How to build a green AI?

Adrien F. Vincent

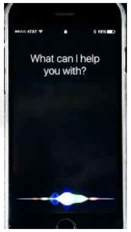
Hardware Artificial Intelligence group

IMS Lab, University of Bordeaux

adrien.vincent@ims-bordeaux.fr

Energy consumption of AI

Artifial Intelligence (AI) may be found in many applications.

- First version: 250 kW
  - One game lasts 2 h
- Second version: 10 kW

~ 850 km!

20h!

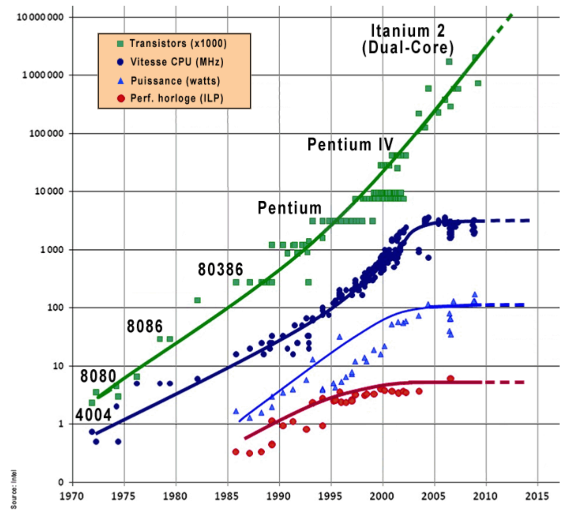
Energy consumption of AI – cont.

- Vehicle battery: 52kW·h (Zoe in 2021)
- Drone Matrice 600 Pro
  - Weight 10kg
  - Payload 6kg during 38’
- ICT
  - 7% of world electricity consumption
  - Forecast for digital electronics: 20% or maybe 50% in 2030

May not be sustainable!

Processors are facing physical limits

- Thermal dissipation
  - 100 W/cm²



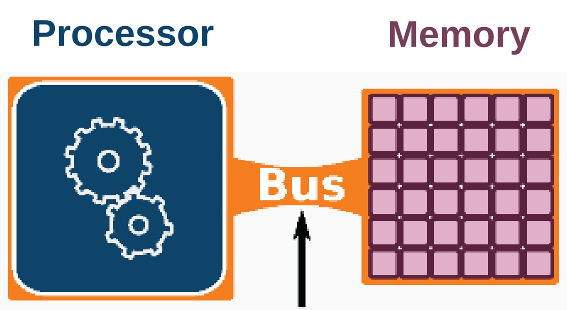
Transistors (x1000)

CPU frequency (MHz)

Power (W)

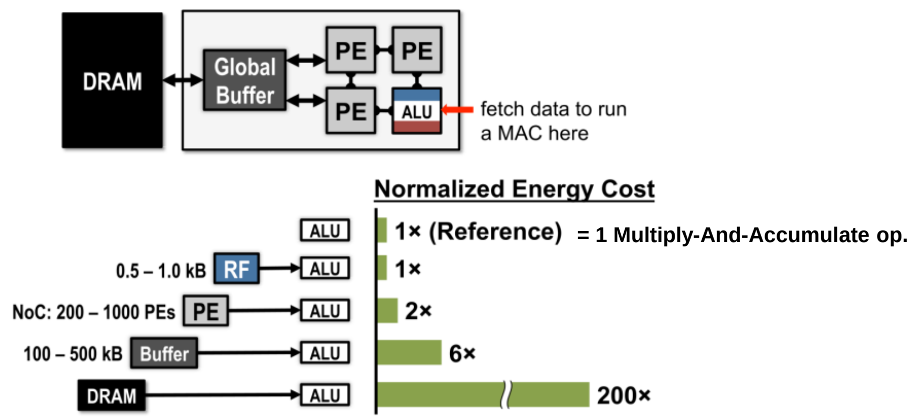
Instr. Level Parallelism

- Memory access



Von Neumann bottleneck

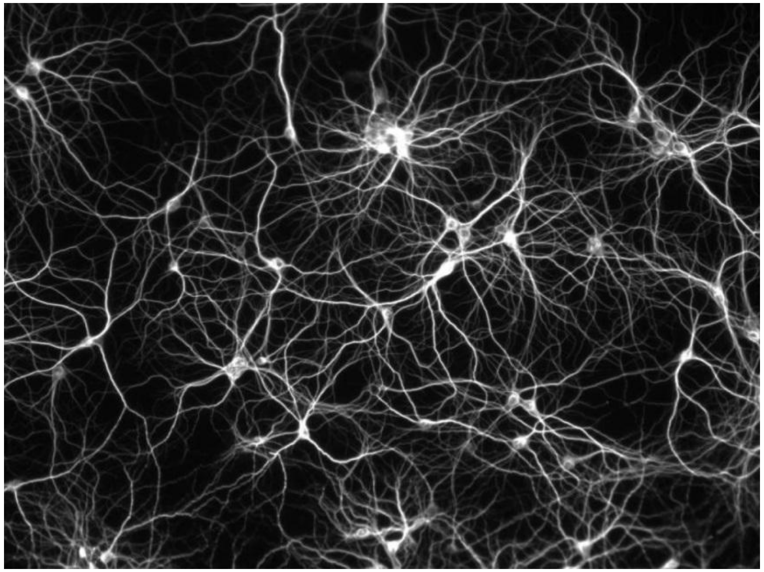
Moving data is expensive!



V. Sze & al., Proc. of IEEE, 2017

An alternative paradigm of computation

- Take inspiration from biological brains



https://wccftech.com/scientists-artificial-neurons-mimics-human-brain-cells/

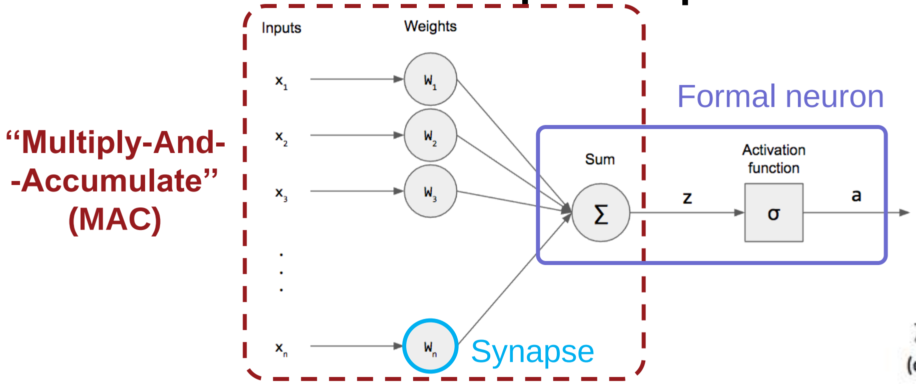
Human brain:

- ~100 billions of neurons
- ~1000 synapses per neuron
- Neuron frequencies ~10Hz to ~100Hz
- Massively parallel computation
- ~20W
- Efficient for high-level tasks (translation, recognition, synthesis)

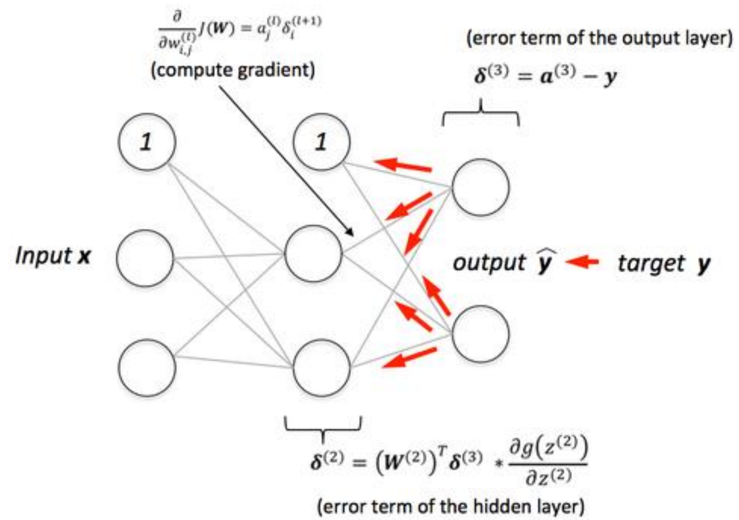
Original computing architectures to explore!

Artificial (and Deep) Neural Networks

- Based on the perceptron concept



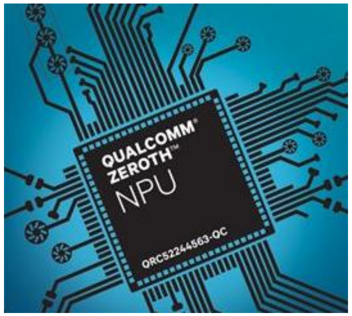
- Supervised learning



- OK but implemented on Von Neumann computers…

Dedicated neural network processors

Qualcomm (2013) : Zeroth



Google (2016) : TPU

Intel (2017) : Myriad

IBM (2014) : TrueNorth

Spiking neural networks

Intel (2017) : Loihi

IC purely CMOS
Learning is not easy

Ideas for building a green(er) AI

Let us have a look at “unconventional computing!”

1) Event-based computing with spiking neural networks

2) Memristor-based disruptive hardware implementations

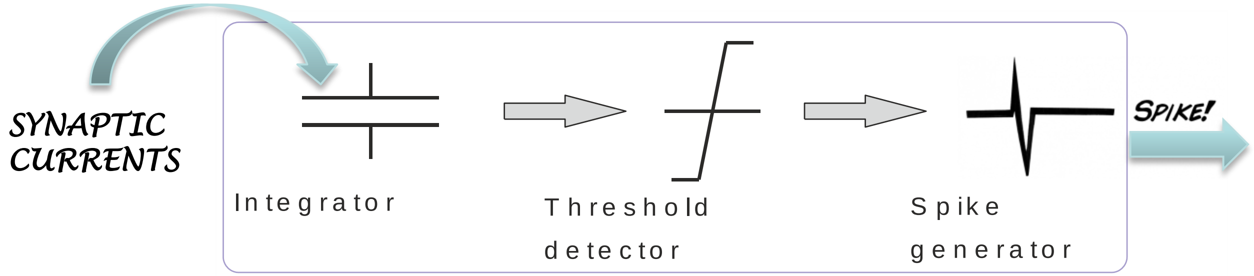
3) Smart System Integration: hardware AI-enhanced sensor
   (ULPEC EU H2020 project)

4) Radio-frequency processing with spintronic nanodevices
   (RadioSpin EU H2020 Project)

Event-based computing with spiking neural networks

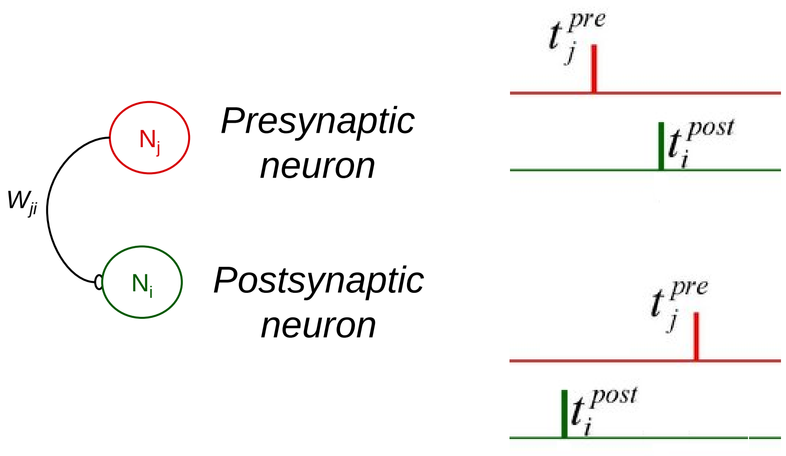
Event-based computing & learning rule

- Spiking neuron = time dependent



- Spike-Timing-Dependent Plasticity (STDP)

Local and unsupervised learning rule!

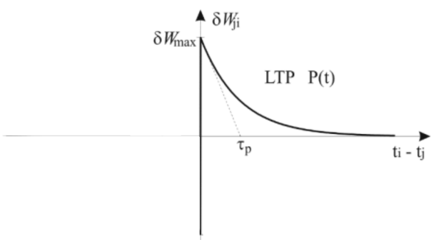


Causality

Increase of synaptic weight

No causality

Decrease of synaptic weight



CMOS implementation of neuromorphic systems

frontiers in
NEUROSCIENCE

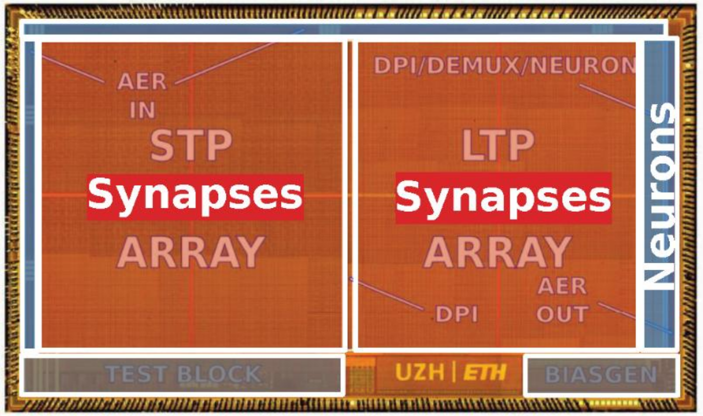
REVIEW ARTICLE
published: 31 May 2011
doi: 10.3389/fnins.2011.00073

Neuromorphic silicon neuron circuits

Giacomo Indiveri[1]*, Bernabé Linares-Barranco[2], Tara Julia Hamilton[3], André van Schaik[4], Ralph Etienne-Cummings[5], Tobi Delbruck[1], Shih-Chii Liu[1], Piotr Dudek[6], Philipp Häfliger[7], Sylvie Renaud[8], Johannes Schemmel[9], Gert Cauwenberghs[10], John Arthur[11], Kai Hynna[11], Fopefolu Folowosele[5], Sylvain Saighi[8], Teresa Serrano-Gotarredona[2], Jayawan Wijekoon[6], Yingxue Wang[12] and Kwabena Boahen[11]

[1] *Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland*
[2] *National Microelectronics Center, Instituto Microelectronica Sevilla, Sevilla, Spain*
[3] *School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW, Australia*
[4] *School of Electrical and Information Engineering, University of Sydney, Sydney, NSW, Australia*
[5] *Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA*
[6] *School of Electrical and Electronic Engineering, University of Manchester, Manchester, UK*
[7] *Department of Informatics, University of Oslo, Oslo, Norway*
[8] *Laboratoire de l'Intégration du Matériau au Système, Bordeaux University and IMS-CNRS Laboratory, Bordeaux, France*
[9] *Kirchhoff Institute for Physics, University of Heidelberg, Heidelberg, Germany*
[10] *Department of Bioengineering and Institute for Neural Computation, University of California San Diego, La Jolla, CA, USA*
[11] *Stanford Bioengineering, Stanford University, Stanford, CA, USA*
[12] *Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA*

Integrable and learning-capable artificial synapses are still a challenge.

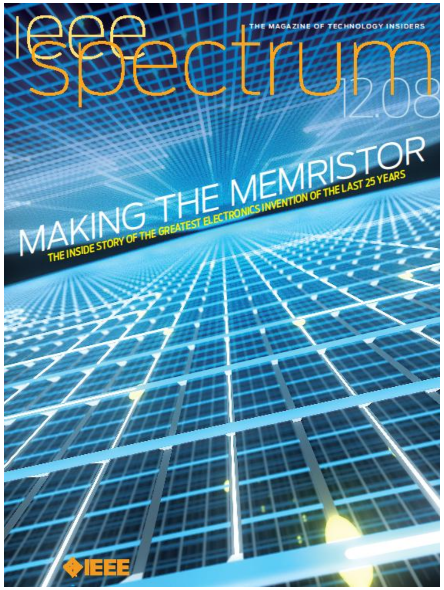


131 072 synapses (~ 65 % surface) **ROLLS** [QIAO, 2015]
256 neurons (~ 3 % surface)

N. Qiao & al., Frontiers in Neuroscience, 2015

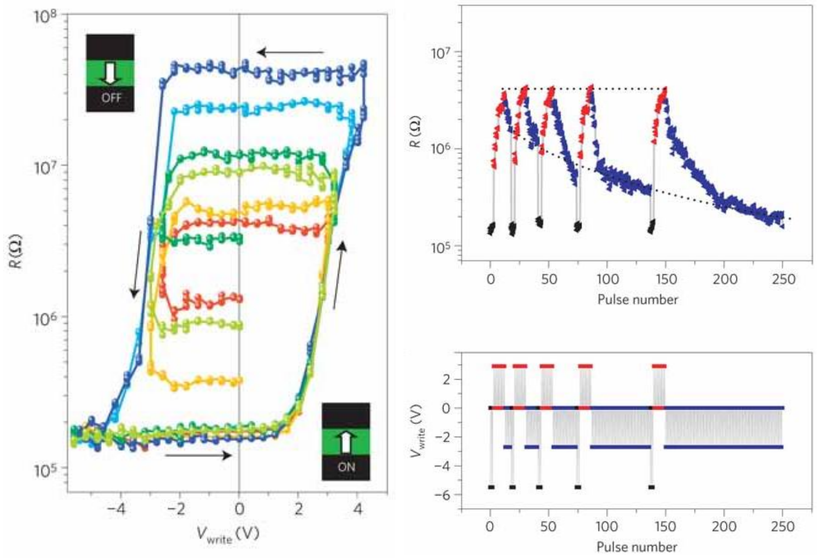
Memristor-based disruptive hardware implementations

Outbreak of memristors



- HP's experimental breakthrough in 2008
- What is a "memristor?"
  - Nanodevice ~ 10s nm x 10s nm
  - "Memory resistor," i.e., a "resistor that learns."
  - Behavior ~ like a biological synapse…

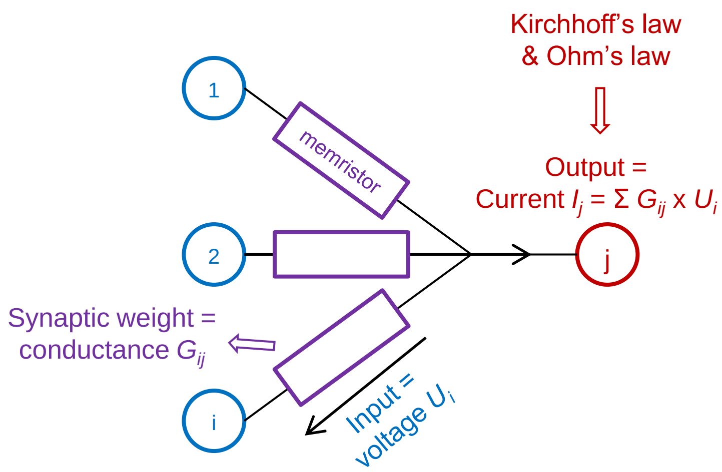
- Non volatile = memory
- Intrinsically plastic
- Several technologies
  (filamentary, phase change memory, spintronics, ferroelectric…)



A. Chanthbouala & al., Nature Materials, 2012

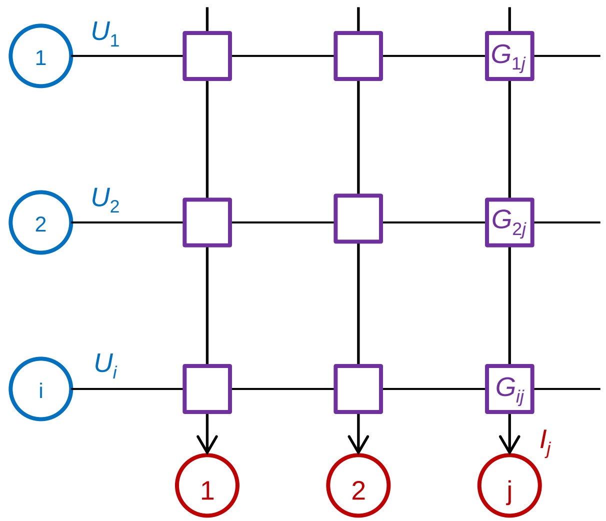
Memristors can help for the inference

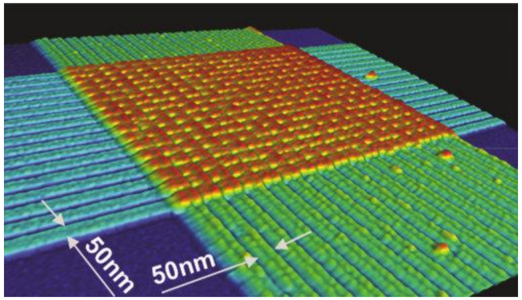
Analog Multiply-And-Accumulate operation



The physics is doing the computing for us!

Memristive crossbar array => Matrix-by-Vector Multiplication

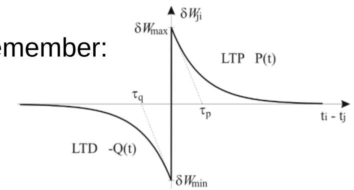




Y. Yang & al., Appl. Phys. Lett., 2012

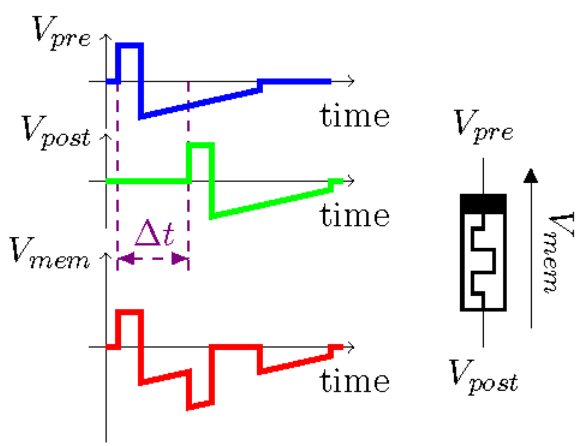
Memristors can also help for the learning

Remember:

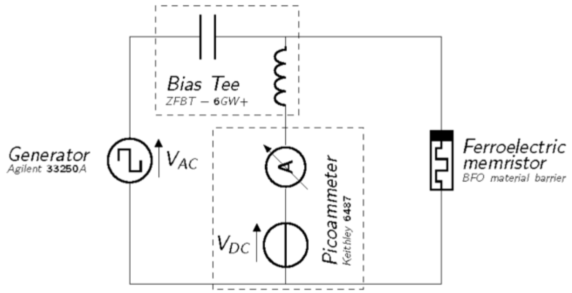


STDP-like behavior in ferroelectric memristors

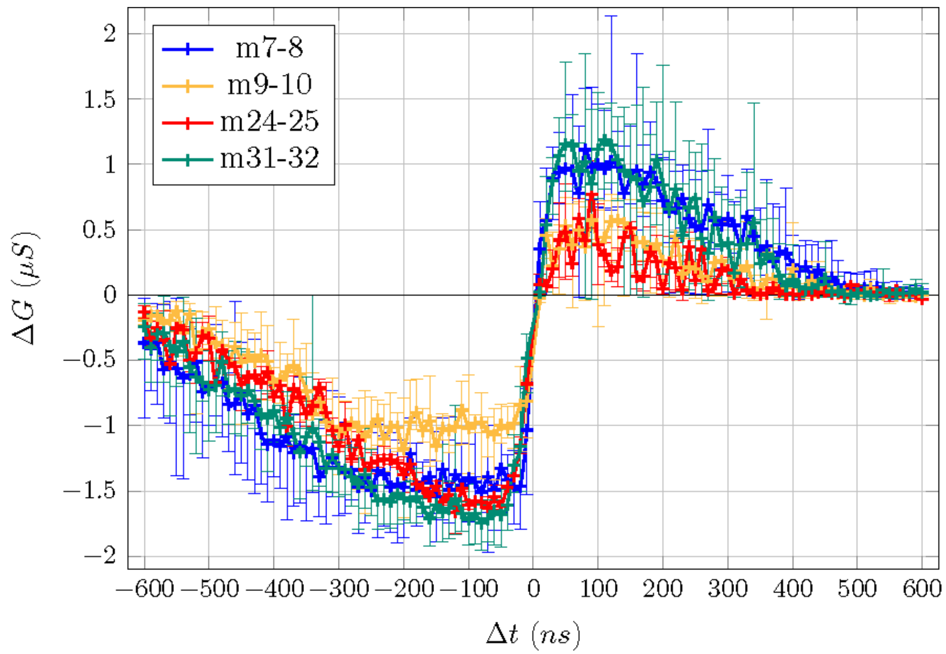
Voltage waveforms



Experimental setup



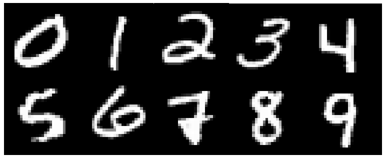
Measurements



Exploiting the intrinsic physics of the device to implement the learning rule.

Memristor-based learning: proof of concept

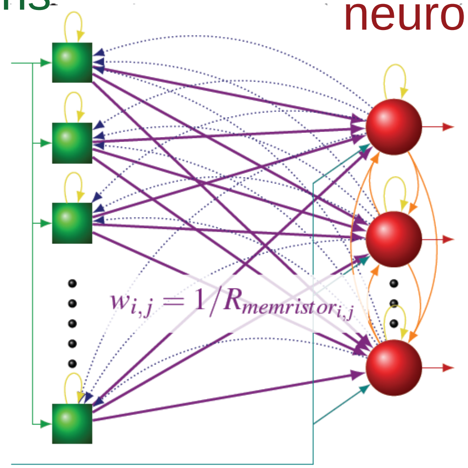
- M-NIST dataset
  (60k pictures for learning and 10k pictures for testing)
- 28 x 28 pixels = 784 input neurons
- 100 output neurons
- Memristor variability is taken into account
- STDP learning rule



784 input neurons

100 output neurons

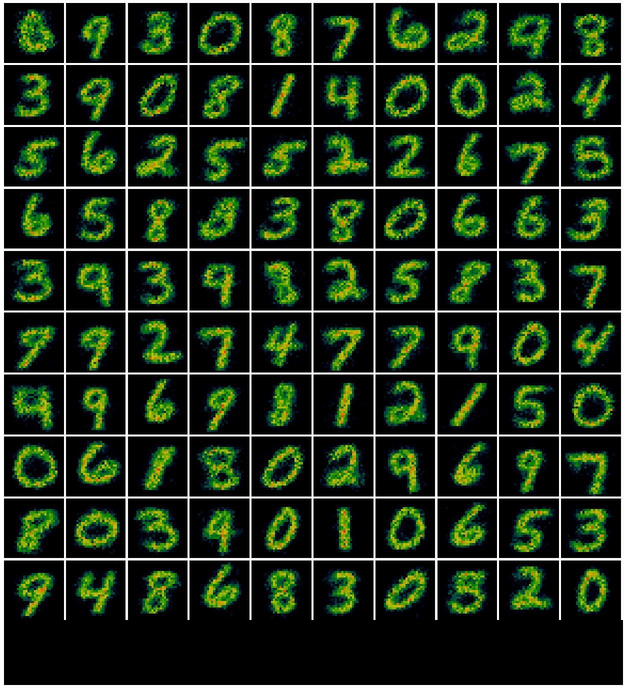
pixel events



- 100 boxes = 100 neurons
- 1 pixel = 1 synapse

After [illegible] pictures

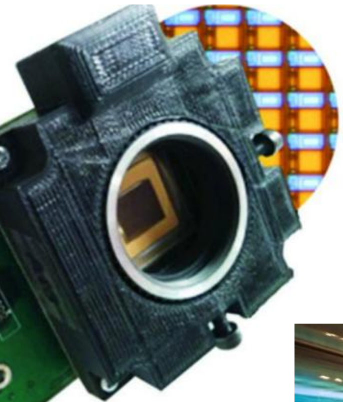
Simulation results
(synaptic conductance maps)



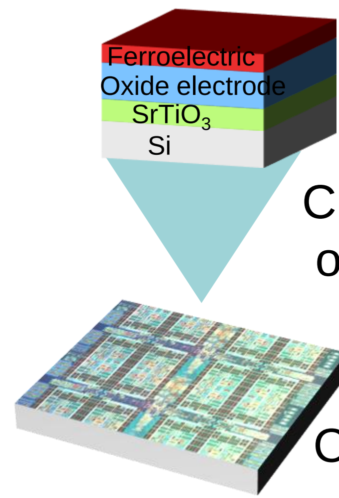
Smart System Integration: hardware AI-enhanced sensor

(ULPEC EU H2020 project, 2017-2021)

ULPEC project



- ULPEC : **U**ltra-**L**ow **P**ower **E**vent based **C**amera
  - Smart System Integration call
  - From sensor to decision taking for driving aid

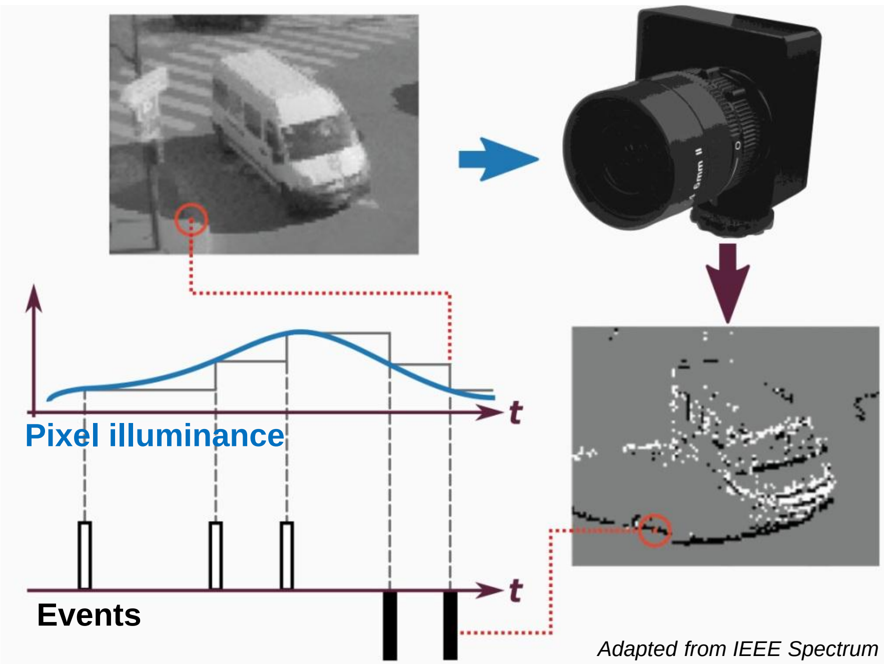


Crossbar array of memristors

CMOS (Sensor + Neurons)

CNRS
BOSCH Invented for life
université de BORDEAUX
SORBONNE UNIVERSITÉ CRÉATEURS DE FUTURS DEPUIS 1257
UNIVERSITY OF TWENTE.
ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich
THALES
UNIVERSITÉ PARIS SUD
PROPHESEE METAVISION FOR MACHINES
IBM Research | Zurich
TSST

Event-based camera: principle & main features



- Event temporal accuracy: sub-millisecond
- Pixel individual event rates: 0 to tens of kHz
- High dynamic range response: >120dB
- Intrinsic data compression

https://www.prophesee.ai/event-based-vision-applications

Example of high-speed detection and tracking for automotive

Detection of changes:
outputs sparse data => low power consumption.

Learning can be difficult in real world…

If one performs simulations with:

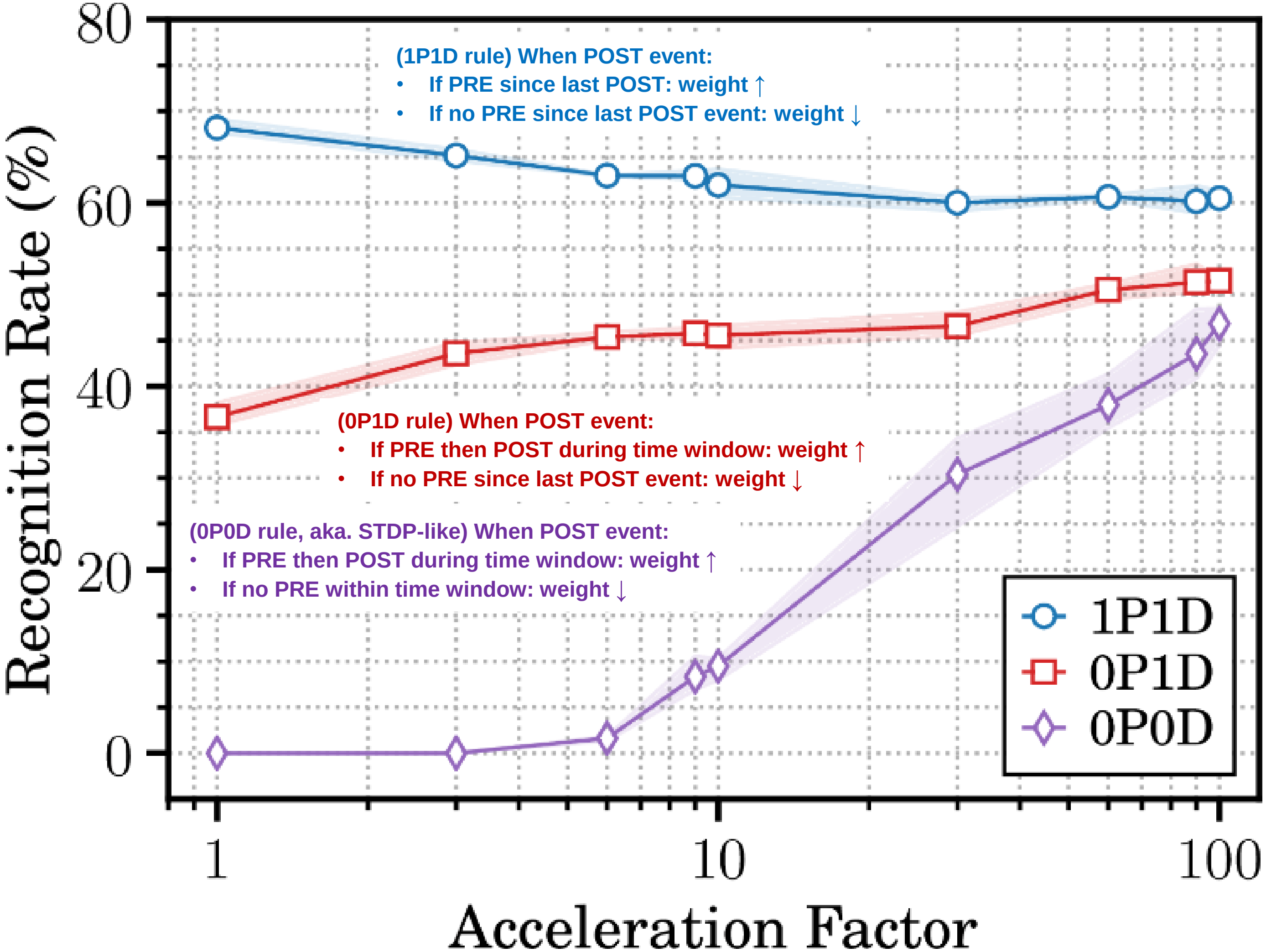
- MNIST inputs filmed with an event-based camera => N-MNIST
  - Camera movement in front of digits (100 ms-saccades)
- STDP-like rule:
  - Time window for STDP: 10 µs (due to analog design constraints)
  - Presynaptic event and then a postsynaptic one in the time window: synaptic weight ↑
  - Postsynaptic event without a presynaptic one in the time window: synaptic weight ↓

0 % recognition rate

Not enough events in the STDP time window for learning.

One may need to use other learning rules

Recognition rate vs. artificial acceleration of events



P. Lewden & al., IJCNN, 2020

Radio-frequency processing with spintronic nanodevices

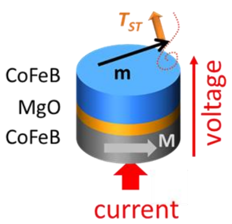
(RadioSpin EU H2020 Project, 2021-2025)

Magnetic Tunnel Junction (MTJ)
A multifunctional spintronic nanodevice

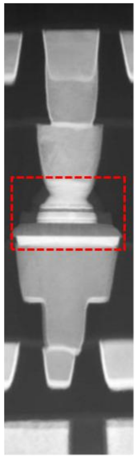
DC: ''direct current,'' i.e., 0 Hz
RF: radio-frequency

N. Leroux & al., Physical Review Applied 15, 034067, 2021

Magnetic tunnel junction



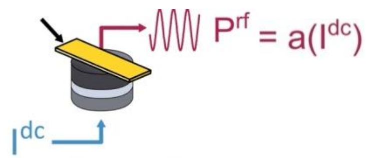
Samsung IEDM 2018, 28nm Logic

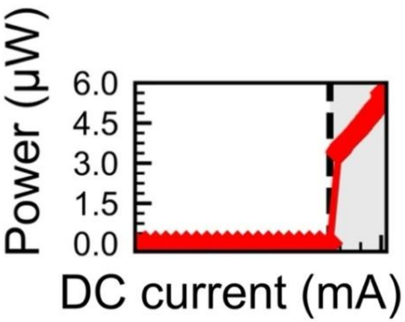


Non-volatility
Endurance
Reproducibility

J. Grollier & al., Nature Electronics 3, 360, 2021

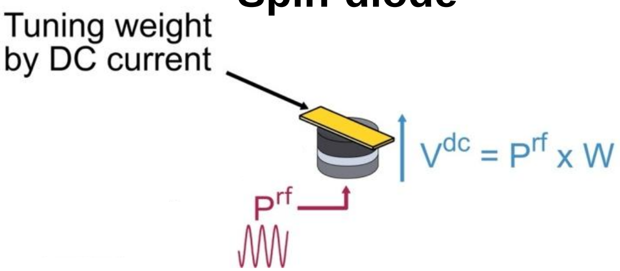
Spin-torque nano-oscillator

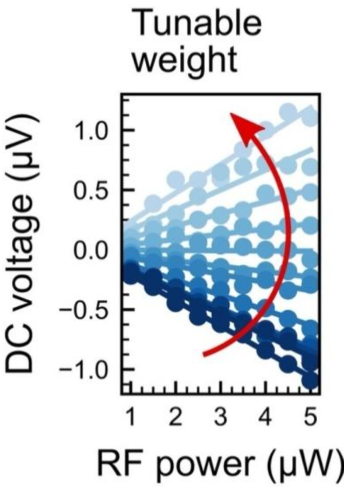




Neuron
(DC → RF)

Spin-diode





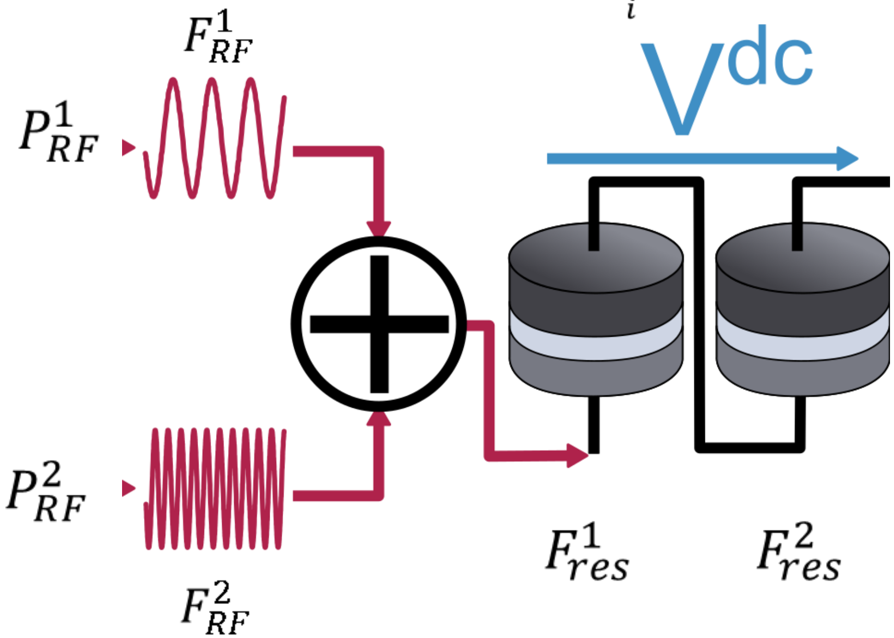
Synapse
(RF → DC)

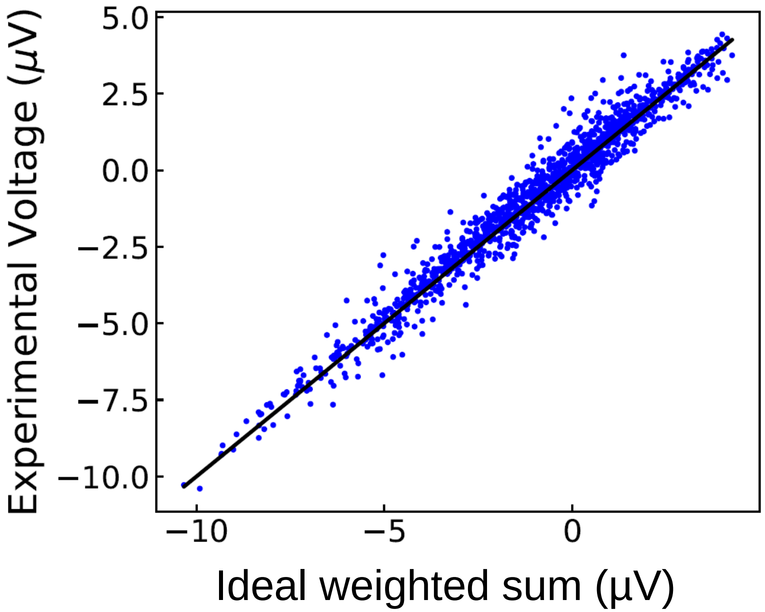
The same technology can be used both for neurons and for synapses.

Combining several synaptic MTJs

N. Leroux & al., Neuromorph. Comput. Eng. **1** 011001, 2021

$$V_{dc} = \sum_i P_{RF}^i \times W(F_{res}^{1i})$$

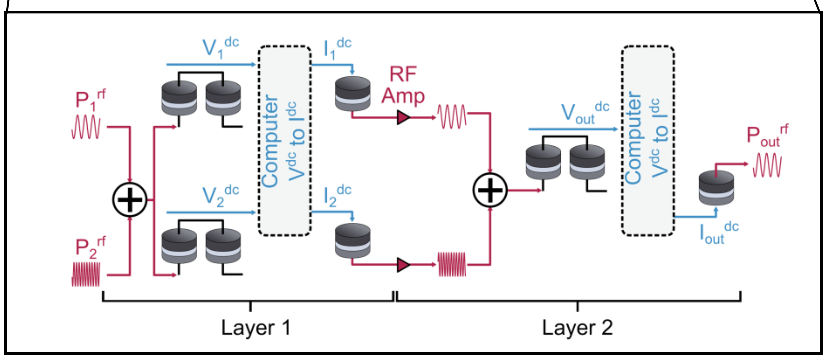




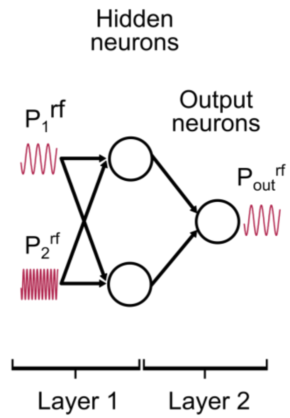
Multiply-And-Accumulate operations on RF signals are possible in hardware!

Fully spintronic hardware neural network

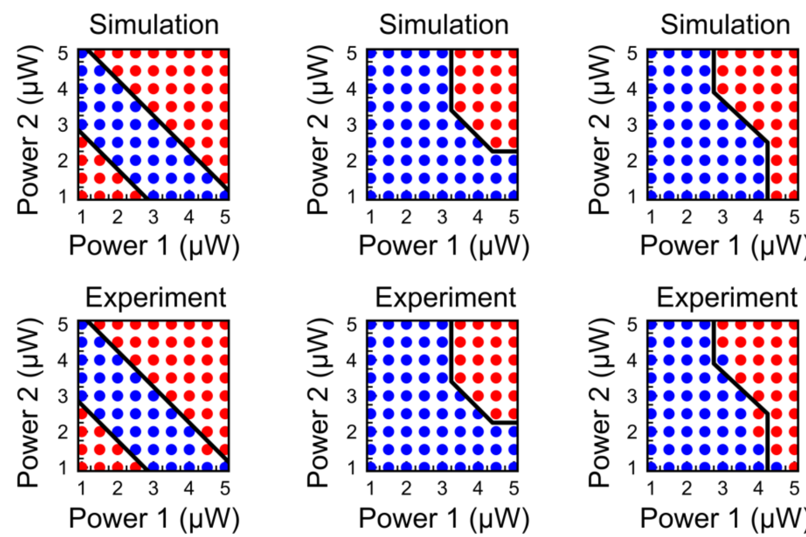
A. Ross & al., ArXiv 2211.03659, under review, 2022



=

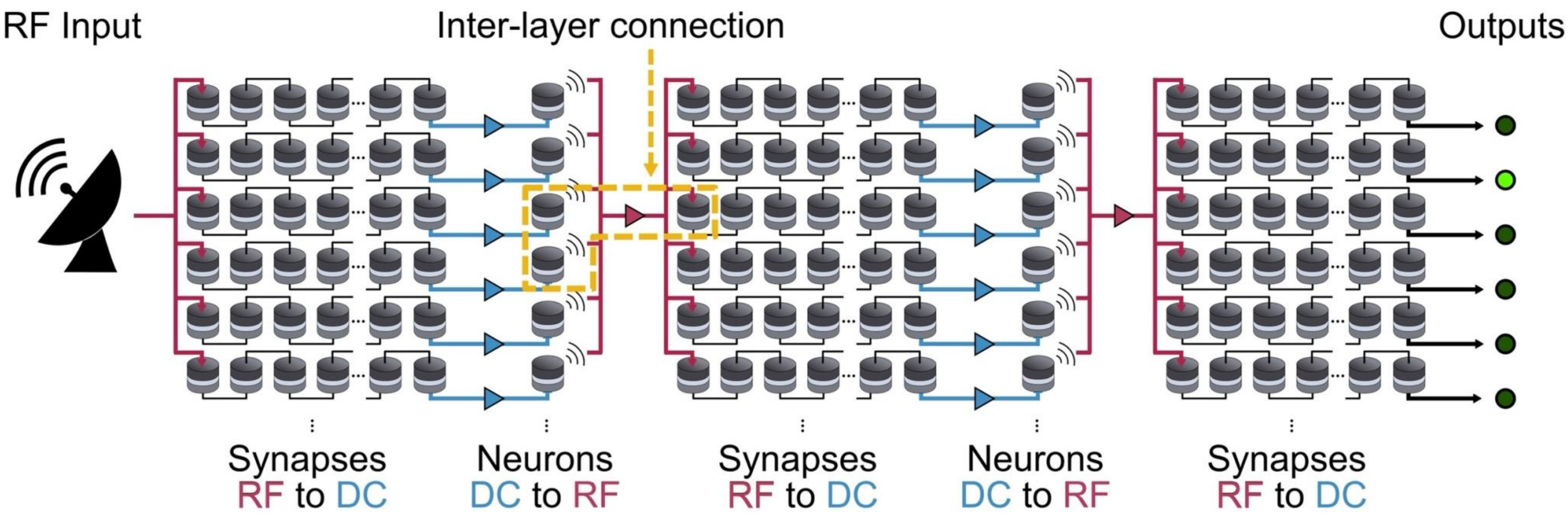


97.7 % accuracy



Towards deep RF spintronics neural networks

A. Ross & al., ArXiv 2211.03659, under review, 2022



Energy consumption estimate:
~ 10 fJ/synapse and ~ 100 fJ/neuron for MTJs with 20 nm diameter:
100 fold energy gain compared to GPUs.

Conclusion

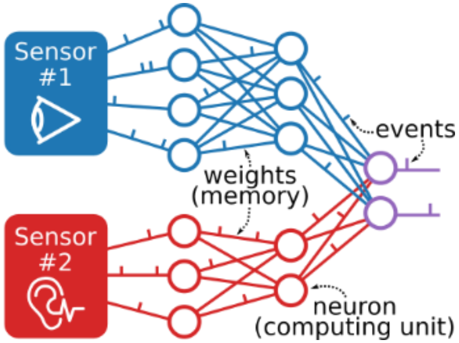
Conclusions and outlook

- Need for energy sustainability of AI before great disillusion.
- Event-based computation (neuromorphic) is a good candidate: mainly for edge-computing.
- CMOS will be not enough: unconventional nanolectronics is likely to play a major role in Green AI.
- It will likely be a long way before industrial products… however that is the purpose of research!
- In any case: best energy savings are not generating irrelevant data!

Currently in Bordeaux: Green AI project

- GrAI: chair on energy-efficient hardware AI
  supported by the French National Research Agency

- Objective: designing better sensors and computing architectures for low-power spiking neural networks.
  - Building new event-based sensors
  - Optimize spiking neural network architectures
  - Study sensor fusion
  - Digital or analog fashion

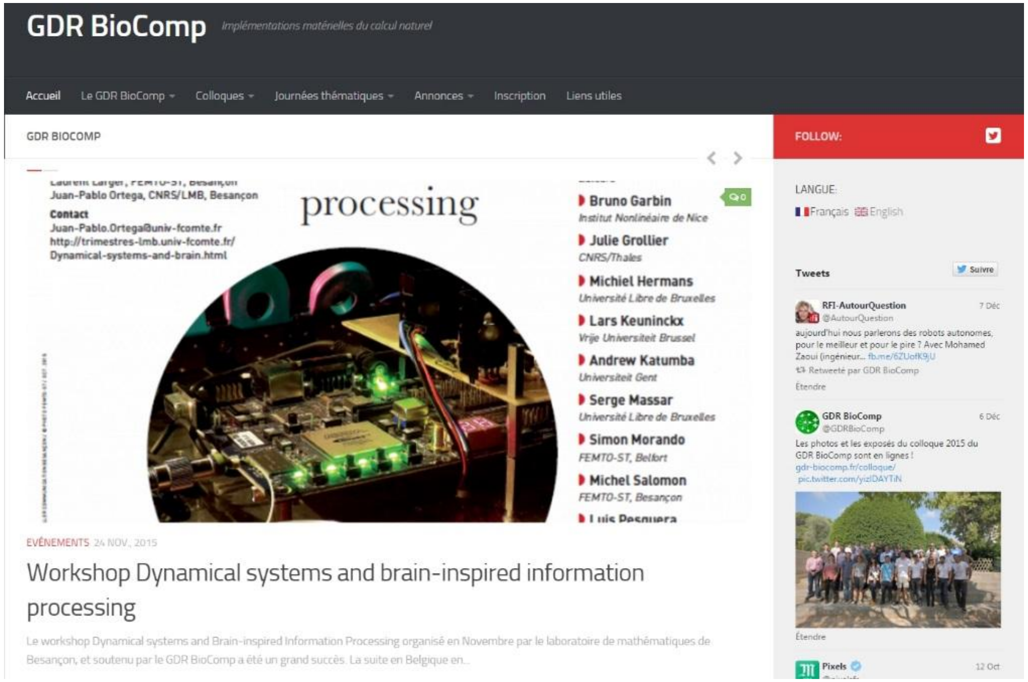


- Led by Sylvain Saïghi (sylvain.saighi@ims-bordeaux.fr)

To go further… at the French level

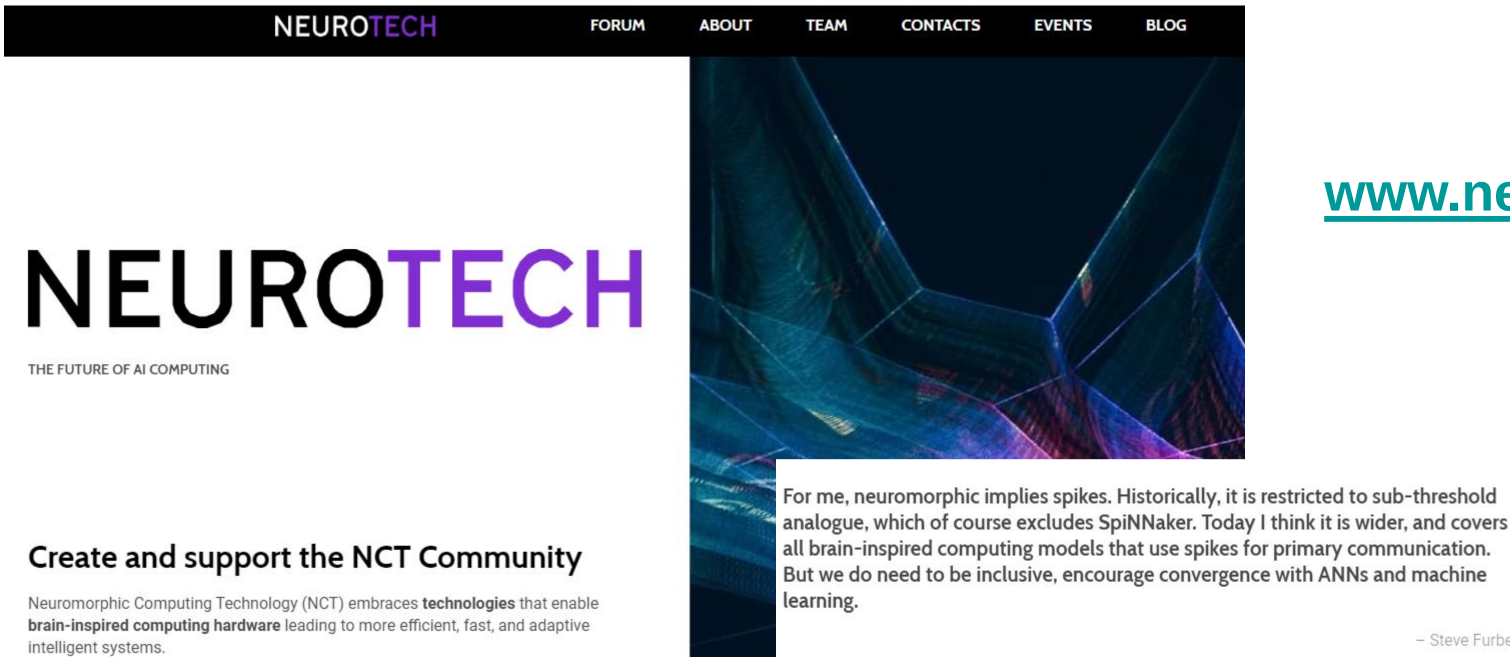
- GDR BioComp
  - Hardware implementation of natural computation
  - 5 CNRS institutes

www.gdr-biocomp.fr



To go further… at the European level

- NeuroTech EU project
  - Aims to gather all academic and industrial NCT stakeholders
  - Started on Fall 2019 for 3 years (just finished!)



www.neurotechai.eu

Thank you for your attention!

adrien.vincent@ims-bordeaux.fr

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant RadioSpin No 101017098.
This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 732642 (ULPEC project).
This work was supported by a grant overseen by the French National Research Agency (ANR) as part of the ''Chaires IA'' Program (GrAI Project, ANR-19-CHIA-0003).
Financial support from the French Agence Nationale de la Recherche (ANR) through MIRA project is acknowledged.