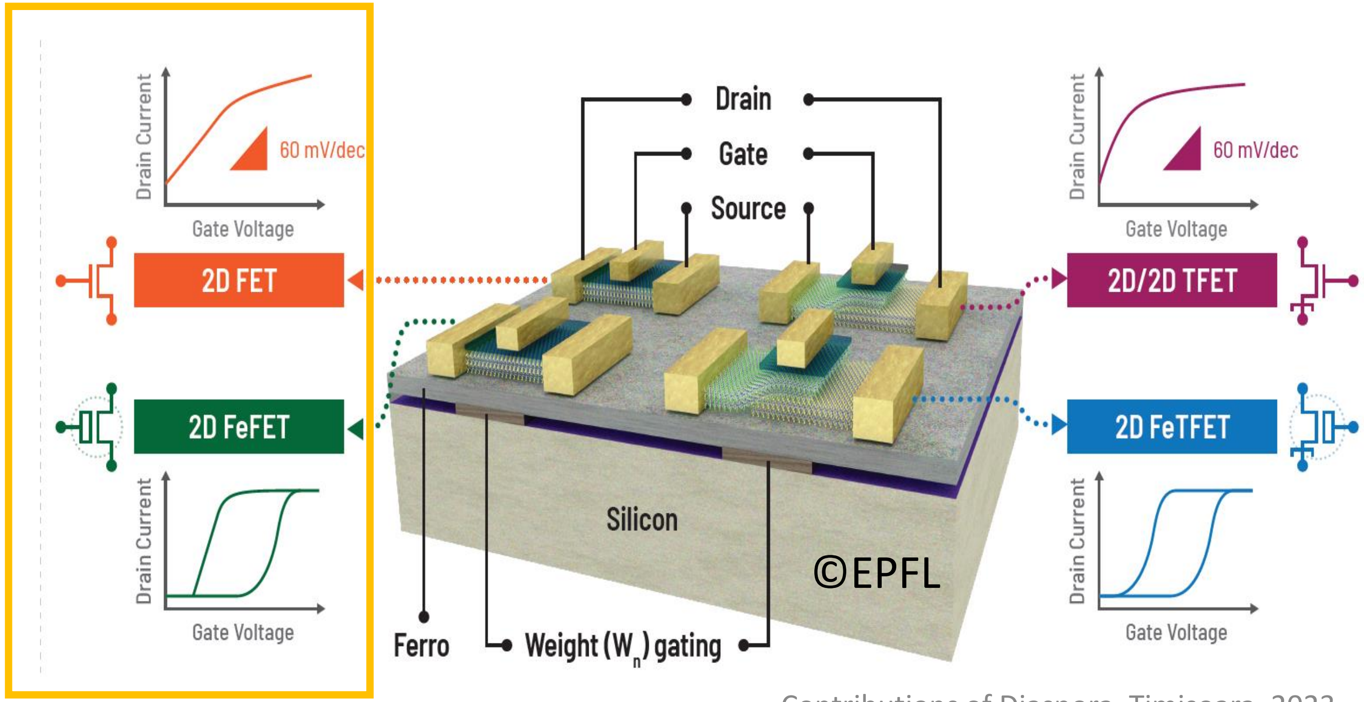


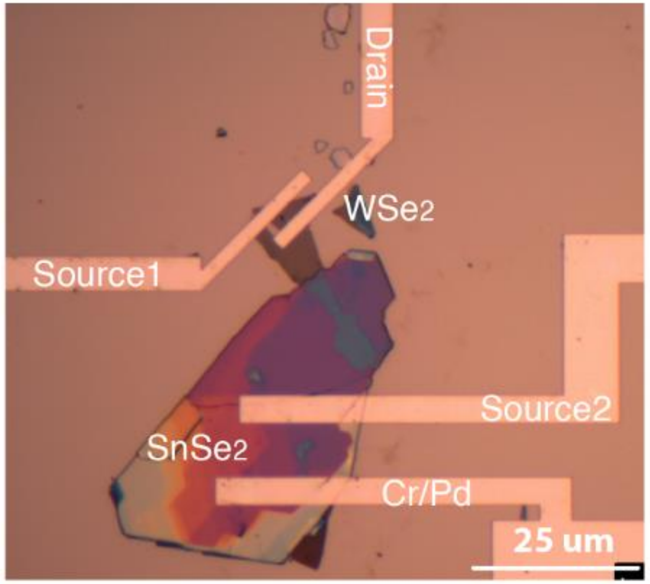


From steep slope switches to low power von-Neuman / neuromorphic technology platforms

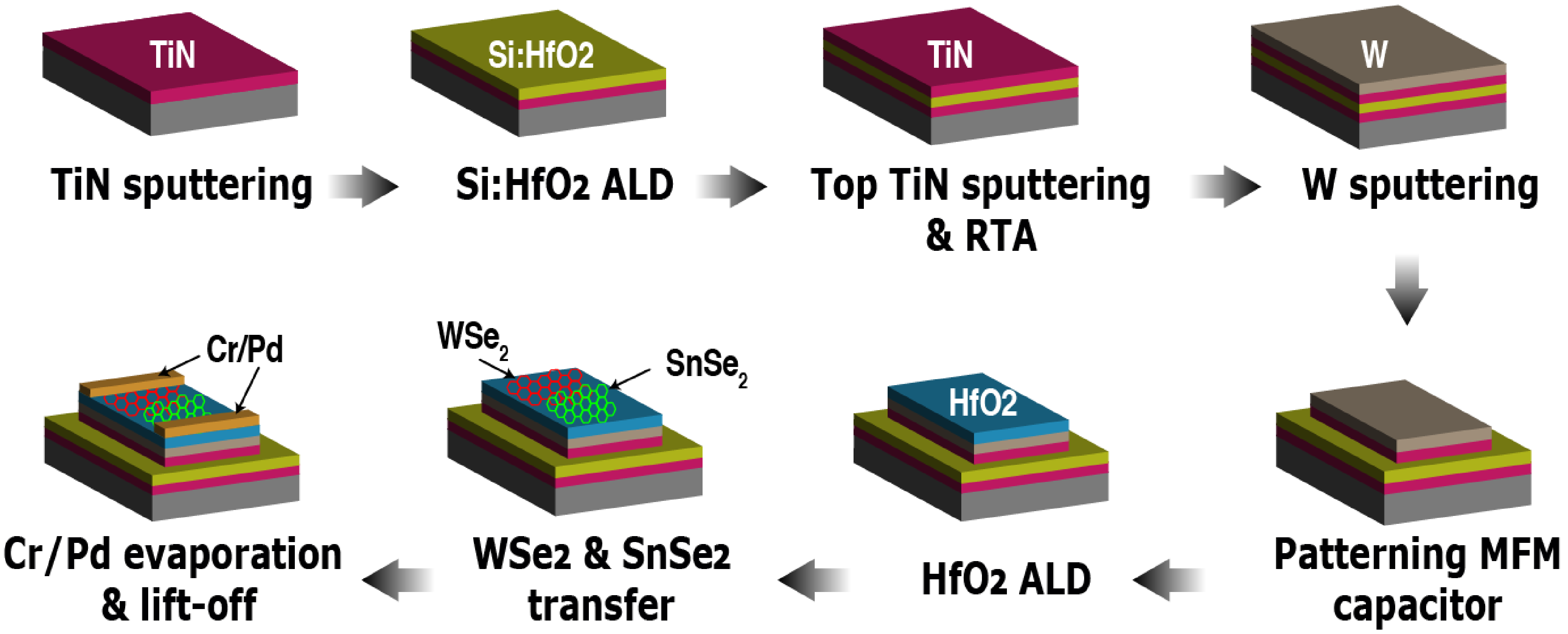
High-k ferroelectrics and 2D materials will from a platform for the co-integration of energy efficient electronics and neuromorphic systems!

- **Challenge: co-integration without performance loss + enlarge the design space!**
- **2D material system: WSe2/SnSe2, ferroelectric: Si-HfO2 or HZO**

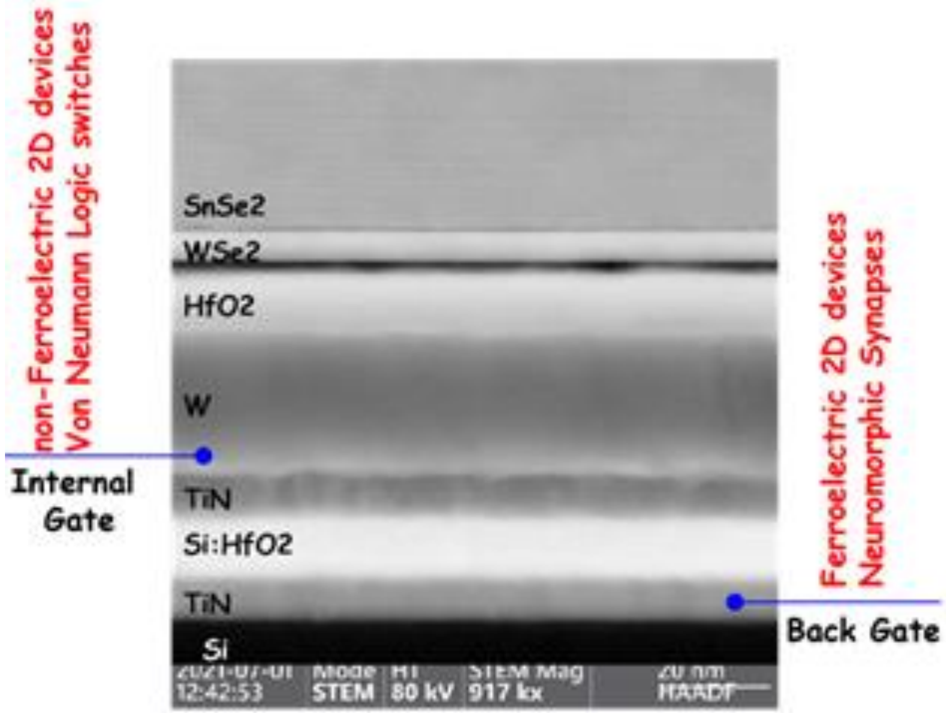


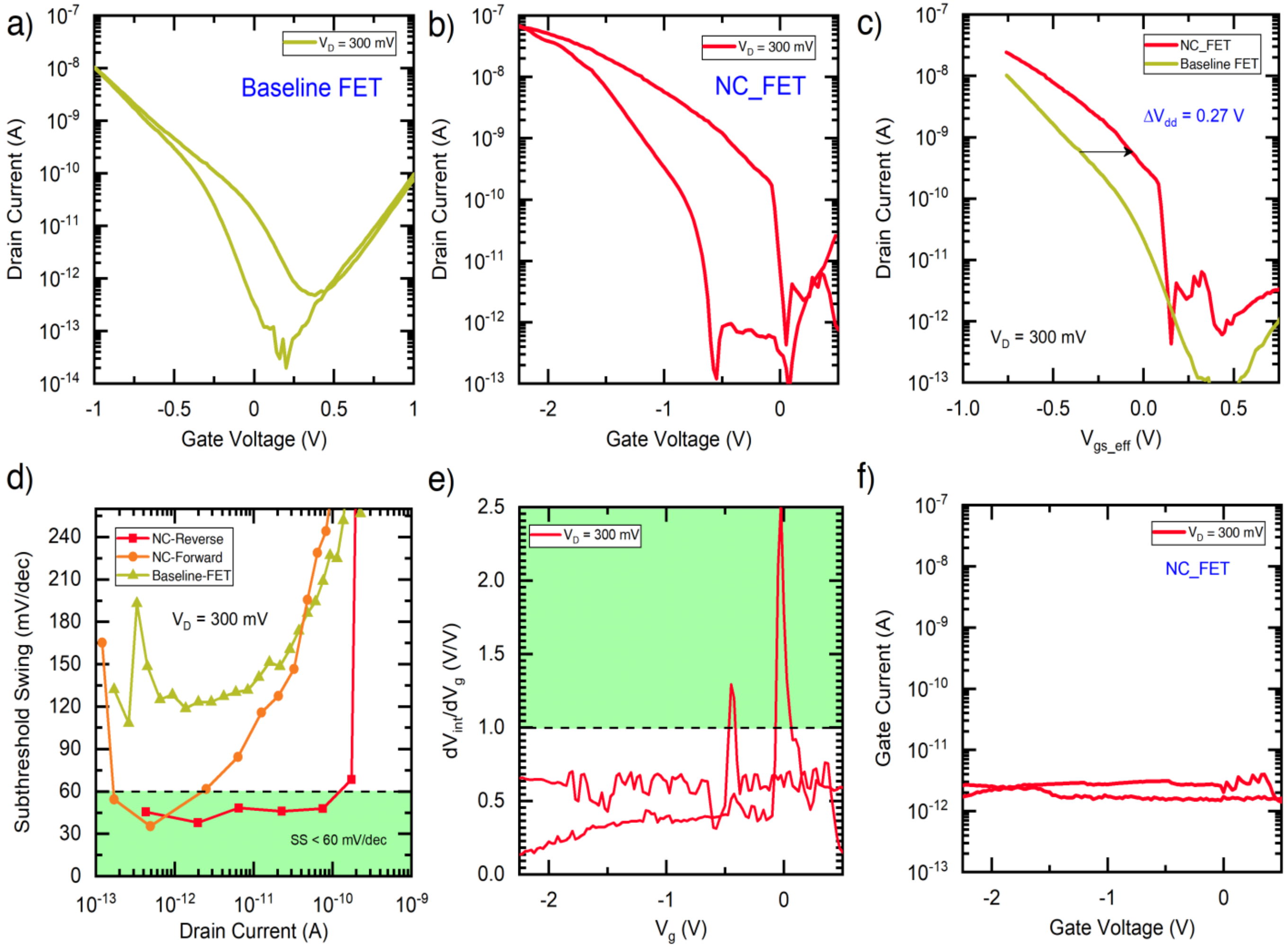


Simplified fabrication of buried Si:HfO2 ferroelectric gating and 2D WSe_2 & $SnSe_2$ transfer

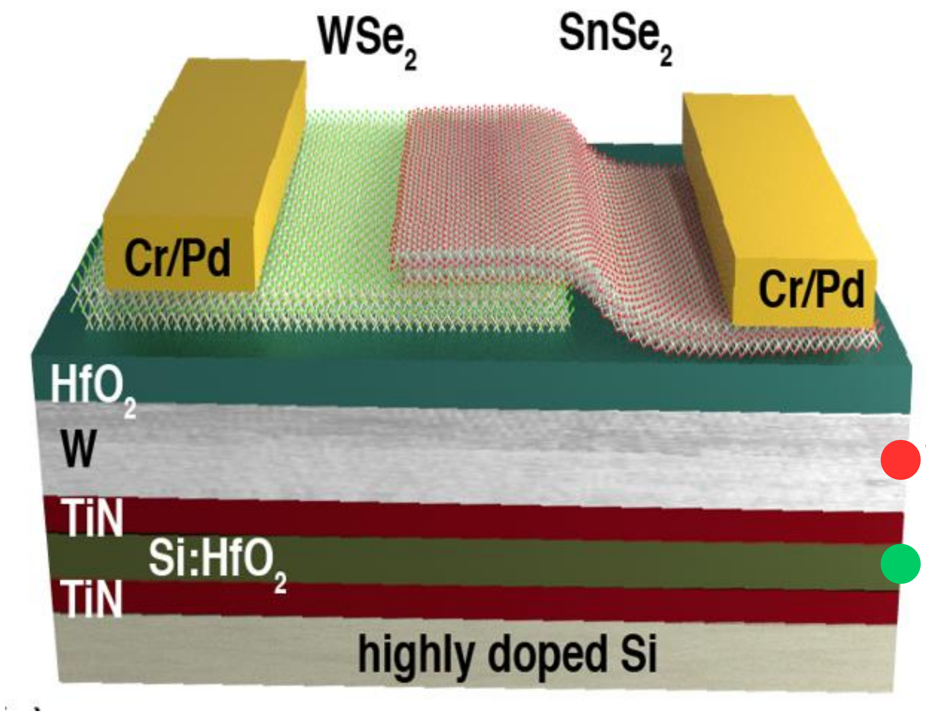


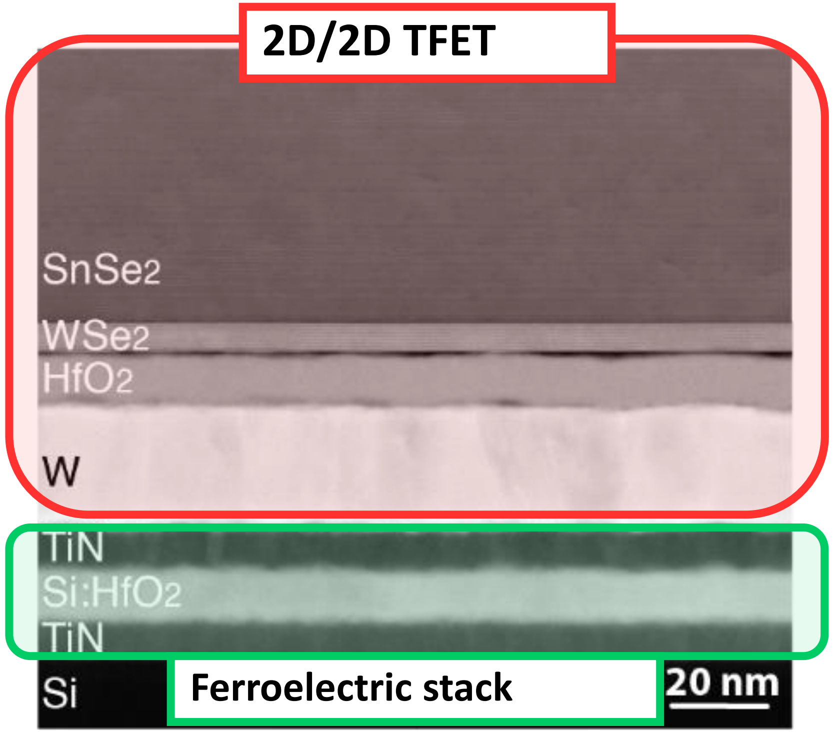
Co-integrated (MOS)FET and NC-MOSFET





2D/2D $WSe_2/SnSe_2$ Tunnel FET with local back gate





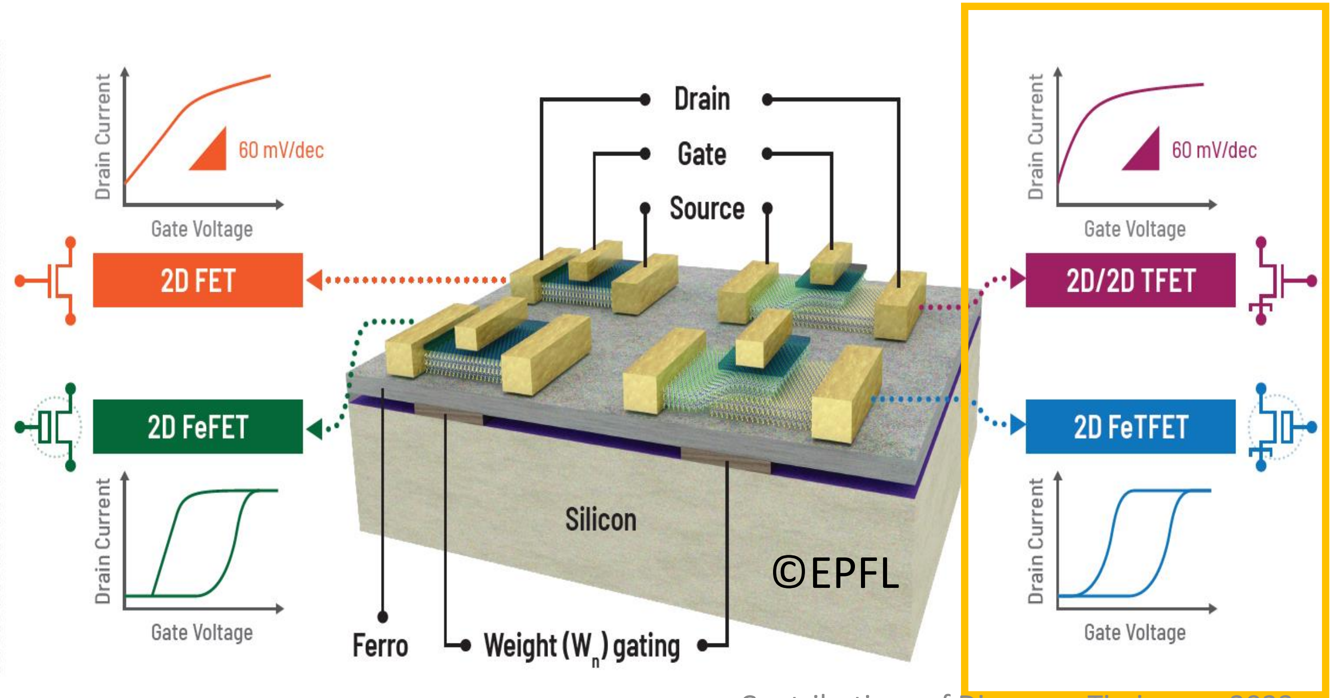
S. Kamaei et al., ESSDERC 2022.

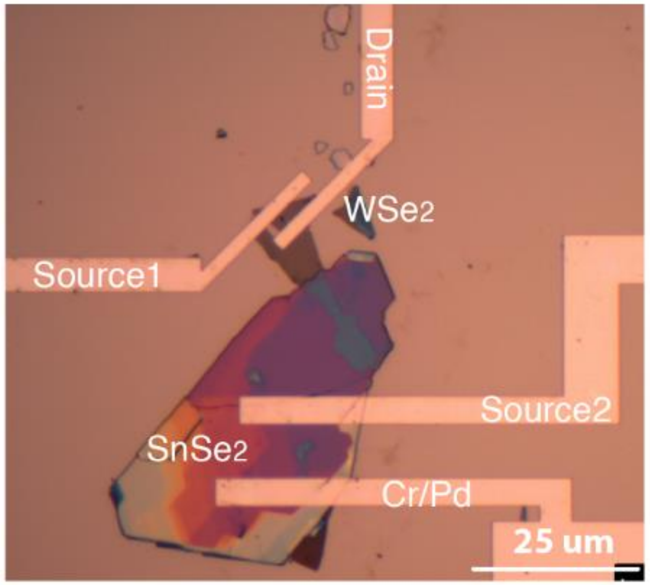
- Deterministic transfer irrespective of lattice mismatch
- Type III, broken-gap band alignment in $WSe_2/SnSe_2$: very good for tunnel FETs

From steep slope switches to low power von-Neuman / neuromorphic technology platforms

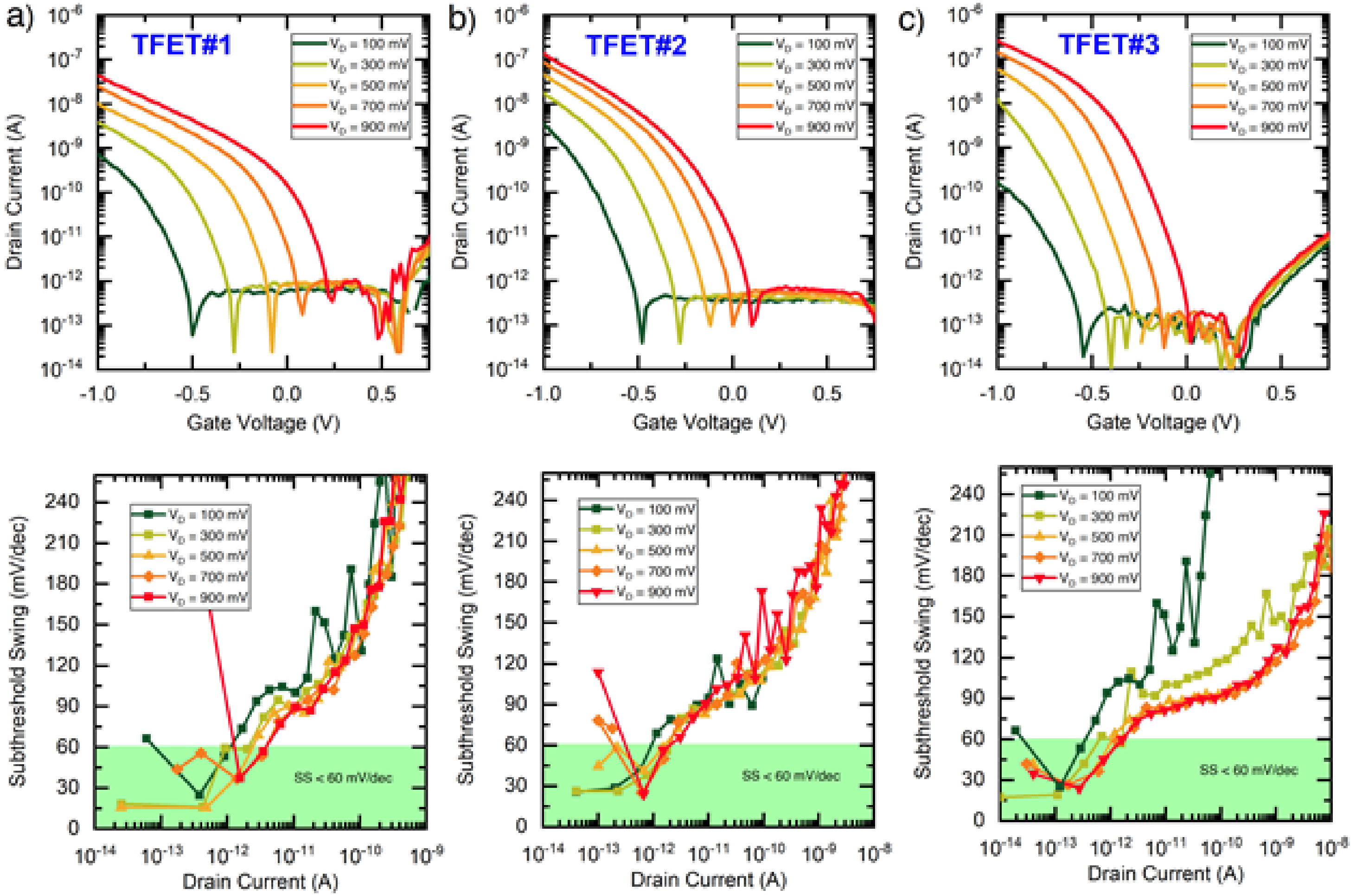
High-k ferroelectrics and 2D materials will from a platform for the co-integration of energy efficient electronics and neuromorphic systems!

- **Challenge: co-integration without performance loss + enlarge the design space!**
- **2D material system: WSe2/SnSe2, ferroelectric: Si-HfO2 or HZO**





2D/2D $WSe_2/SnSe_2$ Tunnel FETs



EPFL Tunnel FET performance:

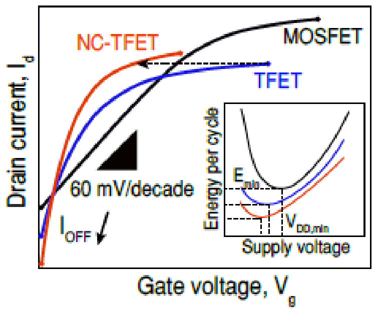
- **SS_{min} of 16 mV/dec**
- **SS_{avg} of 54.5 mV/dec**

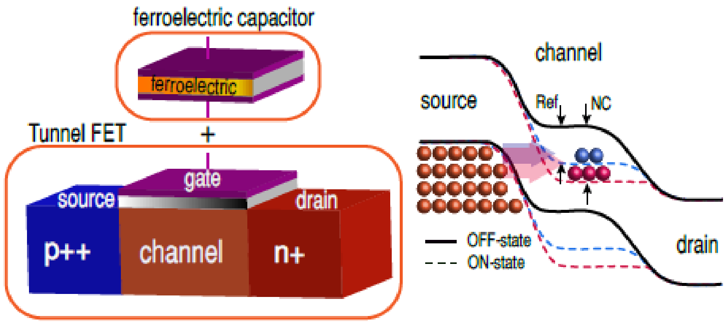
at VD = 300 mV over almost 3 decades of drain current

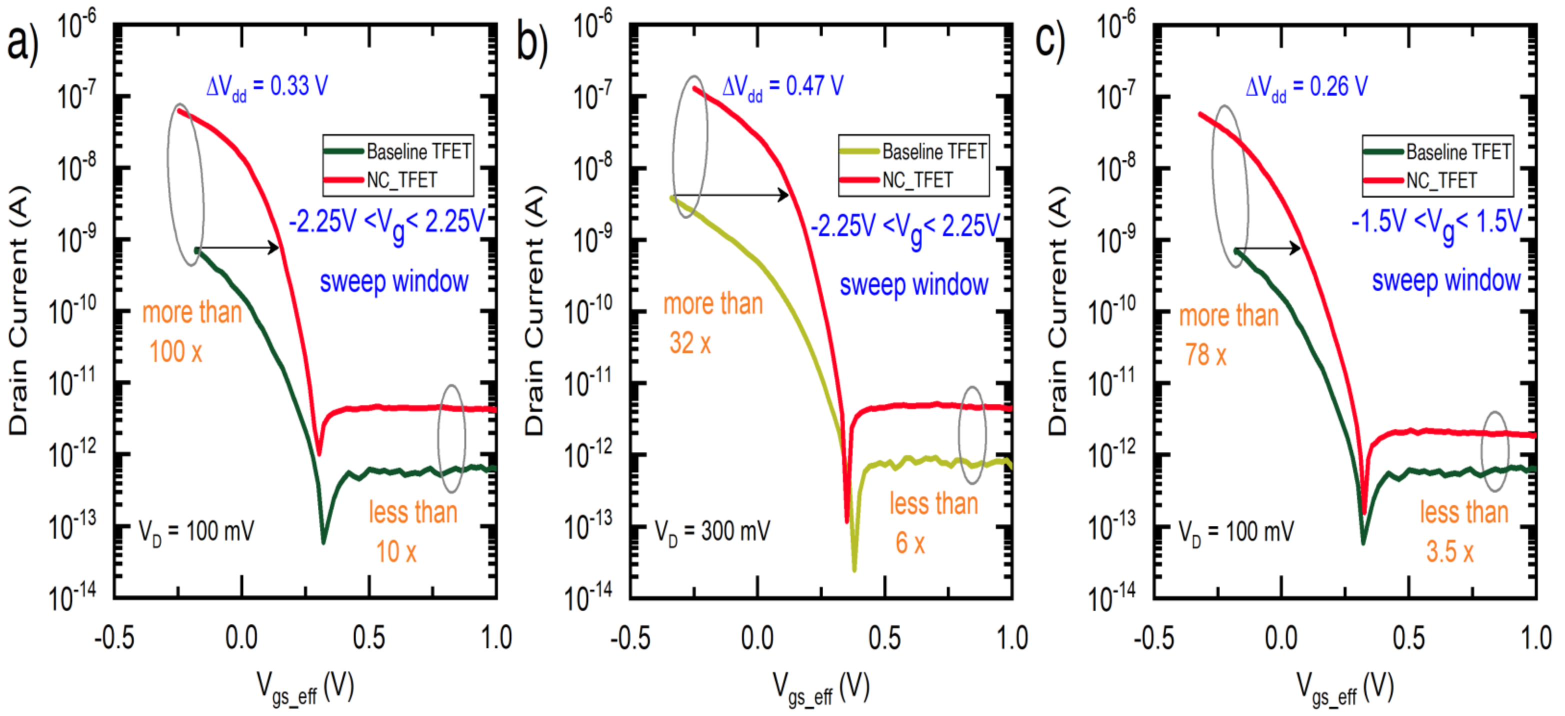
- **Ion/Ioff ~ 10^5**
- Hysteresis ~100mV
- >90% yield
- Variability under study (essentially dictated by flake thickness)

Boosting 2D/2D Tunnel FETs by Negative capacitance with ferroelectric gating: UNIQUE performance gain

$$S = \frac{\partial V_g}{\partial(\log I_d)} = \underbrace{\frac{\partial V_g}{\partial \psi_S}}_{m}\underbrace{\frac{\partial \psi_S}{\partial(\log I_D)}}_{n} = \left(1+\frac{C_s}{C_{ins}}\right)\frac{kT}{q}\ln 10 = \boxed{m \times n}$$



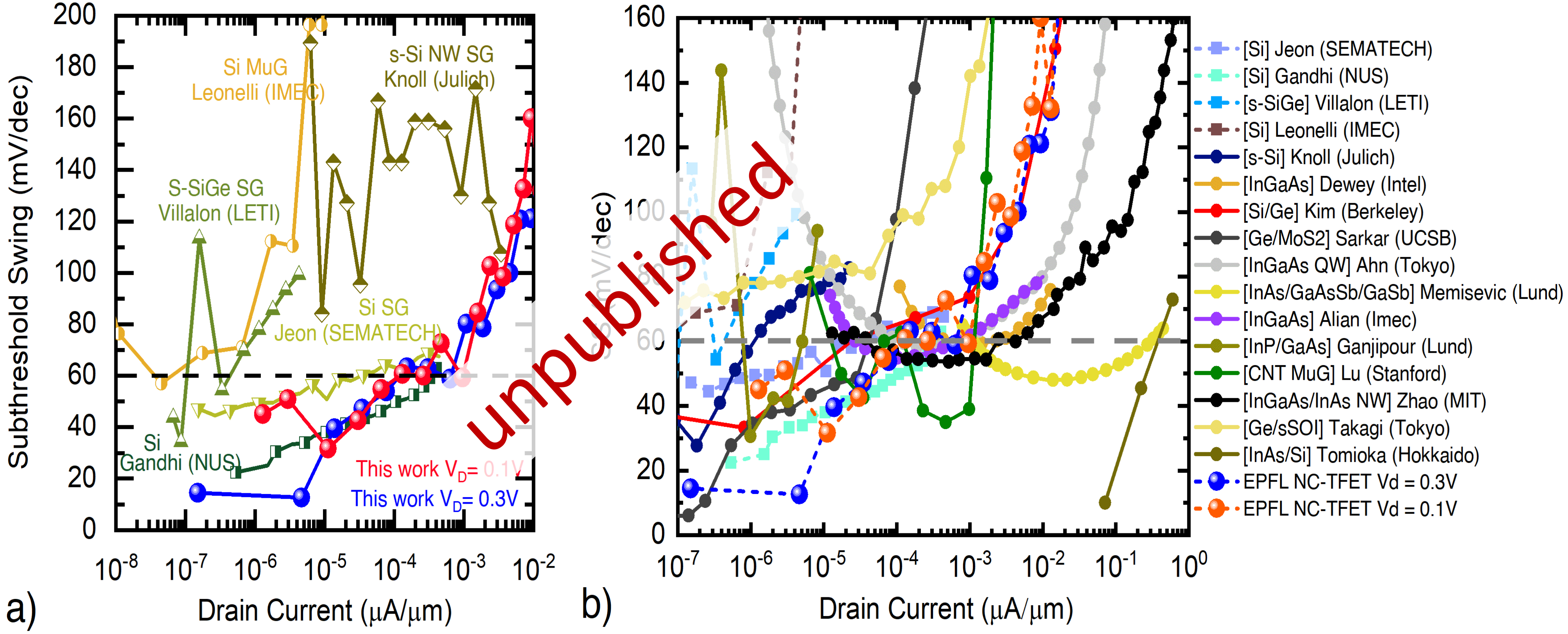




A. Saeidi *et al.*, "Negative Capacitance as Performance Booster for Tunnel FETs and MOSFETs: An Experimental Study," *IEEE Electron Device Letters*, 2017.

- **SS_{min} = 10mV/dec, S_{avg}=18.8 (over 2 dec.)/ 55 (over 4 dec.)**
- **Ion/Ioff= 107, Vdd=0.3V**
- **Sub-60mV/dec region = 4 decades**

Benchmarking: probably one of the best 2D/2D p-type TFET to date (after co-integration)

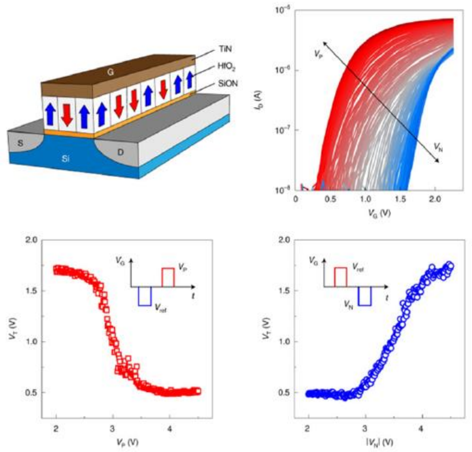


S. Kamaei & A.M. Ionescu, under review (unpublished data 2023).

Neuromorphic co-integrated 2D synapses

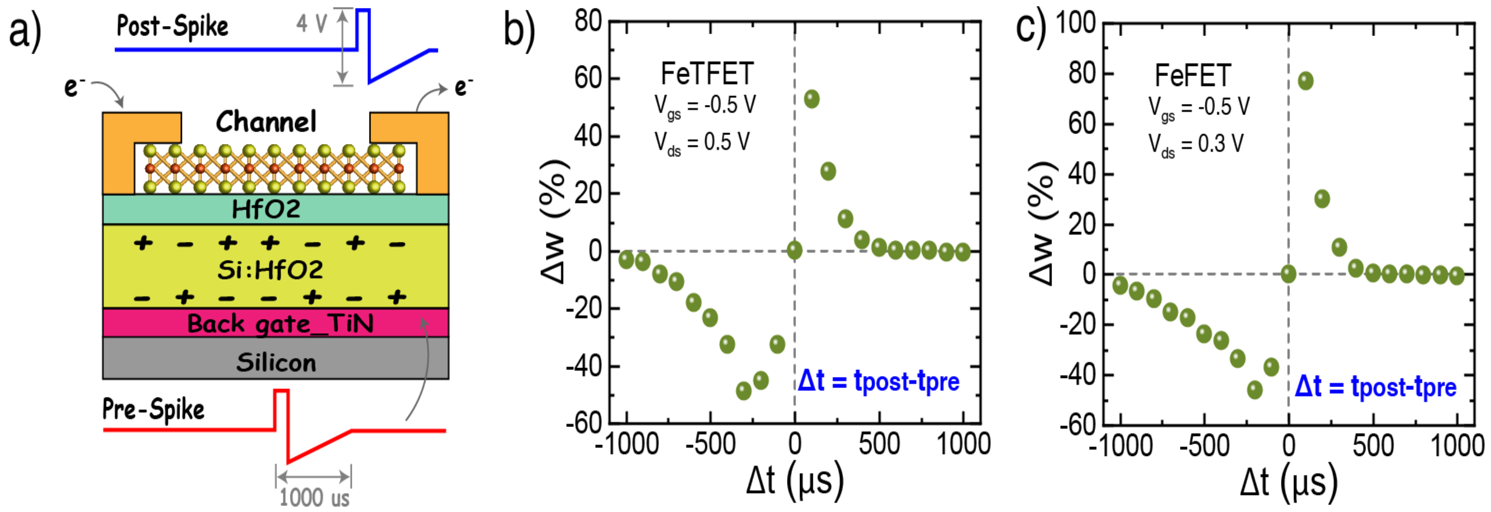
Spike-timing-dependent plasticity (STDP) characterization corresponds to **adjusting the strength of connections between neurons** (mimiking real biological processes).

FeFET synapse concept (NAMLAB)



H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, *Nat. Electron.*, 2020,

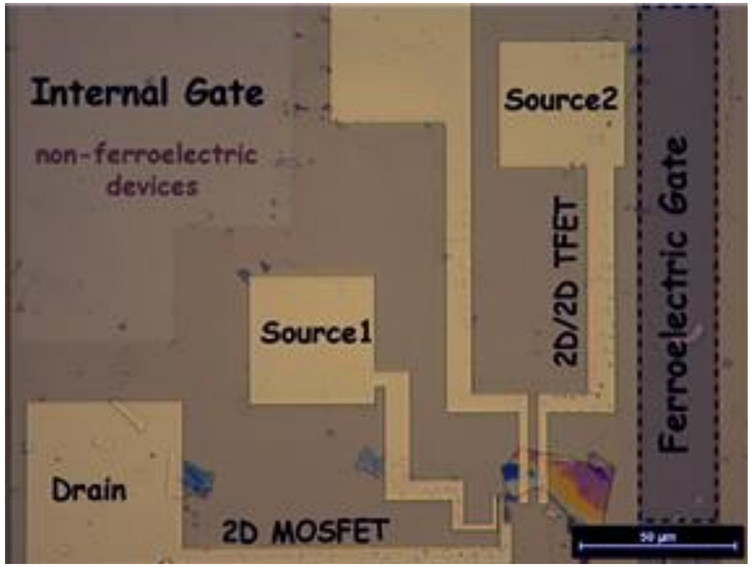
Our 2D synapse characterization



- To emulate STDP learning curves in Fe memtransistors, we apply pre- and post-spikes in the form of voltage pulses with a predefined time difference ($\Delta t = t_{post} - t_{pre}$) to TiN bottom electrode and drain/source electrode, respectively.
- **The synaptic weight (Δw) (= change in channel conductivity) alters based on the timing interval between pulses.**

Summary: future 2D ferroelectric hybrid platforms

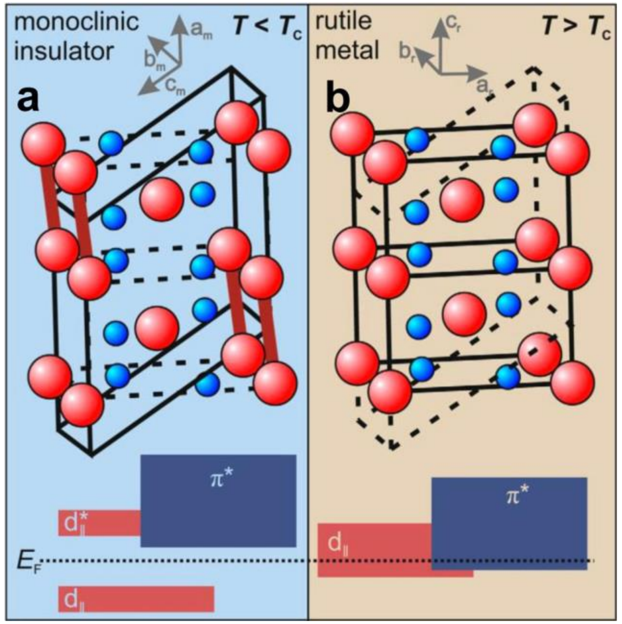
- novel technological co-integration of 2D material systems and ferroelectric gate stacks, **enabling both von Neumann steep slope switches (below 0.3V) and neuromorphic electronic functions**
- **Multi-modal energy efficient operation: four classes of devices on same chip) co-integrated and demonstrated on the same substrate** within a single 2D material system, WSe2/SnSe2, and a single type of gate stack (doped high-k ferroelectric).

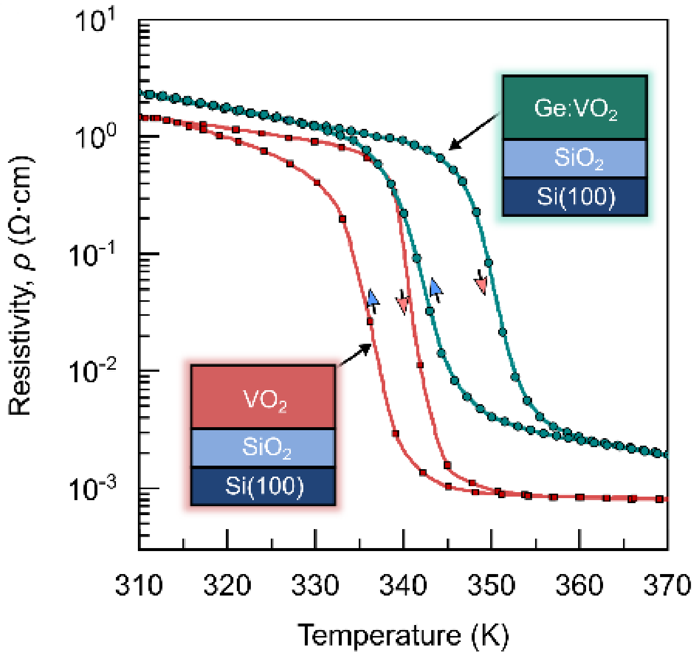


Reversible Metal-Insulator-Transition materials for neuromorphic functions

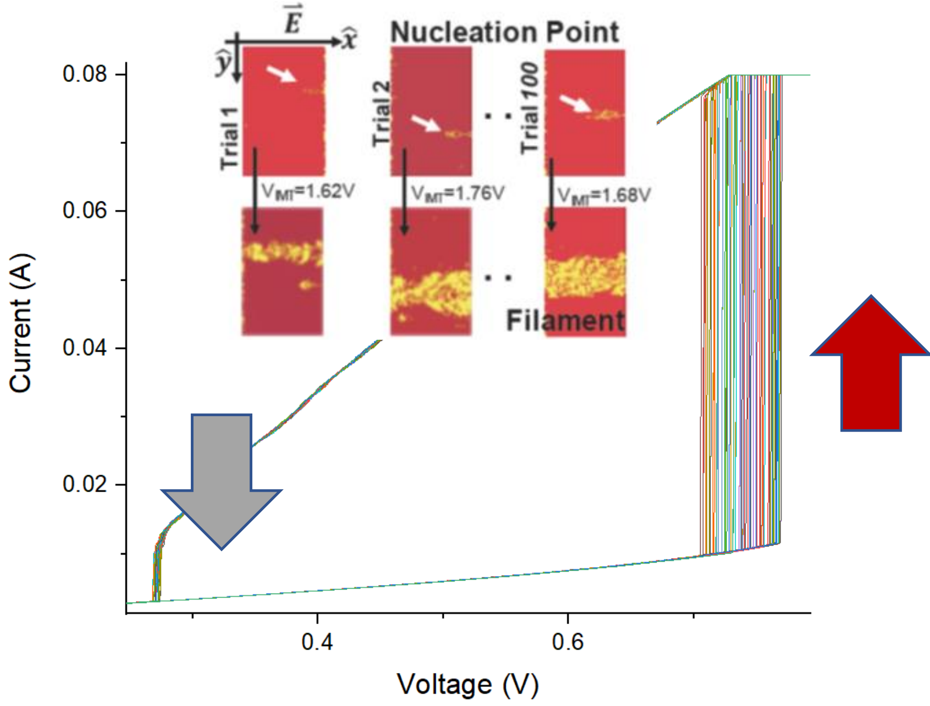
- **Combined Mott-Peierls stochastic IMT/MIT phase transitions** in correlated oxides like vanadium dioxide (VO_2) can be exploited to build memristive sensors.

Temperature induced MIT-IMT



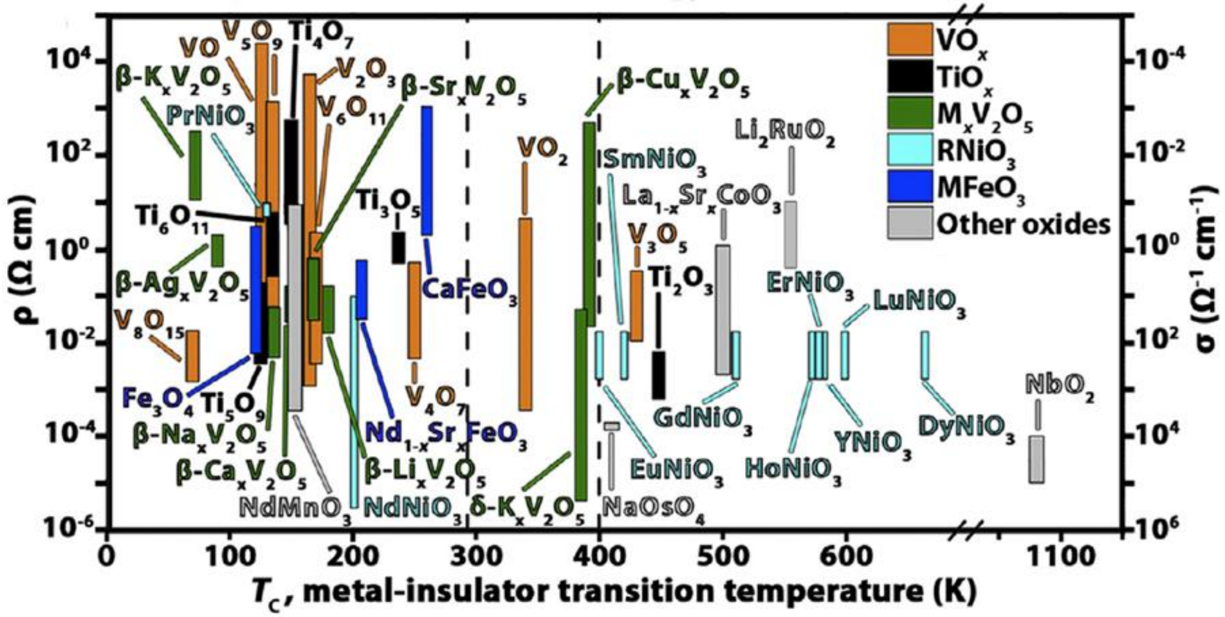


Electrically induced MIT-IMT



MIT materials: practical choices and sustainability

Many emerging IMT/MIT material choices for the future for brain-inspired logic circuits.



J. L. Andrews, et al., Building brain-inspired logic circuits from dynamically switchable transition-metal oxides. Trends Chem. **1**, 711 (2019)

In the earth's crust vanadium is a rather abundant element. It shows a concentration of just under 100 ppm.

Contents lists available at ScienceDirect

Waste Management

journal homepage: www.elsevier.com/locate/wasman

ELSEVIER

Vanadium sustainability in the context of innovative recycling and sourcing development

M. Petranikova [a,*], A.H. Tkaczyk [b], A. Bartl [c], A. Amato [d], V. Lapkovskis [e], C. Tunsu [a]

[a] Chalmers University of Technology, Department of Chemistry and Chemical Engineering, Kemivägen 4, 421 96 Gothenburg, Sweden
[b] University of Tartu, Institute of Technology, Ravila Street 14a, 50411 Tartu, Estonia
[c] TU Wien, Institute of Chemical Engineering, Getreidemarkt 9/166, 1060 Vienna, Austria
[d] Polytechnic University of Marche, Department of Life and Environmental Sciences-DiSVA, Via Brecce Bianche, 60131 Ancona, Italy
[e] Riga Technical University, Scientific Laboratory of Powder Materials & Institute of Aeronautics, 6B Kipsalas Str, Lab. 110, LV-1048 Riga, Latvia

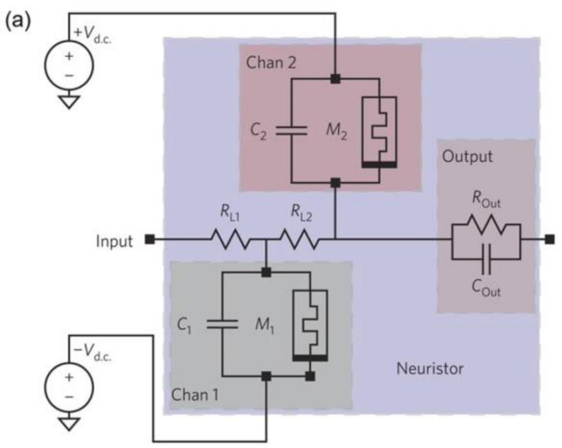
Table 1
Supply risk of selected critical elements and their concentration in the upper continental crust, Earth's crust, and seawater.

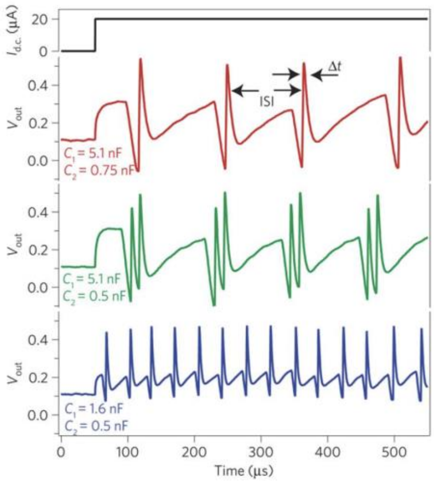
Element	Supply risk* [–]	Upper continentEC,al crust* [ppm]	Earth's crust [ppm]**	Seawater** [μg/m^3]
In	2.4	0.06	0.25	0.10
Bi	3.8	0.16	0.009	60
Ta	1.0	0.9	2.5	<2.50
Ge	1.9	1.4	1.4	5.00
W	1.8	1.9	1.3	10
Be	2.4	2.1	3.8	0.21
As	–	4.8	1.7	1200
Hf	1.3	5.3	1	3.40
Nb	3.1	12	20	<5.00
Co	1.6	17	18	1.20
Ga	1.4	17	19	1.20
V	**1.6**	**97**	**90**	**2,000**

* Threshold for criticality: ≥1.0 (EC, 2017a).

VO_2 neuristors and 1T-1Rspiking oscillators

VO_2 neuristor

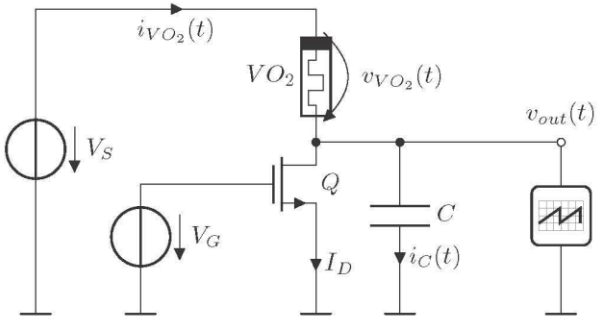


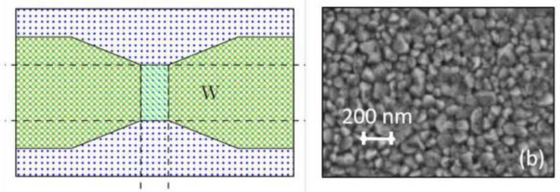


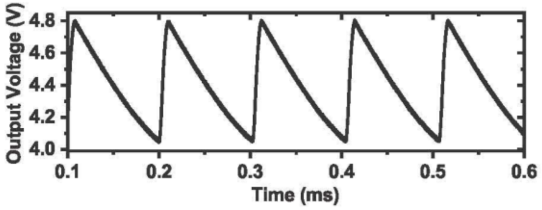
Three types of spiking patterns: regular spiking, chattering, and fast spiking could be achieved in single circuit by adjusting the values of capacitors.

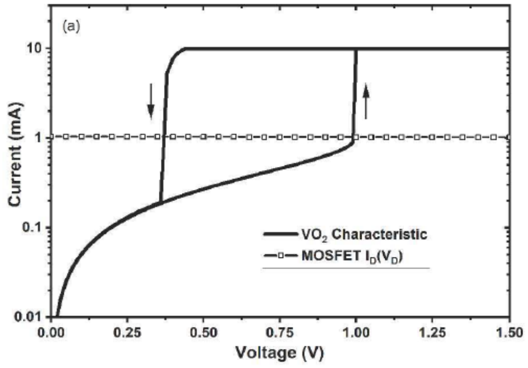
M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, ‘‘A scalable neuristor built with Mott memristors,’’ Nature Mater., 2013.

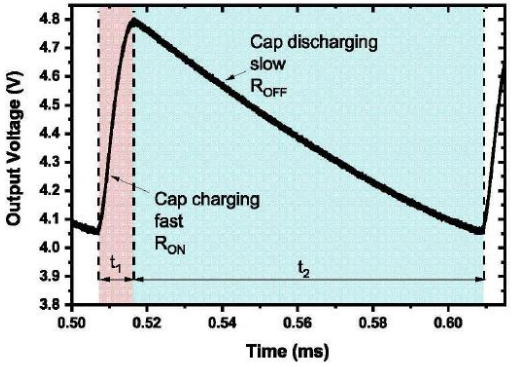
Highly tunable VO_2 spiking oscillator











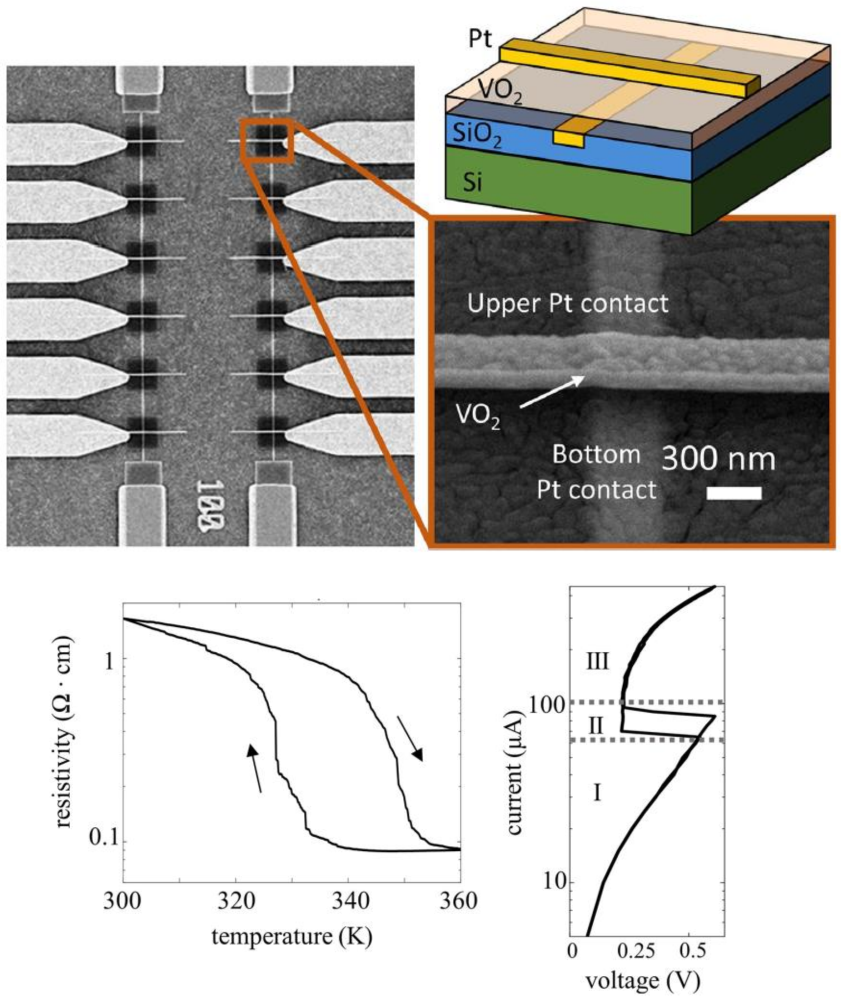
Physical Model:

$$t_1(I_D) = -R_{ON}C \cdot ln\frac{V_{MIT} - R_{ON}I_D}{V_{IMT} - R_{ON}I_D}$$

$$t_2(I_D) = -R_{OFF}C \cdot ln\frac{V_{IMT} - R_{OFF}I_D}{V_{MIT} - R_{OFF}I_D}$$

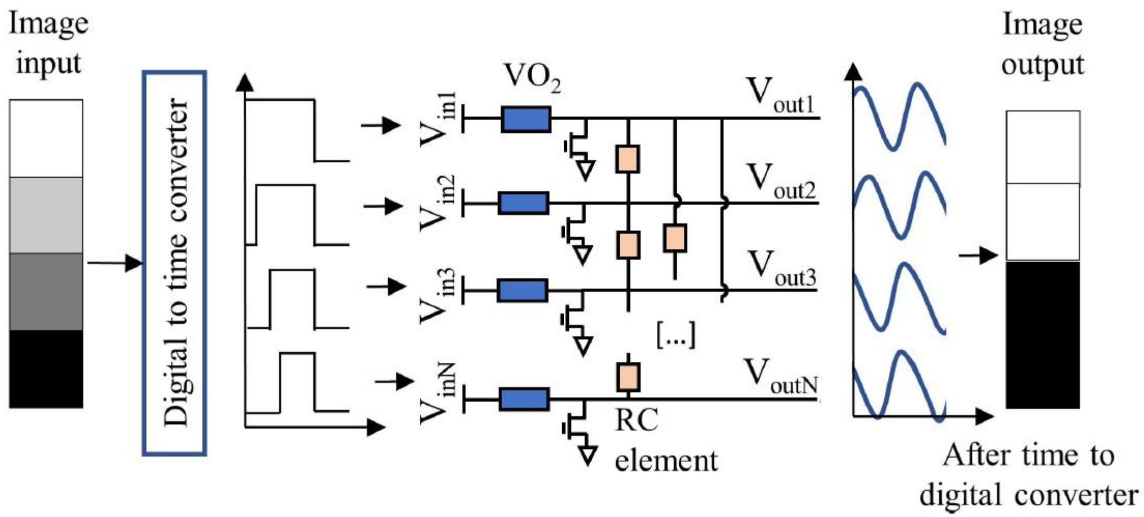
T. Rosca, F. Qaderi and A. M. Ionescu, ESSCIRC 2021.

Coupled VO2 Oscillators Circuit as Analog First Layer Filter in Convolutional Neural Networks



E. Corti et al., Neurosci., 2021.

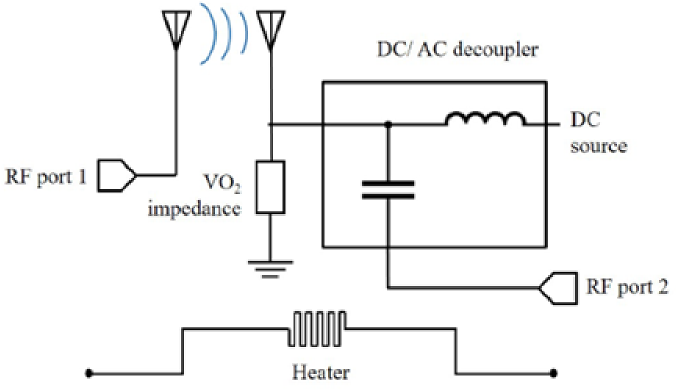
- **in-memory computing** platform based on coupled VO2 oscillators fabricated in **crossbar configuration**
- significant improvements: area density and frequency.
- neuromorphic computing capabilities using the phase relation of the oscillators.
- Application: **replace digital filtering operation in a convolutional neural network** with oscillating circuits.

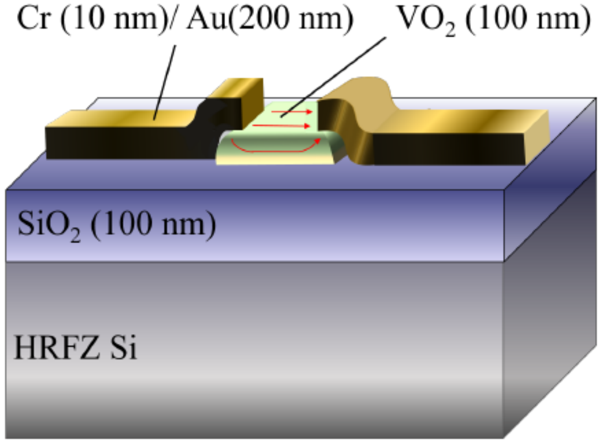


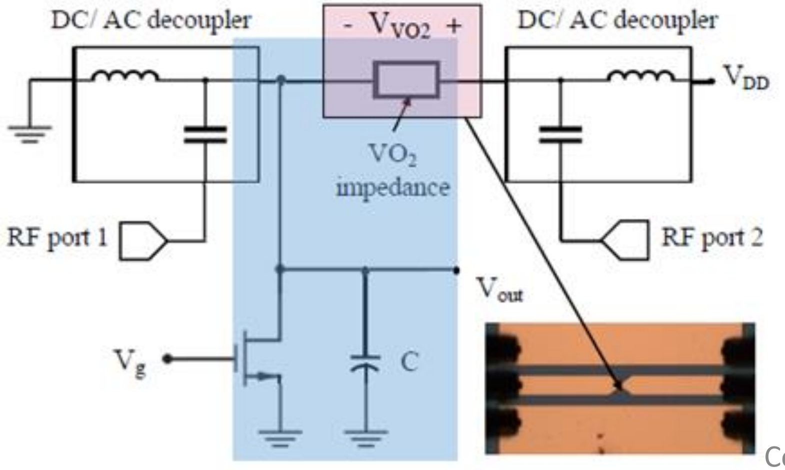
Spiking electromagnetic power sensors (1)

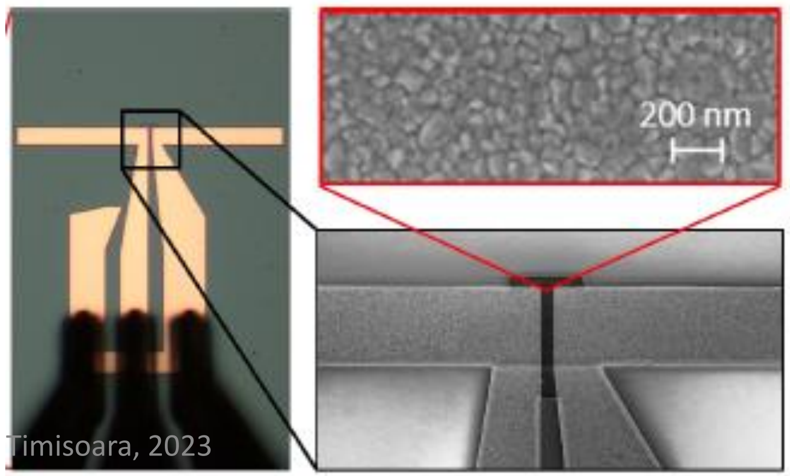
1T-1R power astable oscillator - spiking sensor

Sensitivity to low-energy photons in phase change materials enables the development of efficient millimeter-wave (mm-wave) and terahertz (THz) detectors.



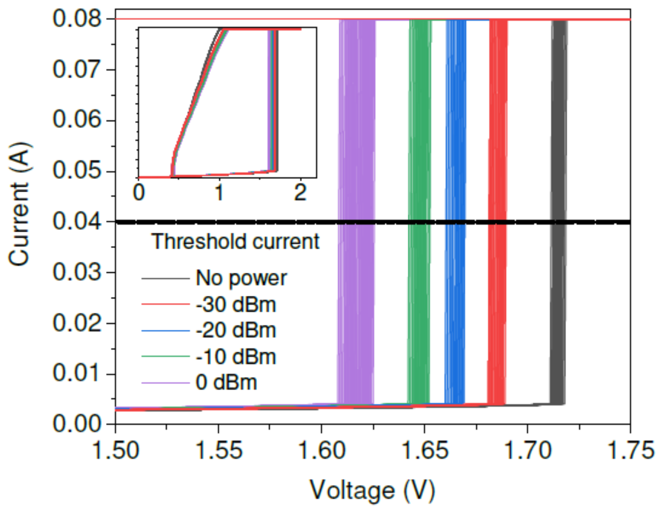


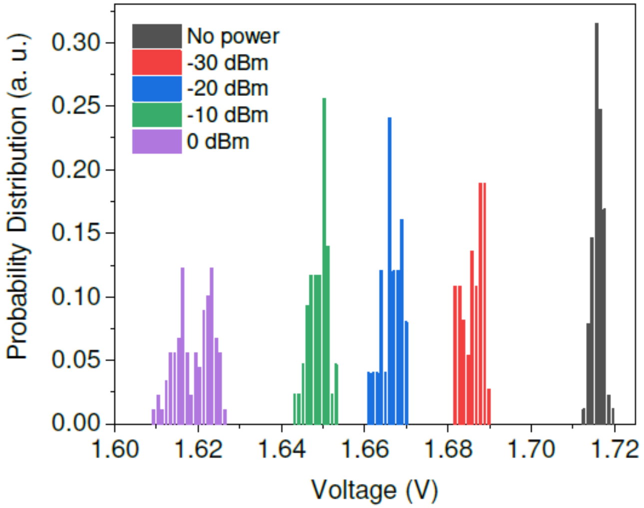


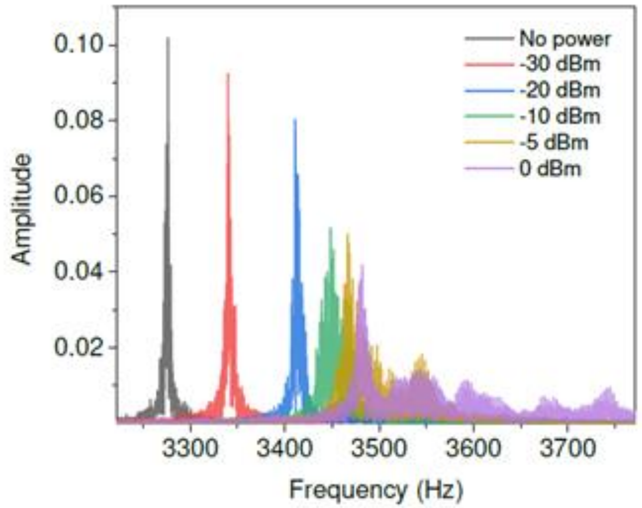


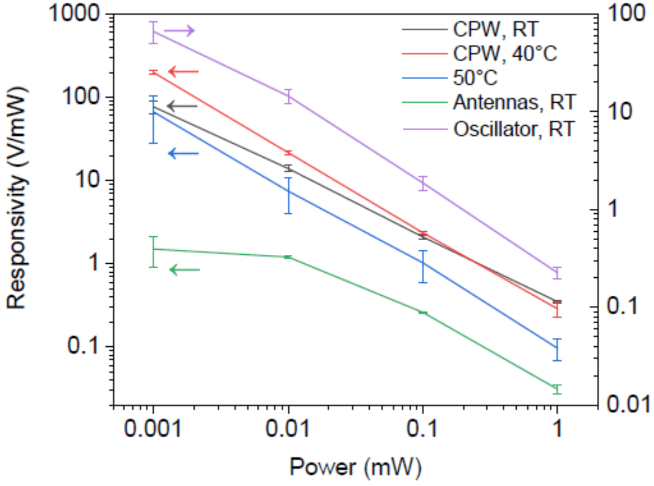
Spiking electromagnetic power sensors (2)

Concept of **uncooled mm-wave detection based on the sensitivity of IMT threshold voltage to the incident** wave by exploiting the characteristics of reversible insulator-to-metal transition (IMT) in Vanadium dioxide (VO2) thin film devices.





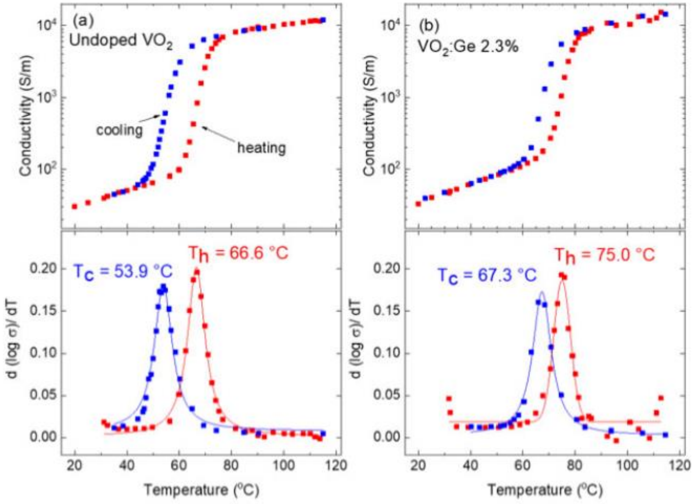


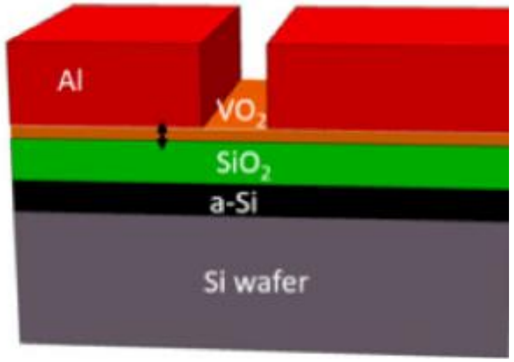


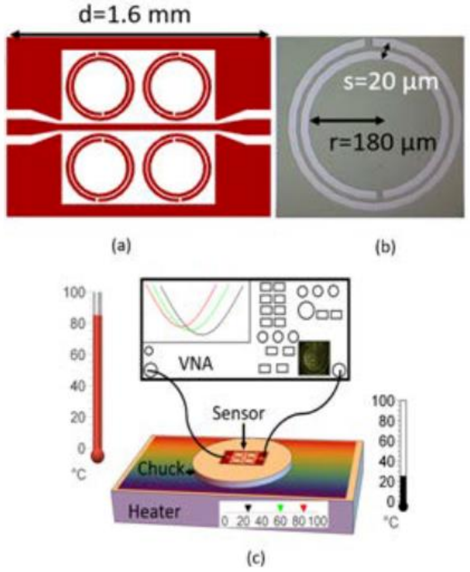
F- Qaderi et al., 'Millimeter-wave to near-THz sensors based on reversible insulator-to-metal transition in vanadium dioxide', to appear, Communications Materials, April 2023.

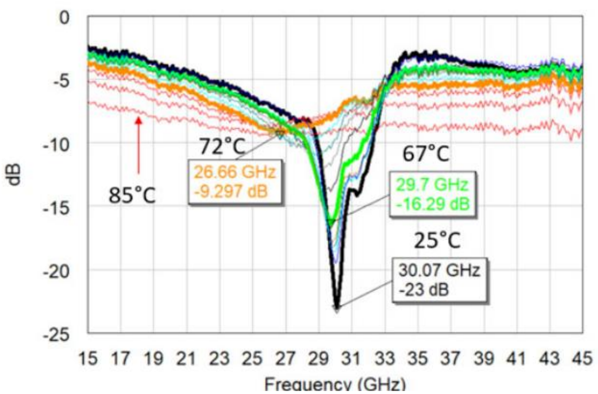
Other applications: VO_2 split ring RF resonator as highly sensitive temperature sensors

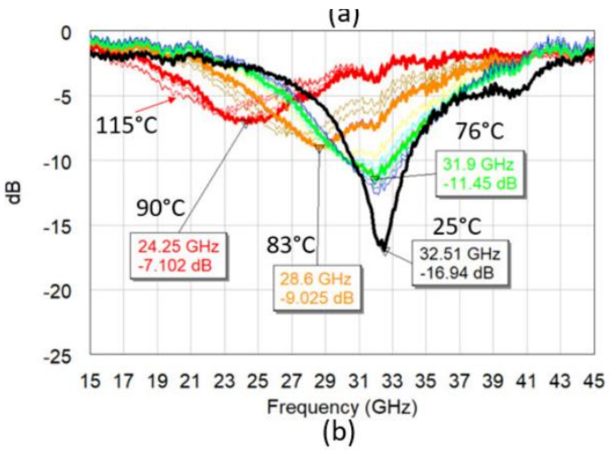
- **split-ring resonator transducers** on 40 nm thick undoped VO2 and on 2.3% Ge-doped (VO2 ALD thin films below coplanar waveguide - CPW).
- **temperature sensing principle is based on the non-linear dielectric constant variation of VO2 around its transition temperature.**











A. A. Muller et al., 2021 21st International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers).

Conclusion

- Edge AI applications need **sustainable technologies** for deployment in large numbers: energy efficiency, reduced data proliferation, abundant and non-toxic materials and processes.
- Novel electronic functionalities for **hybridization of traditional and neuromorphic hardware** can be achieved based on some emerging material and device innovations such as: **ferroelectricity in doped high-k dielectrics, multi-gated 2D semiconducting devices and memristive phase change (MIT) materials and devices**.
- future technological effort from material to system level may permit the 3D **co-integration of spiking neuromorphic hardware** on advanced CMOS chips.