NEUROMORPHIC ARCHITECTURES & OPPORTUNITIES FOR NVM TECHNOLOGIES

Etienne Nowak, Head of Non Volatile Memory Laboratory - CEA-Leti

Flash Memory Summit 2018

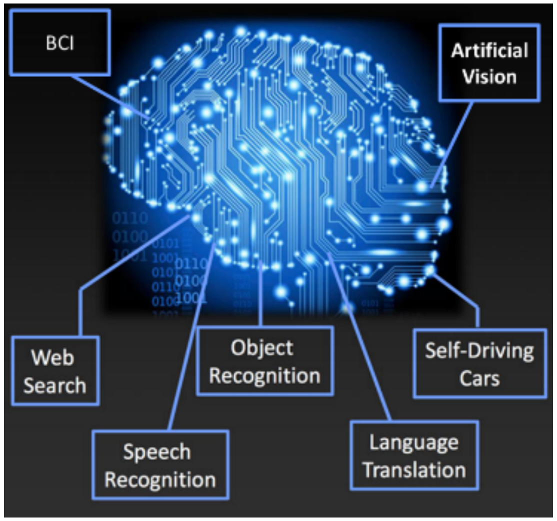
BOOTH #852

Aug. 9th, 2018

Neural Networks :
A huge amount of Applications recently emerged

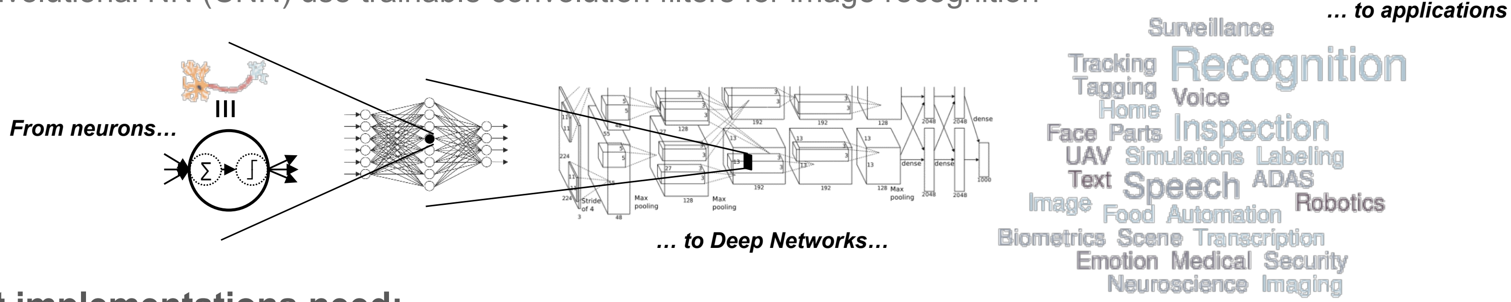
- **Image Recognition**
  - Web (Google, Facebook, …)
  - Autonomous Vehicles (Google, Uber, …)
  - SmartPhones (Qualcomm)
  - Medical application
- **Robotics, drones**
  - Movidius, Aldebaran…
- **Temporal Sequences Recognition**
  - Voice (Google voice + G. assistant, Apple Siri, Microsoft Cortana,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

→**The next general purpose computing ?**



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
  - Recurrent NN (RNN) are effective for sequences recognition (speech)
  - Convolutional NN (CNN) use trainable convolution filters for image recognition

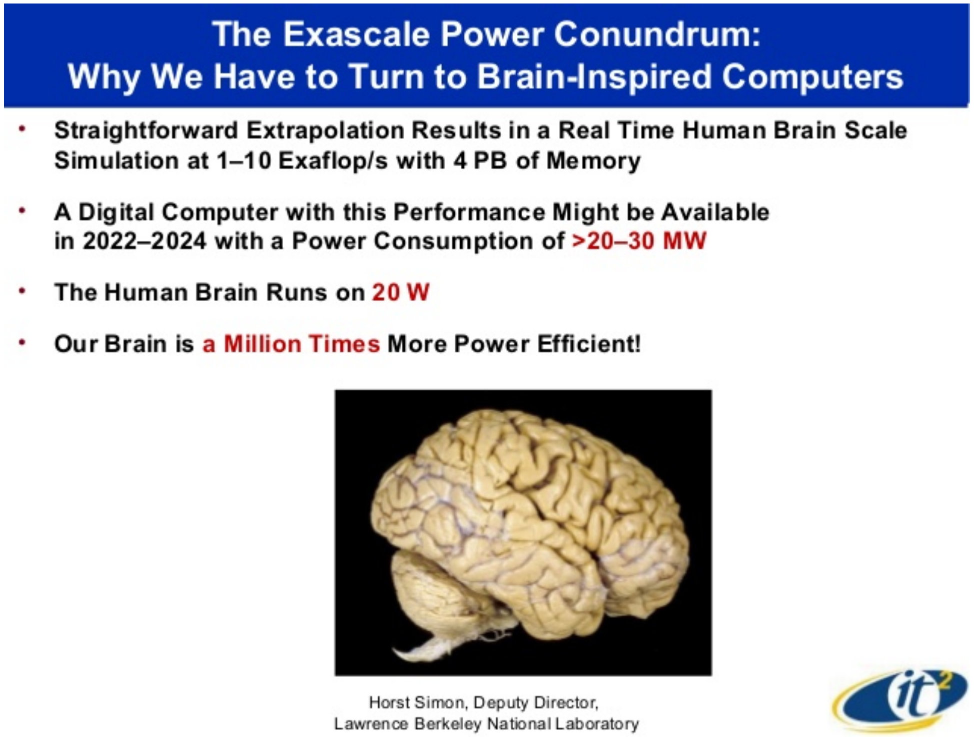


- **Current implementations need:**
  - Large computational power to define network
  - Large labelled data sets for training
  - Access to the large computing system at moment of use

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

→ **Architecture not adapted to distributed or low power embedded data processing**

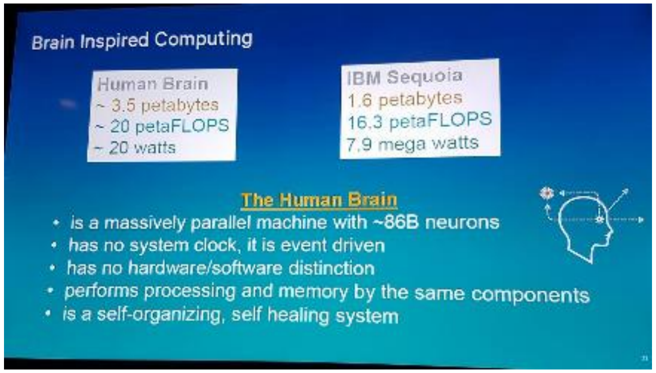
Brain VS. Computer : x 10^6 power discrepancy



The Exascale Power Conundrum: Why We Have to Turn to Brain-Inspired Computers

- Straightforward Extrapolation Results in a Real Time Human Brain Scale Simulation at 1–10 Exaflop/s with 4 PB of Memory
- A Digital Computer with this Performance Might be Available in 2022–2024 with a Power Consumption of >20–30 MW
- The Human Brain Runs on 20 W
- Our Brain is a Million Times More Power Efficient!

Horst Simon, Deputy Director,
Lawrence Berkeley National Laboratory

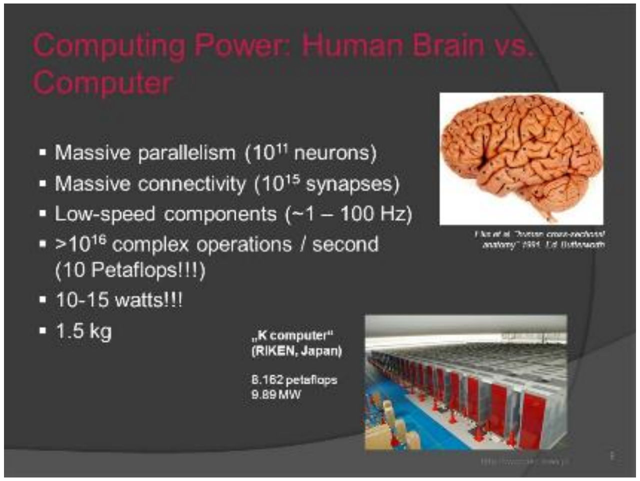


Brain Inspired Computing

Human Brain	IBM Sequoia
~ 3.5 petabytes	1.6 petabytes
~ 20 petaFLOPS	16.3 petaFLOPS
~ 20 watts	7.9 mega watts

The Human Brain

- is a massively parallel machine with ~86B neurons
- has no system clock, it is event driven
- has no hardware/software distinction
- performs processing and memory by the same components
- is a self-organizing, self healing system



Computing Power: Human Brain vs. Computer

- Massive parallelism (10^{11} neurons)
- Massive connectivity (10^{15} synapses)
- Low-speed components (~1 – 100 Hz)
- >10^{16} complex operations / second (10 Petaflops!!!)
- 10-15 watts!!!
- 1.5 kg

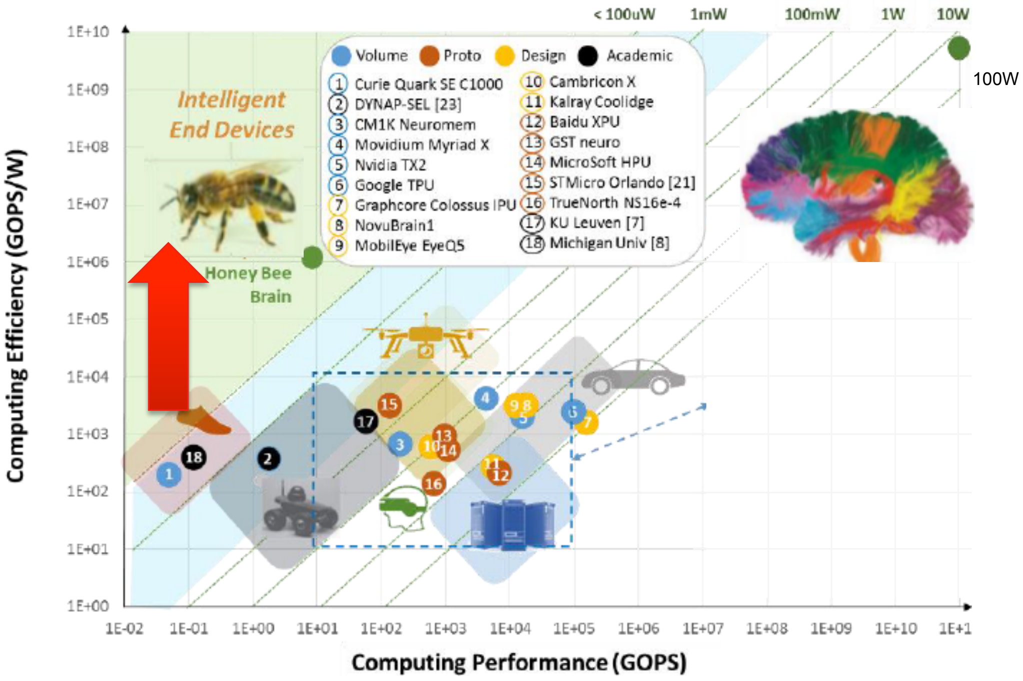
[illegible] et al. "Human cross-sectional anatomy" 1991, Ed. Butterworth

„K computer“ (RIKEN, Japan)

8.162 petaflops
9.89 MW

PROVIDE A LONG ROADMAP FOR COMPUTING EFFICIENCY

- Basic brain elements have the similar performance than today CMOS and NVM architecture
- Biological system computation are 3 to 6 order more energy efficient than current dedicated silicon system

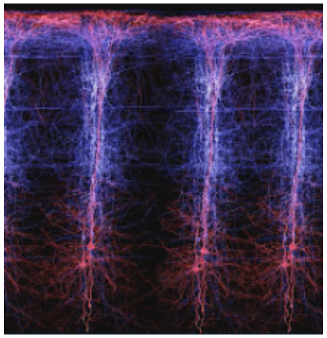


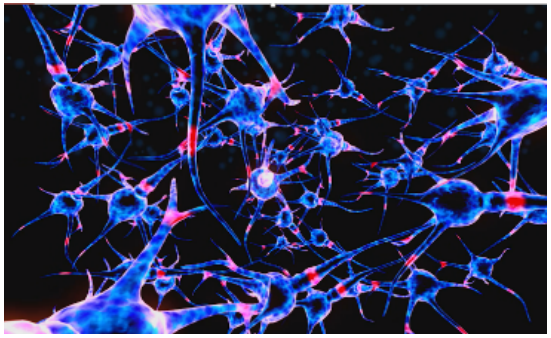
NEURON : A UNIVERSAL NON VOLATILE MEMORY BUILDING BLOCK THAT IS NOT SO SMALL AND ENERGY EFFICIENT

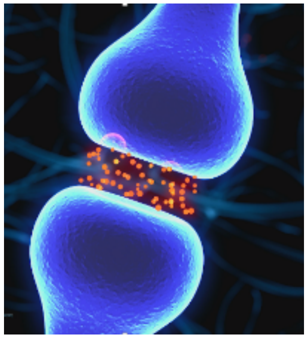
- 1 spike ~ 120pJ
- 1 neuron ~ $20x20x20um^3$
- 10^4 memory elements per neuron

Current NVM has better efficiency

NAND Flash has as smaller size







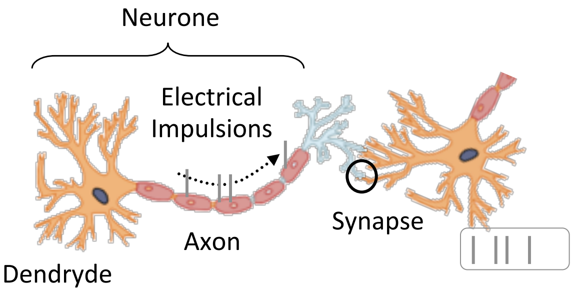
- Opportunity : System are highly scalable and « general purpose »
  - Mouse brain : 10^7 Neurons, 10^{11} Synapses (=memory element)
  - Cat brain :10^9 Neuron , 10^{13} Synapses (= memory element)
  - Human brain : 10^{11} Neuron , 10^{15} Synapses (= memory element)

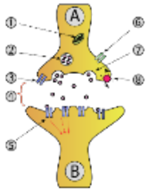
HOW BIOLOGICAL SYSTEMS CAN INSPIRE US MORE?

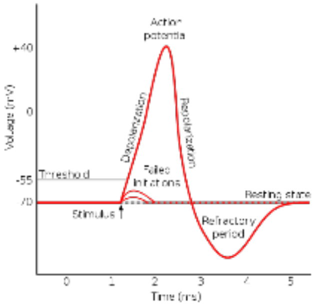
- **Network**
  - Set of neurones
  - Interconnected through synapses
  - **3D connected**

- **Neurone**
  - **Compute** elements
    → 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



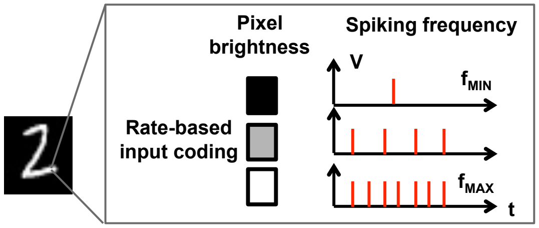




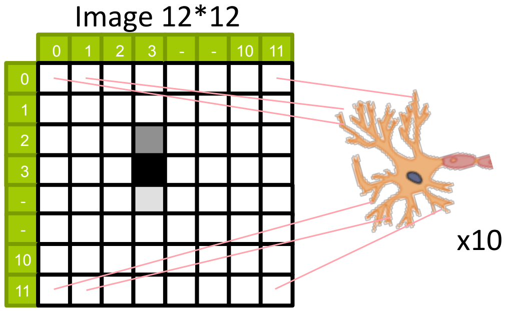
Action potential

→ **Low frequency (1-10 kHz) usage but huge connectivity**
→ **Require NVM elements to enable computation**

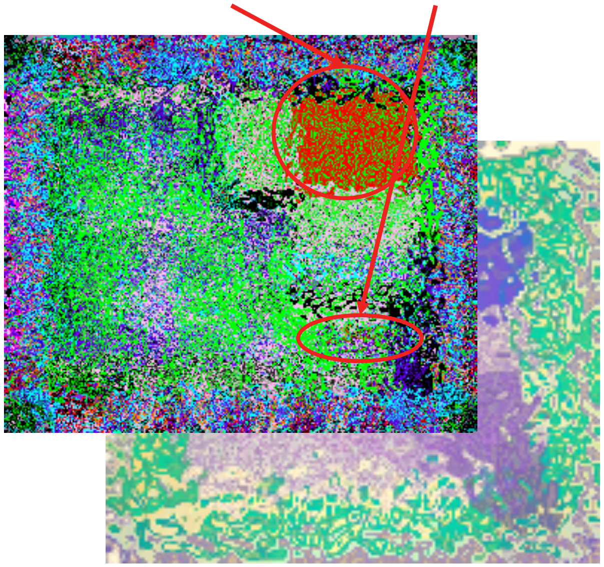
BIOLOGICAL INSPIRED NEURONES USING OXRAM



- **Classification of handwritten numbers**
- **Small resolution image**
  - 12*12 pixels
- **Fully-connected network**
  - 10 neurones : 1 neurone / class
  - 144 synapses



- **130nm CMOS + ReRAM,**
- **Clock frequency: 50 MHz**
- **10 neurones**
- **10*144 synapses = 11,5 kOxRAMs**

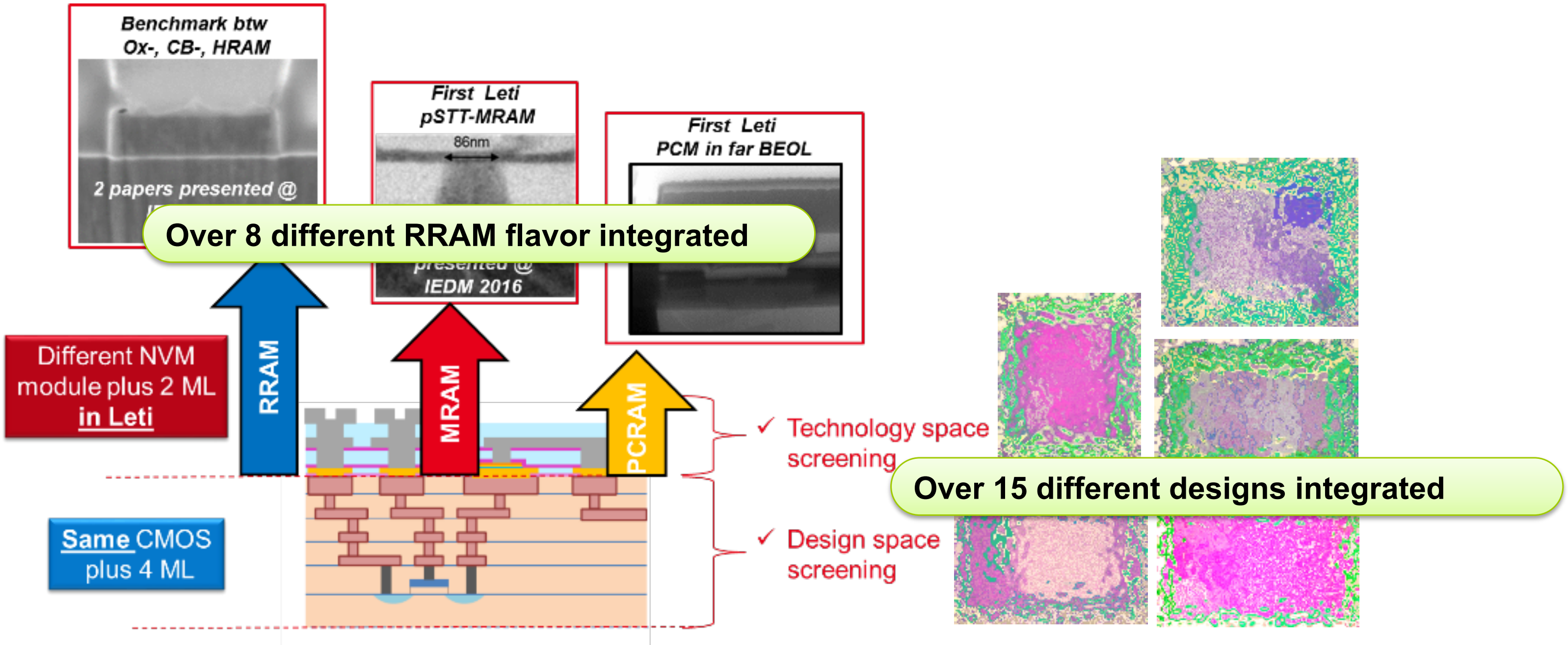


Fabricated circuit /under test

→ Capability to design functional circuit based on ReRAM and Spike-driven

MEMORY ADVANCED DEMONSTRATOR (MAD)
FOR DESIGN AND TECHNOLOGY EXPLORATION

- → Open to all designers in 200mm of HfO2 based ReRAM https://mycmp.fr
- → Contact Leti if needs alternative ReRAM flavor or want to provide yours
- → New in 2019 : **300mm integration** for access to more efficient CMOS



Over 8 different RRAM flavor integrated

Over 15 different designs integrated

CHALLENGES AHEAD

- **Logic – NVM Connectivity**

→ Compatible NVM and Logic Scaling

→ Cost Effective 3D solutions

- **Parameters loading - Learning**

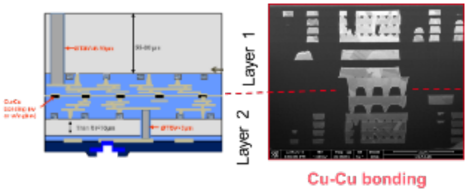
→ Programmation methods

→ Unsupervised learning

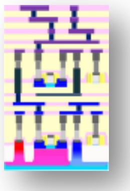
- **Operation**

→ Spike based system with NVM

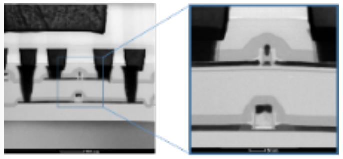
→ Denser 3D system



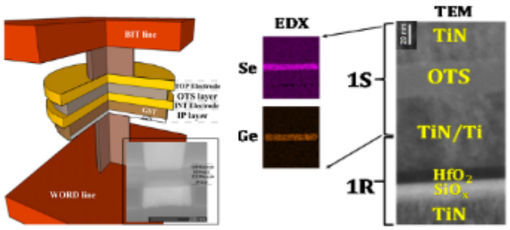
Wafer stacking
Alignment 50-100nm



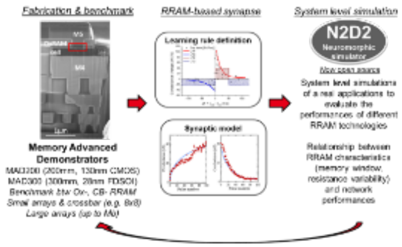
CoolCube™



CoolCube™
Alignment <3nm

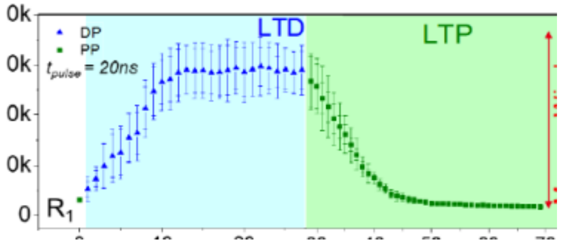


Selector + NVM integration

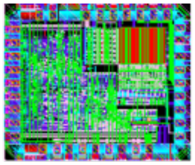

Fabrication & benchmark
RRAM-based synapse
System level simulation
Learning rule definition
Synaptic model
N2D2
Neuromorphic simulator
Now open source
Memory Advanced Demonstrators
MAD200 (200mm, 130nm CMOS)
MAD300 (300mm, 28nm FDSOI)
Benchmark btw Ox-, CB- RRAM
Small arrays & crossbar (e.g. 8x8)
Large arrays (up to Mb)
System level simulations of a real applications to evaluate the performances of different RRAM technologies
Relationship between RRAM characteristics (memory window, resistance variability) and network performances

Full framework for spike based design

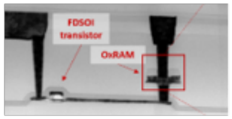
NVM as analog synapse



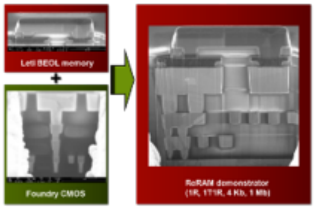
WANT TO BE PREPARE FOR THE NEXT REVOLUTION IN COMPUTING EFFICIENCY ? BOOTH #852



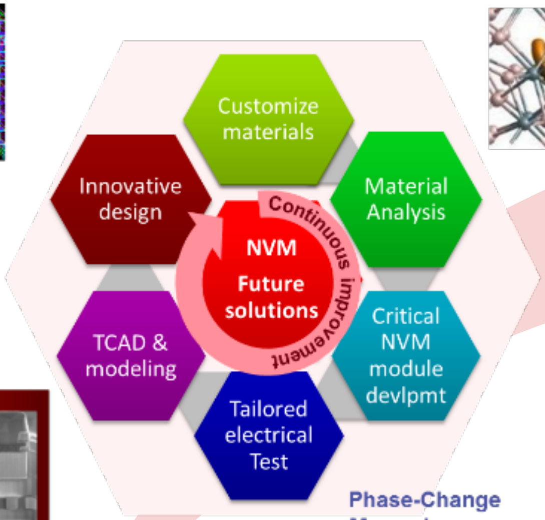
Dedicated NVM design

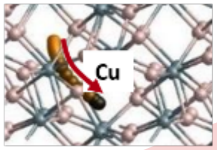


300mm 1T1R integration

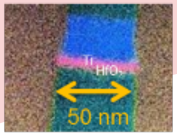


Memory Advanced Demonstrator (MAD)



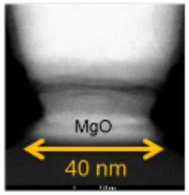


Ab-initio simulation

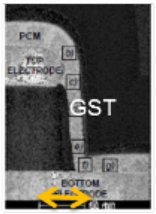


Resistive-RAM OxRAM & CBRAM

MRAM



Phase-Change Memories



A wide toolbox for customized research and benchmark between different BEOL NVM technologies

x 10^6
Your power efficiency !

Meet us @ Booth #852