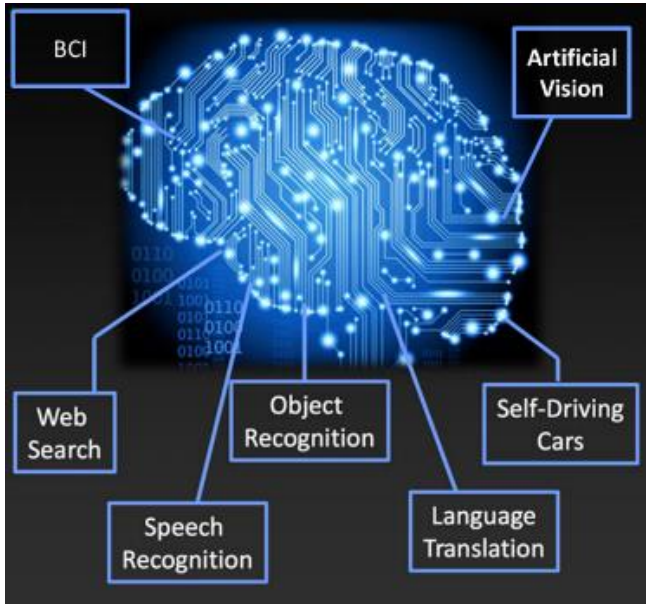
ARCHITECTURE À SPIKE ET MÉMOIRES OXRAM

Outils, technologies et composants pour l'IA

Frédéric Heitzmann & Alexandre VALENTIAN – CEA-LETI | 22nd May 2019

NEURAL NETWORKS: A HUGE AMOUNT OF APPLICATIONS RECENTLY EMERGED

- **Image Recognition**
  - Web (Google, Facebook, …)
  - Autonomous Vehicles (Google, Uber, …)
  - Smartphones (Qualcomm)
  - Medical application
- **Robotics, drones**
  - Movidius, Aldebaran…
- **Temporal Sequences Recognition**
  - Voice (Google voice + G. assistant, Apple Siri, Microsoft Cortana, Amazon Alexa, Samsung Viv)
- **Security/Monitoring**
  - Industrial Process (GST, General Vision)
  - Video Camera Networks
- **Data mining**
  - Smart City (IBM Watson, Schneider Electric)
- **Healthcare and Medicine**
  - Deep Mind, Nvidia Horus …

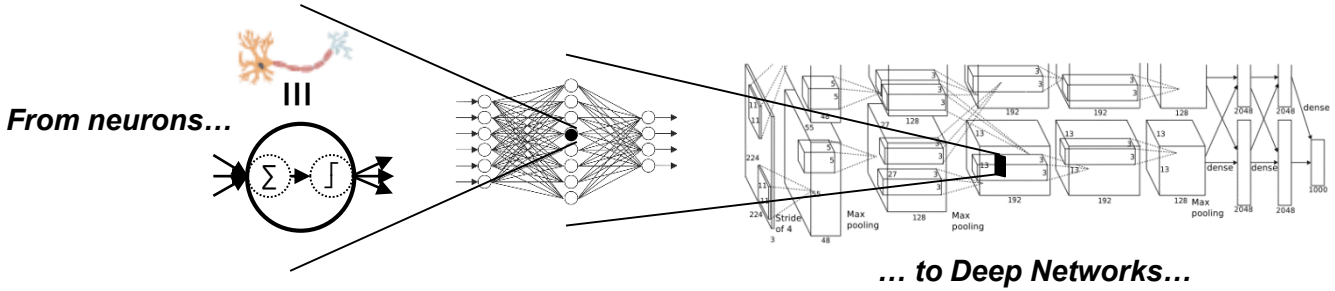


→ **The next general purpose computing?**

Confidential

NEURAL NETWORKS: PROMISE OF A BREAKTHROUGH

- **From neurons to Deep Neural Networks (NN) and Deep Learning**
  - Scaled-up NN contains millions of neurons and billions of synapses
  - Trained with huge datasets (up to millions of images) with gradient descent technics
  - Convolutional NN (CNN) use convolution filters for image recognition
  - Recurrent NN (RNN) are effective for sequences recognition (speech)



- **Current implementations need**
  - Very large computational power for the training phase
  - Large computing power for inference phase

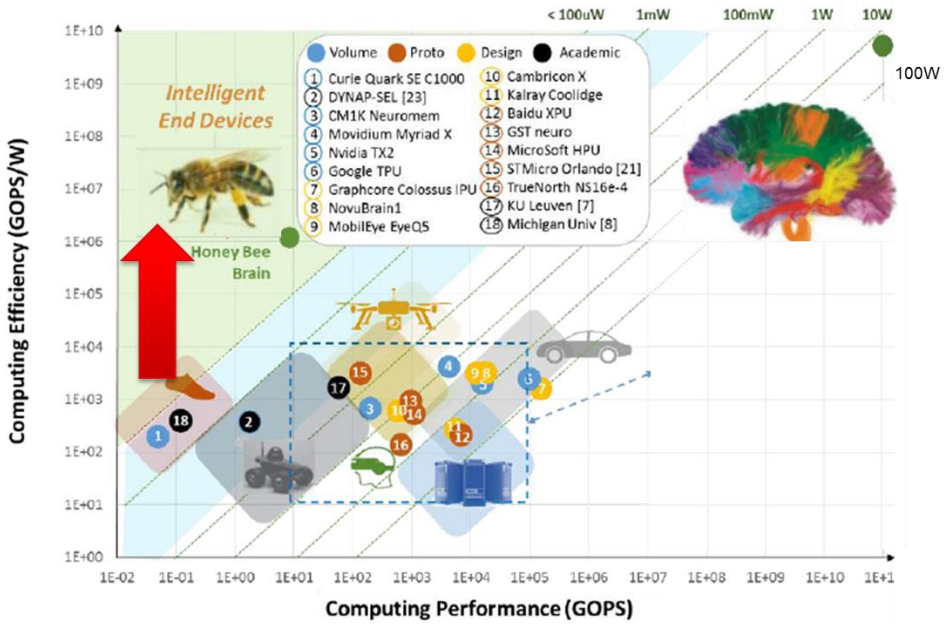
→ **Very high energy consumption due to data movement**

→ **Architecture not adapted to distributed, low-power data processing**

BRAIN VS. COMPUTER: X 10^6 POWER DISCREPANCY

- **Biological system computations are**
  - 3 to 6 order more energy efficient than current dedicated silicon system

- **Brain-inspired computing might just be the key!**



- **Human brain is**
  - Massively parallel
    - 86B neurons and 10^4 more synapses
  - Doing processing using memory elements
  - Event-driven, spike based induced activity
    - No system-clock
  - Self-learning, self-organizing

- **Embedded brain-inspired solutions needs**
  - High density storage, close to neurons
  - Computational storage
  - A time-code will be a must
  - Scalability, re-configurability
    - Online learning to come

Confidential

OUTLINE

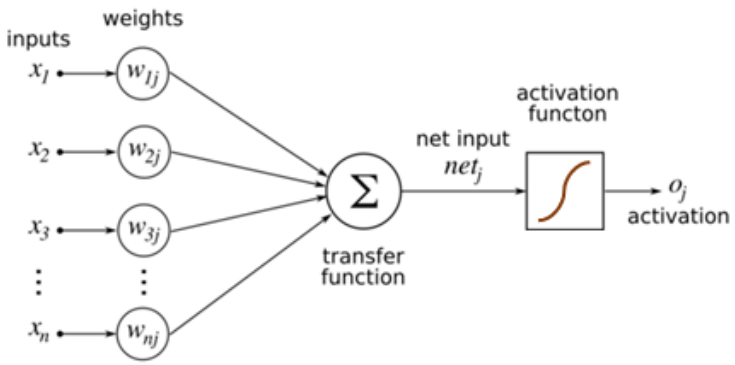
- **Opportunity**
  - Combine Spike-coding and RRAM technology
- **LETI RRAM technology**
- **Circuit**
  - Learning strategy
  - Architecture
- **Conclusion**

INDUSTRIAL NEURAL NETWORKS

- **Abstraction from biology**
  - The spike train is converted into a value representing its mean frequency

- **Neuron**
  - MAC operation
    - Multiplication-Accumulation
  - Non-linear activation function
    - Sigmoid, ReLU …

- **Synapse**
  - Weight stored into DRAM



THOUGHTS

- **We should not compete with those industrial solutions**
  - Ex.: Nvidia V100
    - 21.1B transistors, 12nm FinFET, 815mm² die
    - 16GB HBM2 memory, 900 GB/s, 2.5D integration
    - 300W

- **This is state of the art <u>BUT</u>**
  - It consumes a lot!
  - Computation and memory are not intertwined

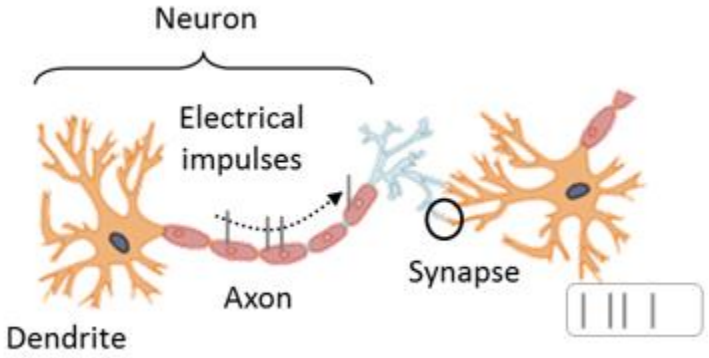
- **Brain works a lot differently**
  - Computation is analog
    - Neuron soma = synaptic current integrator
  - Communication is digital
    - Spikes = unary events, very robust to noise
  - Compute and memory cells are co-located

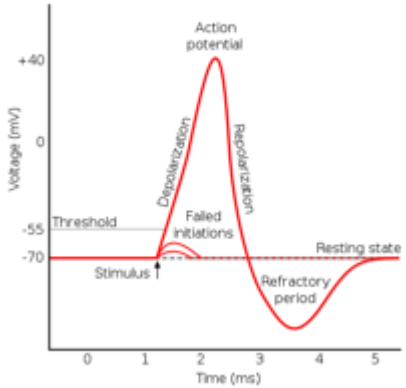
HOW BIOLOGICAL SYSTEMS CAN INSPIRE US MORE?

- **Network**
  - Set of neurons
  - Interconnected through synapses
  - **3D connected**

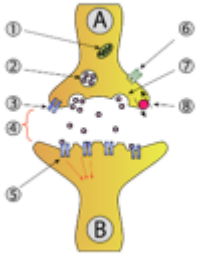
- **Neuron**
  - **Compute** element
    → Integration of inputs
  - 1k – 10k inputs
  - 1 output only but with **very high Fan-out**

- **Synapse**
  - **Memory** element
    → Modulation of inputs
  - Define the function of the network





Action potential = spike



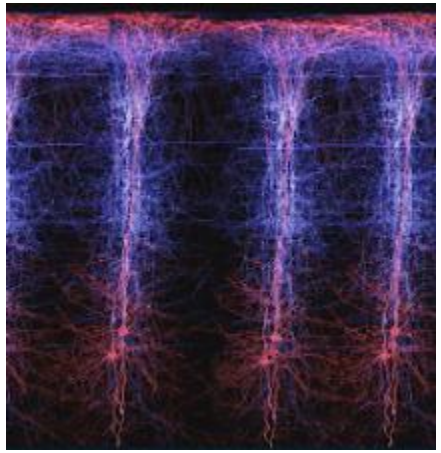
→ **Low frequency (1-10 kHz) usage but huge connectivity**

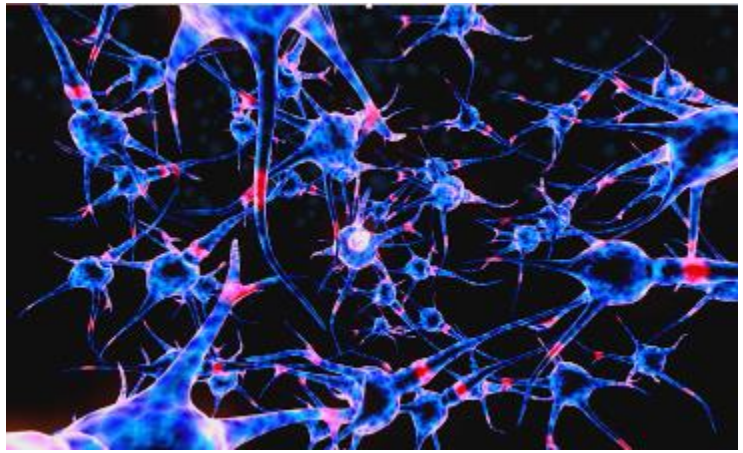
→ **Require NVM elements to enable computation**

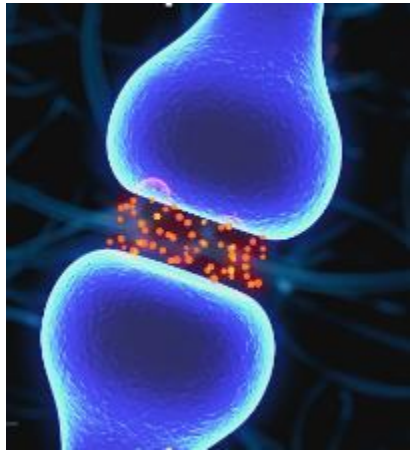
NEURON : A UNIVERSAL NON VOLATILE MEMORY BUILDING BLOCK THAT IS NOT SO SMALL AND ENERGY EFFICIENT

- 1 spike ~ 120 pJ
- 1 neuron ~ 20x20x20 μm^3
- 10^4 memory elements per neuron

Current NVM has better efficiency





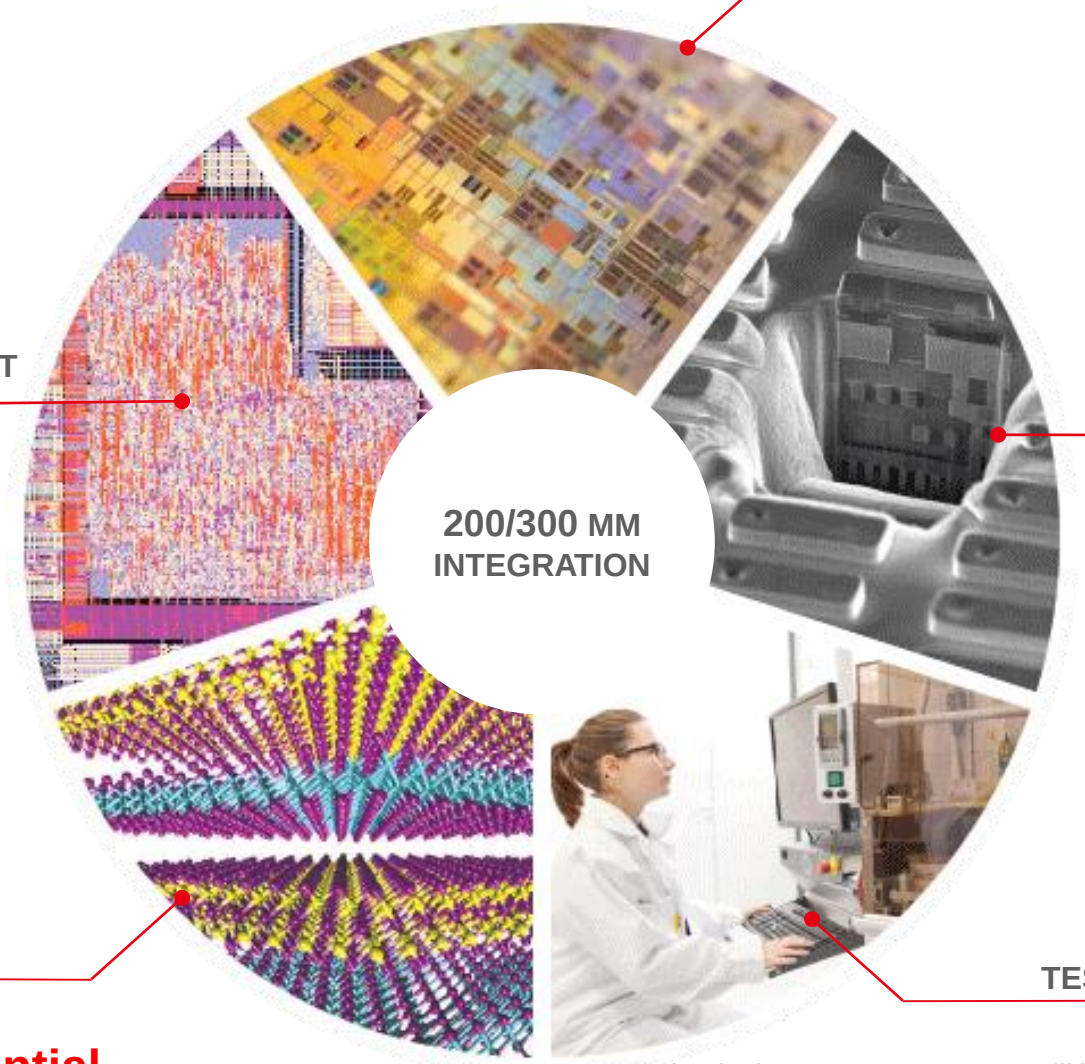


- **Opportunity: Systems are highly scalable and "general purpose"**
  - Mouse brain: 10^7 Neurons, 10^{11} Synapses (= memory elements)
  - Cat brain: 10^9 Neurons, 10^{13} Synapses
  - Human brain: 10^{11} Neurons, 10^{15} Synapses

OUTLINE

- Opportunity
  - Combine Spike-coding and RRAM technology
- **LETI RRAM technology**
- Circuit
  - Learning strategy
  - Architecture
- Conclusion

MEMORY: A UNIQUE VALUE PROPOSITION



Large variety of materials available

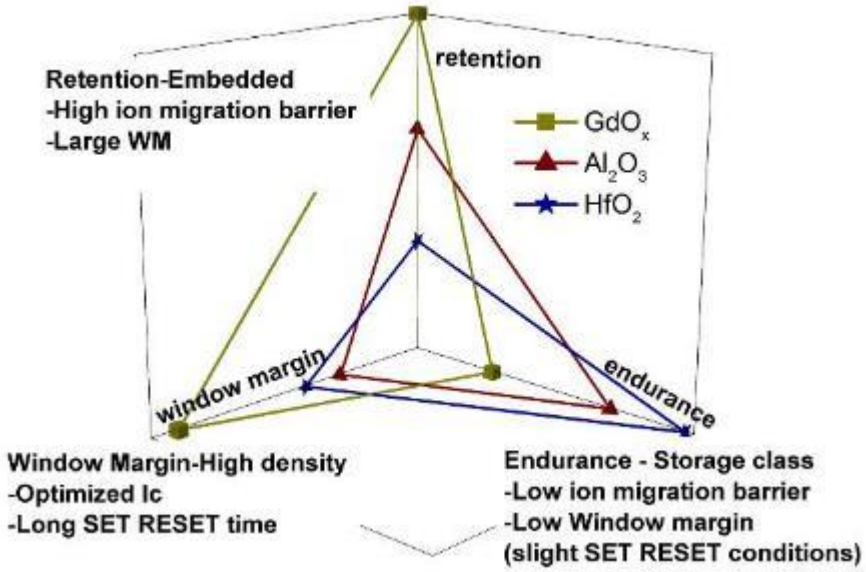
HfAlxOy
SiOx
TaOx
ZrOx
AlOx
VOx

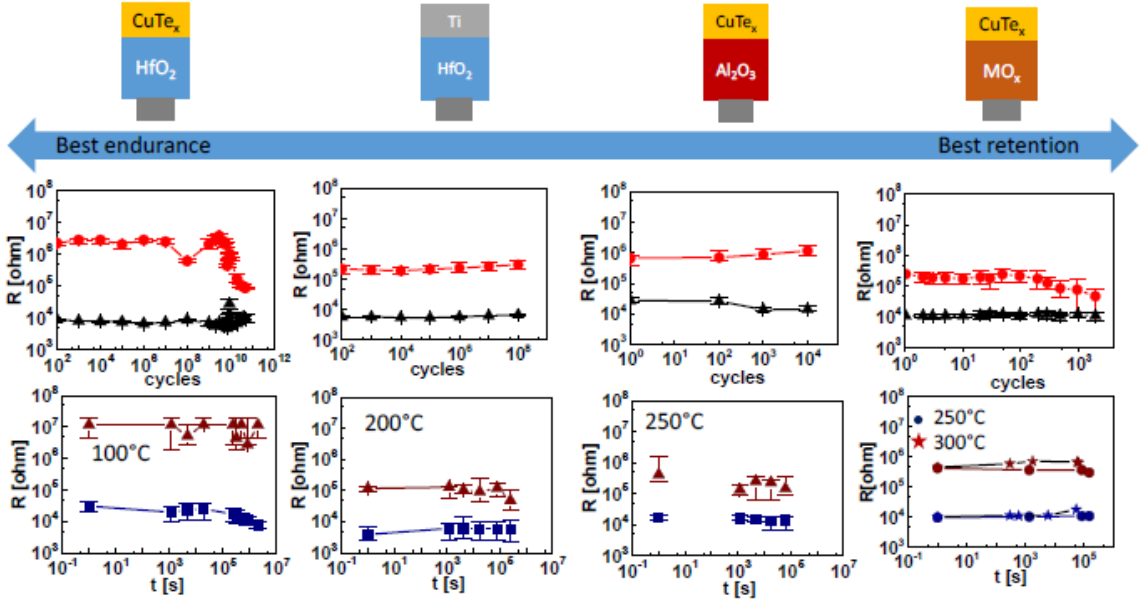
GeSbTe
GeAsSbTe

Large variety of Memories available

Conductive Bridge RAM
Oxide Resistive RAM
Ferro-electric RAM
Phase – Change Memory
pSTT-Magnetic RAM

RRAM BENCHMARK FOR TRADEOFF UNDERSTANDING





VLSI 2018
IMW 2018
EDL 2018
IRPS 2018

Towards circuit implementation

- Via Collaborations
- MAD Shuttle

LETI AND CMP ANNOUNCE WORLD'S FIRST MULTI-PROJECT WAFER SERVICE WITH INTEGRATED SILICON OXRAM

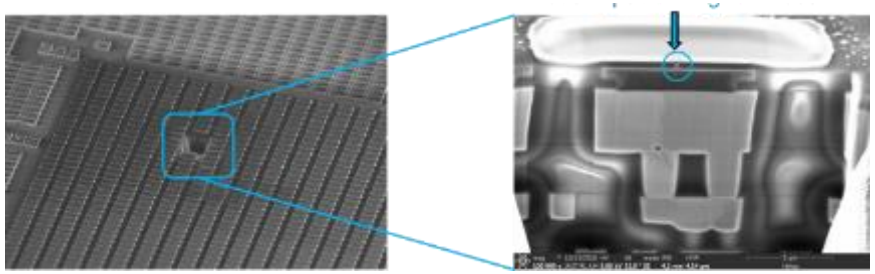
Weebit Nano and Leti extend their agreement to fast track commercialisation

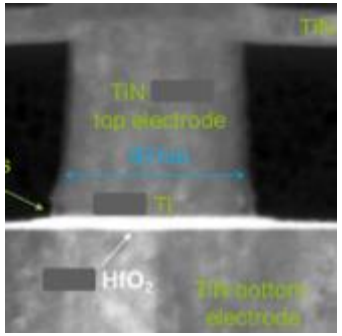
16 May 2018 – Weebit Nano (ASX: WBT), the Israel-based semiconductor company seeking to develop and commercialise the next generation of memory technology, today announced an extension of the agreement with its partner Leti, the French research institute recognised as a global leader in the field of micro-electronics, to further develop and optimise Weebit's ReRAM memory technology.

Confidential

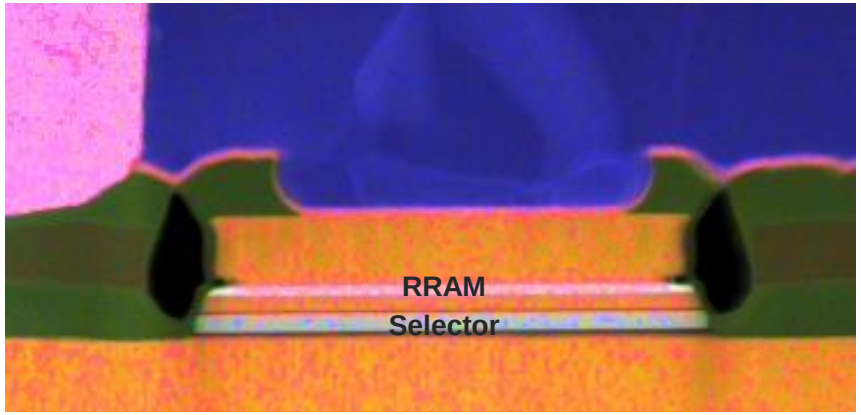
RRAM CONTINUOUS IMPROVEMENT

- **RRAM scaling and cost reduction**
  - Scalable down to 40nm x 40nm





- **Co-integration with Selector**
  - Enables Mbit scale crossbar, instead of kbit

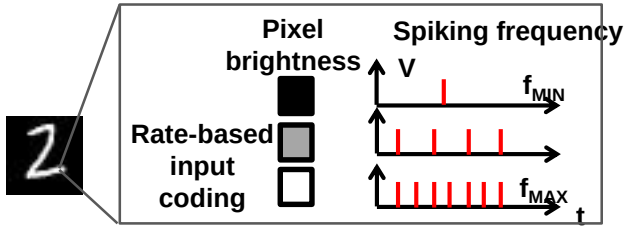


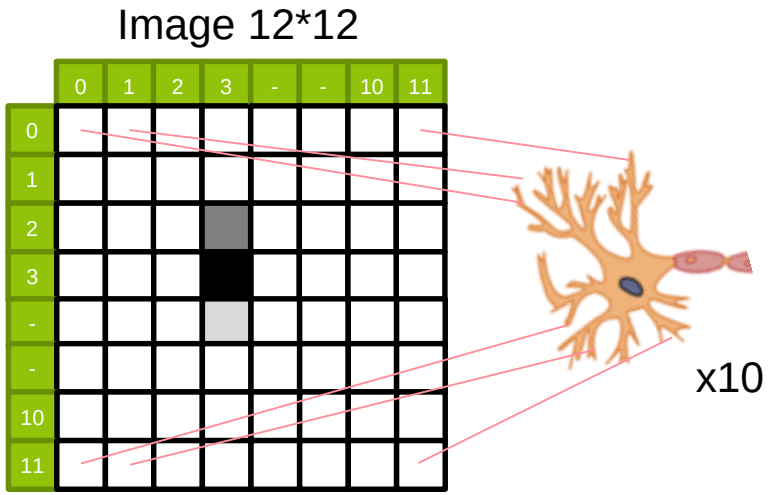
OUTLINE

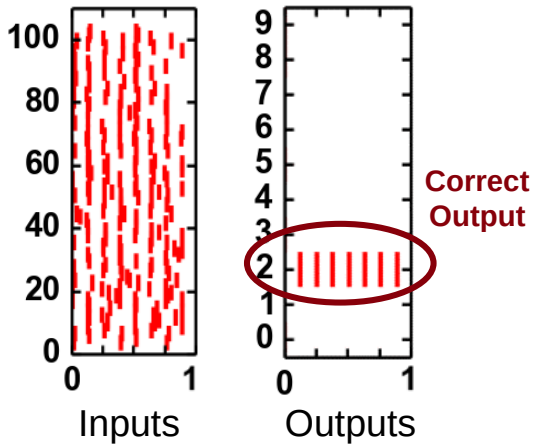
- **Opportunity**
  - Combine Spike-coding and RRAM technology
- **LETI RRAM technology**
- **Circuit**
  - Learning strategy
  - Architecture
- **Conclusion**

APPLICATION

- **Hand-written digit classification**
  - MNIST Database
- **Reduced image size**
  - 12*12 pixels
  - Pixel grey level encoding
- **Fully-connected neural network topology**
  - 10 output neurons : 1 neuron / class
  - Each neuro is connected to the entire image : 144 synapses



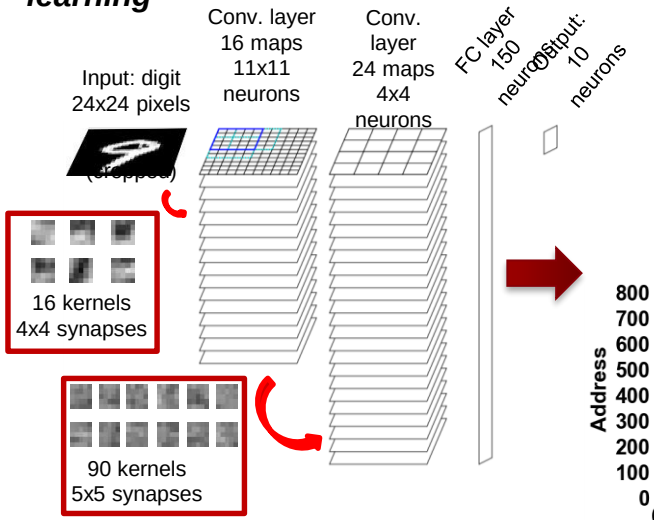




THE PROMISES OF SPIKE-CODING NN

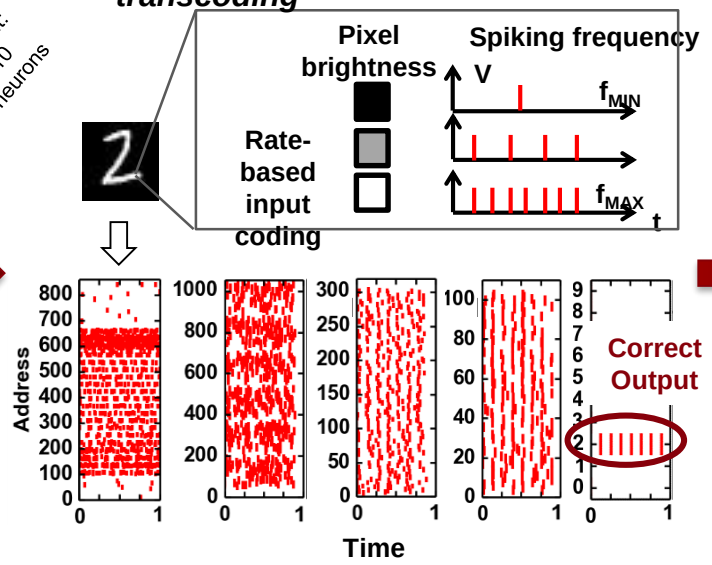
- Reduced computing complexity and natural temporal and spatial parallelism
- Simple and efficient performance tunability capabilities
- Spiking NN best exploit NVMs such as RRAM, for massively parallel synaptic memory

1) Standard CNN topology, offline learning

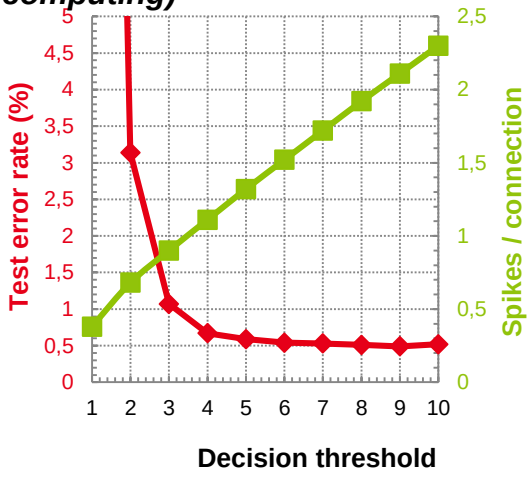


[O. Bichler et al., IEDM, 2015]

2) Lossless spike transcoding

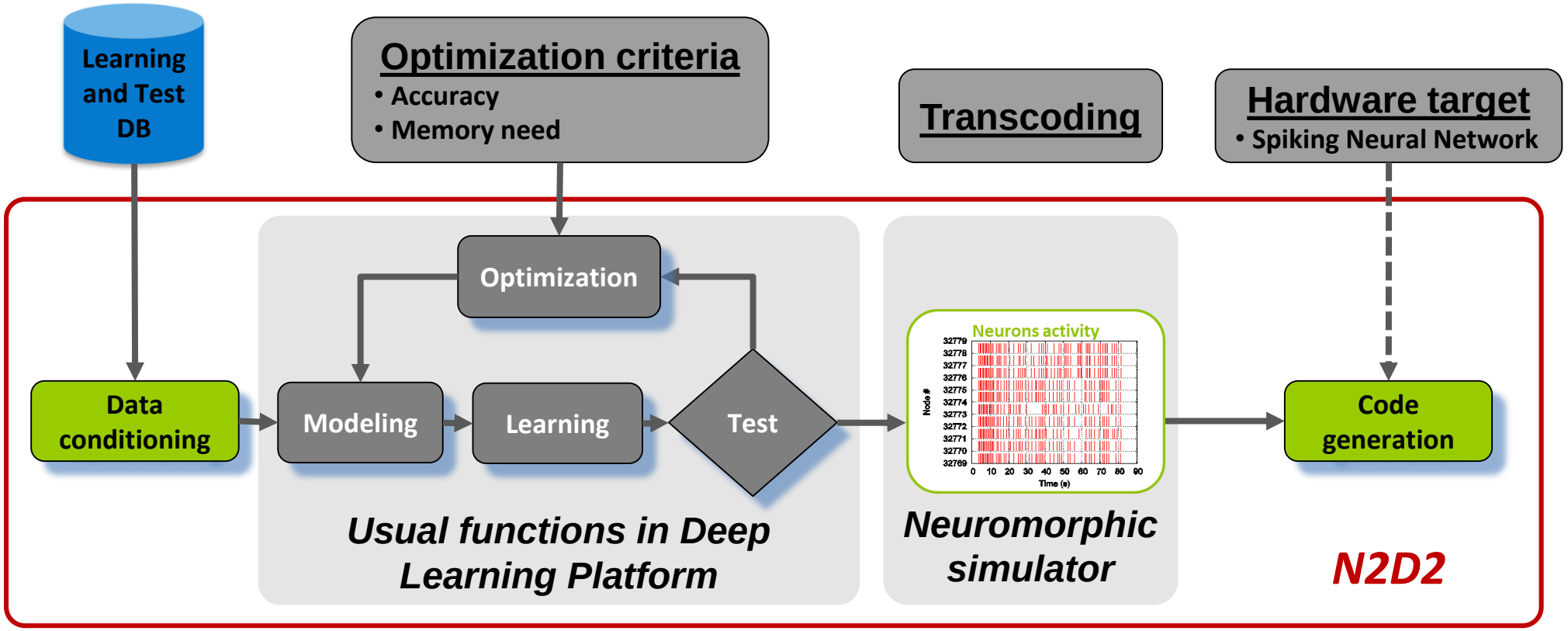


3) Performance vs computing time tunability (approximated computing)



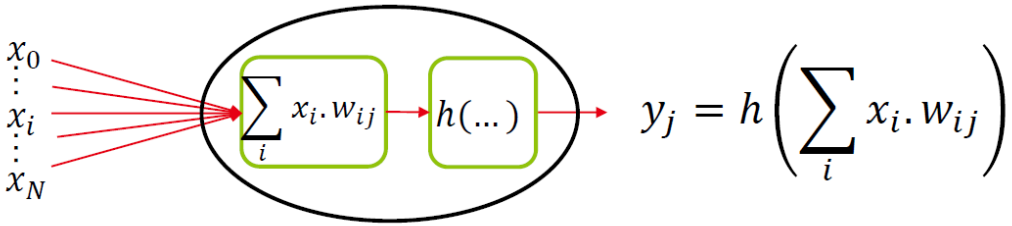
LEARNING STRATEGY

- Offline learning

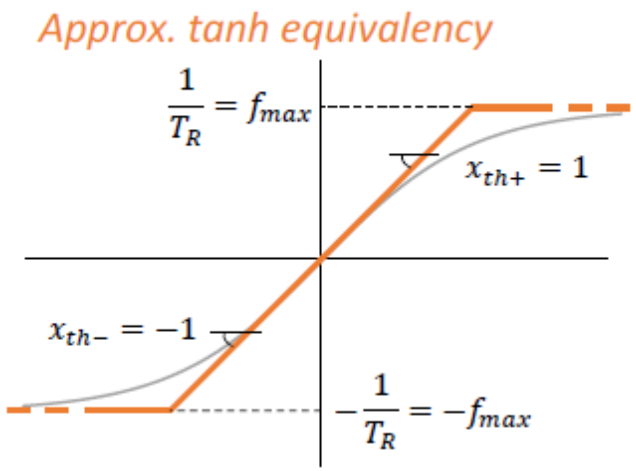


MATHEMATICAL EQUIVALENCE

- **“Formal” neural network model**
  - Multiply-Accumulate (MAC) + non-linear operation

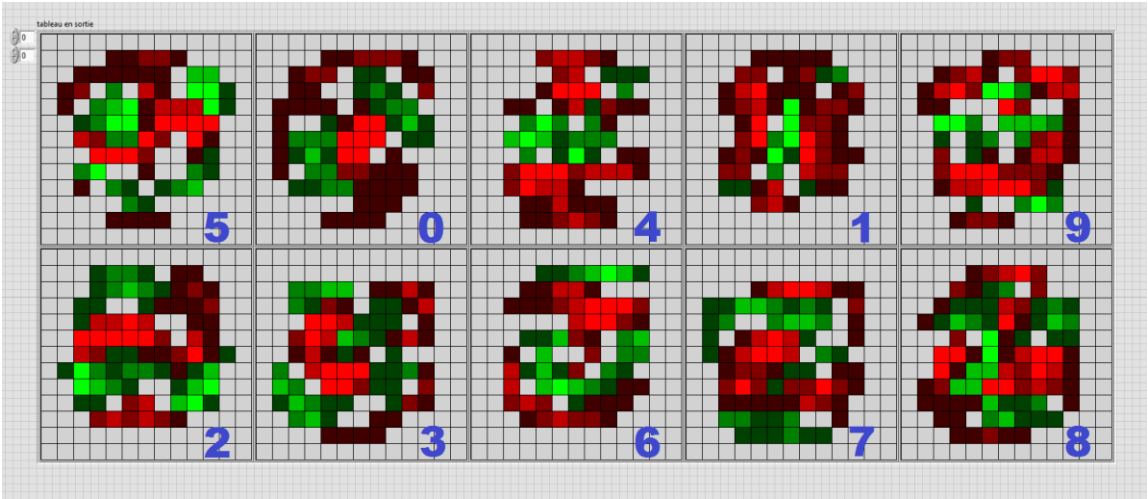


- **“Spiking”, rate-based equivalence**
  - Specially-designed Integrate & Fire (IF) analog neuron
    - For ensuring mathematical equivalent to classical-coding neuron with TANH activation function
- **Features**
  - Neuron with two thresholds, positive and negative
    - Instead of a unique one
  - Thresholds equal to maximum synaptic weight (resp. excitatory and inhibitory)
  - Subtraction or Addition of a maximum synaptic weight, when the neuron spikes
    - Instead of resetting the Soma to 0

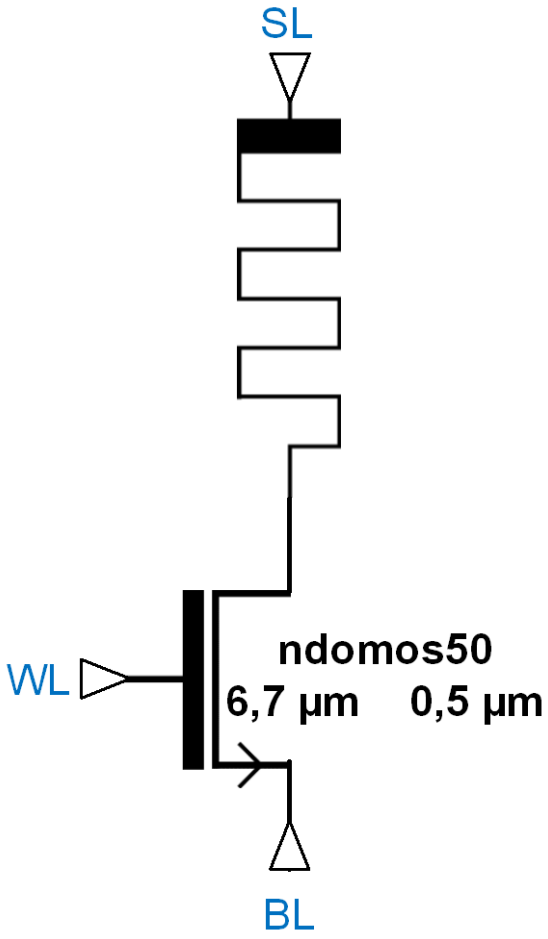


LEARNED NEURONS RECEPTIVE FIELDS

- **Excitatory synapses are represented in green**
  - The greener, the higher
- **Inhibitory synapses are represented in red**
  - The more red, the higher



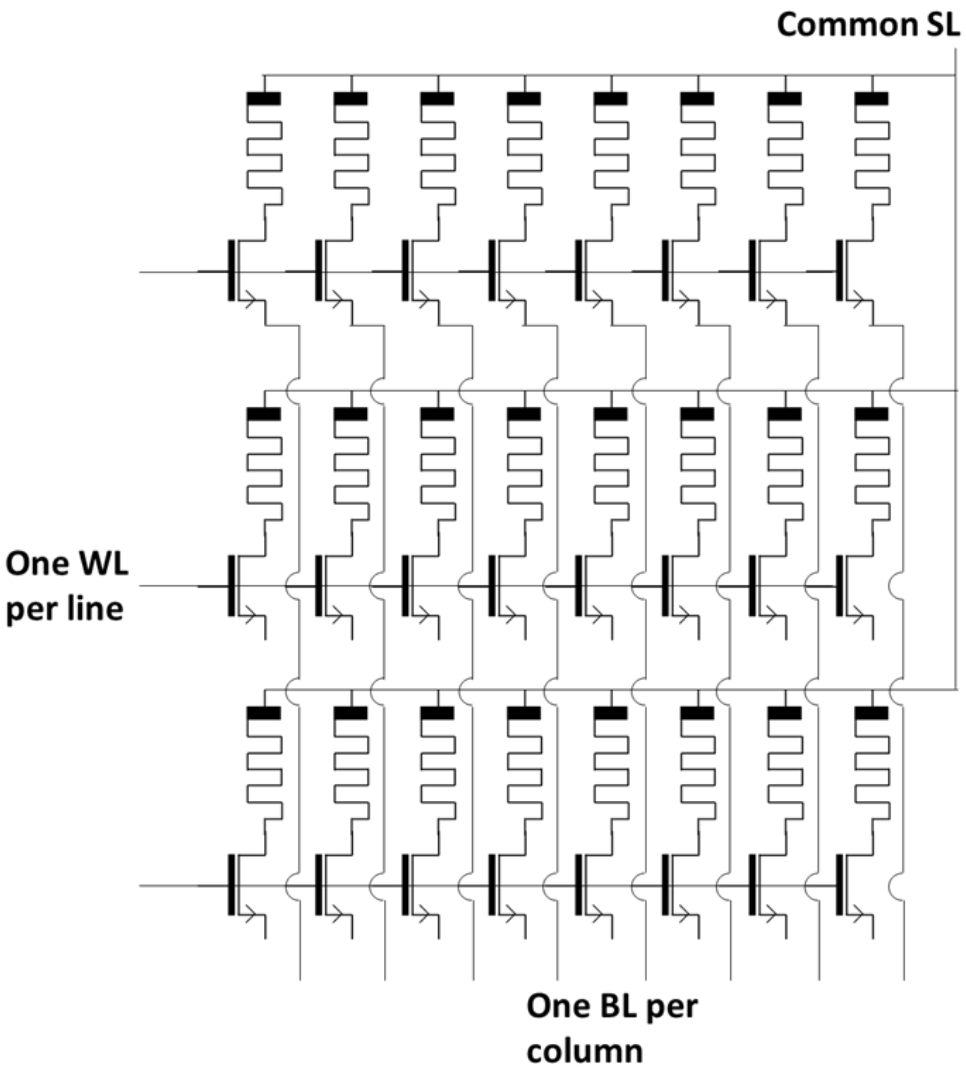
1T-1R: THE BUILDING BLOCK



- **With:**
  - **SL for “Source Line”**
  - **WL for “Word Line”**
  - **BL for “Bit Line”**

- **Access transistor sizing**
  - Corresponds to an optimal occupation of the area available below an OxRAM

OXRAMS MEMORY ARRAY



- **Write operation**
  - Can be done OxRAM per OxRAM
  - By selecting a line and a column
- **Read operation**
  - Is done either line by line, or on a synapse basis

OUTLINE

- **Opportunity**
  - Combine Spike-coding and RRAM technology
- **LETI RRAM technology**
- **CIrcuit**
  - Learning strategy
  - Architecture
- **Conclusion**

CONCLUSION

- **Brain-inspiration can bring huge energy gains**

- **Need for**
  - Very dense computational memories
  - Physically located close to the neurons, eventually on top of them

- **RRAM is a natural fit in that respect**

- **LETI can do Design-Technology Co-Optimization**
  - It designed and fabricated a spiking neural network with RRAM synapses
  - Showed high energy gains

- **Future work**
  - Large scale application circuit in 28nm FDSOI + RRAM
    - LIDAR inference task …

THANK YOU

Leti, technology research institute
Commissariat à l'énergie atomique et aux énergies alternatives
Minatec Campus | 17 avenue des Martyrs | 38054 Grenoble Cedex | France
www.leti-cea.com