How new Non Volatile Memories Are Enabling Disruptive Computing Architectures ?

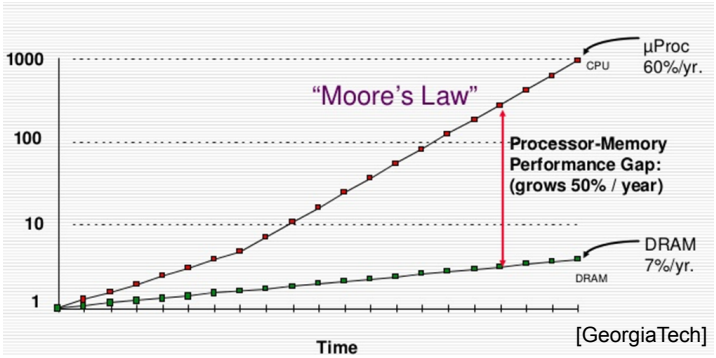
Denis Dutoit, CEA-LETI

B. Giraud, M. Harrand, H. Metras, E. Novak, E. Richet, A. Valentian

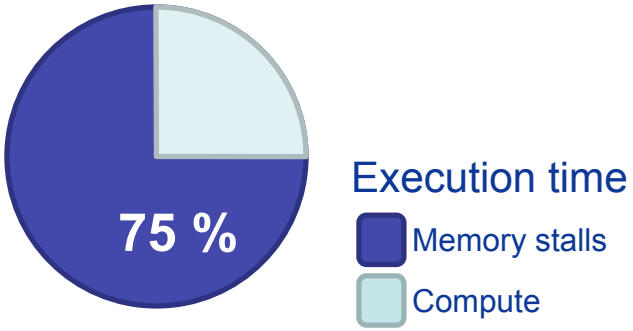
Session: 301-A "Life Beyond Flash: New Non-Volatile Memory Technologies"

The memory wall

Memory bandwidth does not scale at processor bandwidth speed.



CPU is stalled while waiting for a cache refill.
Frequent swap to slow storage memory.

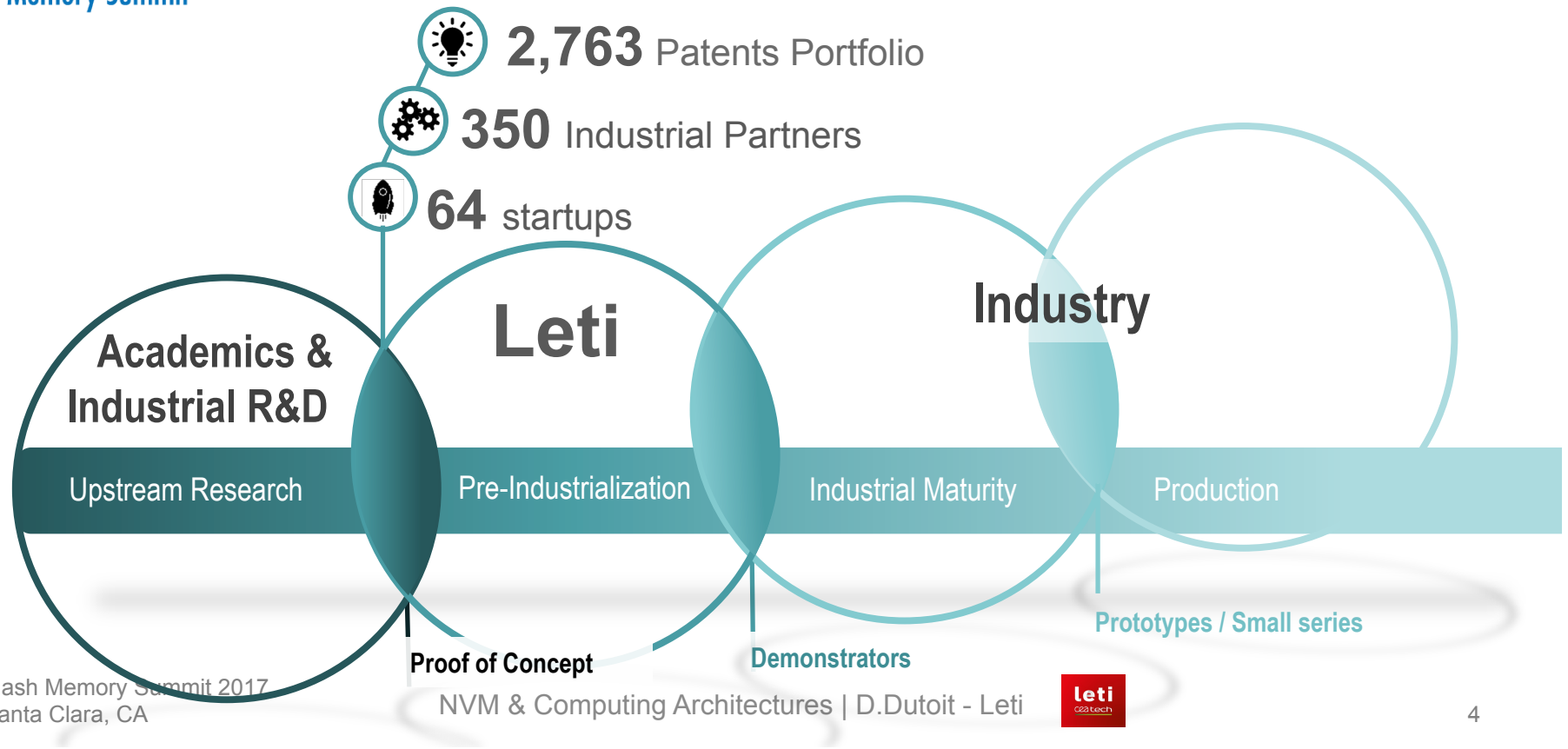


Filling the gap between compute and storage

Outline

- Highlights on Leti's memory technology portfolio and design activities
- How new Non Volatile Memories can help to solve current computing system limitations ?
  - Storage,
  - Main memory,
  - Embedded caches,
  - Neuromorphic architectures.

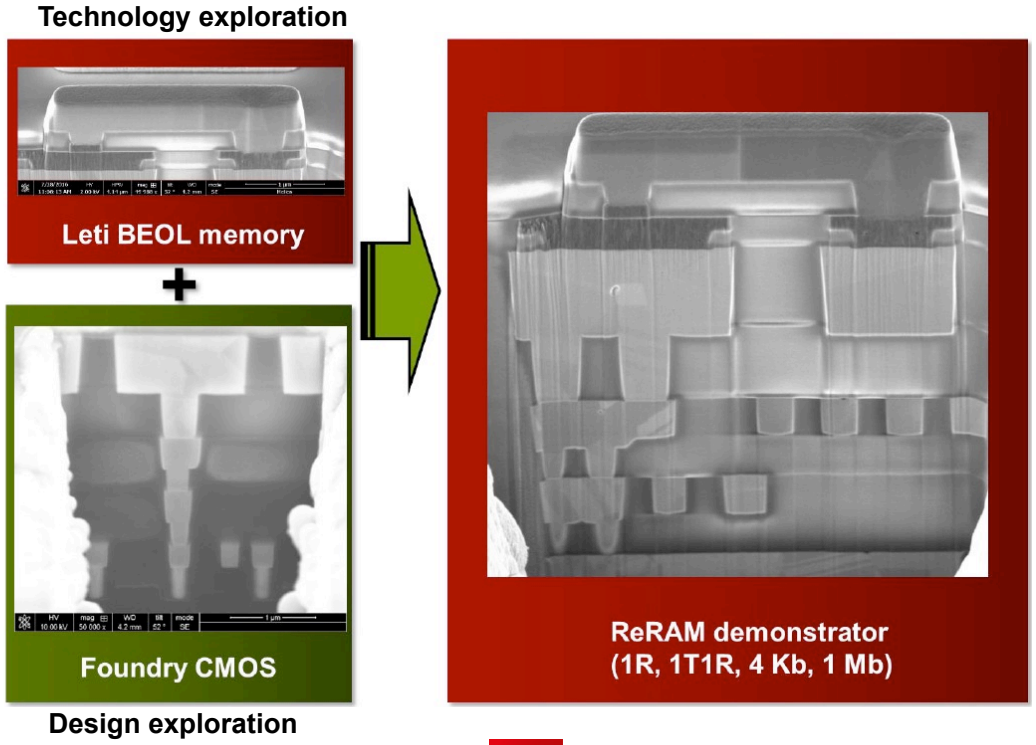
Leti at the heart of innovation



Memory technology portfolio from Leti

- MAD (Memory Advanced Demonstrators) test vehicle to integrate different back-end memory technologies on the same platform.
- Use of foundry CMOS for design exploration with mature circuit model

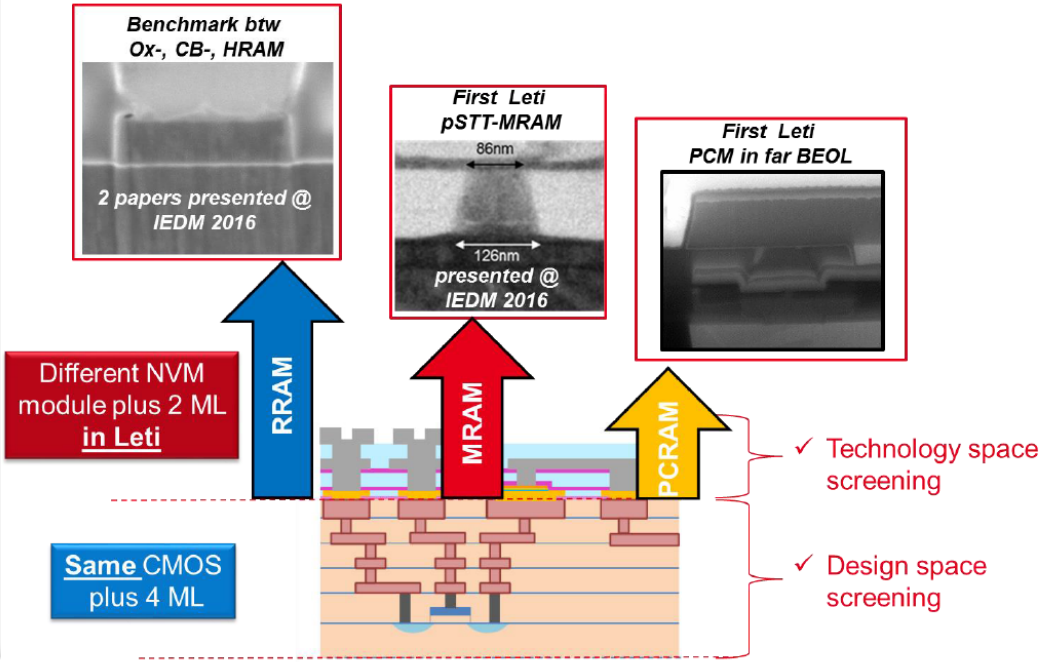
➔ Early **silicon** evaluation of alternative technology or design solutions



2016 Leti NVM achievements

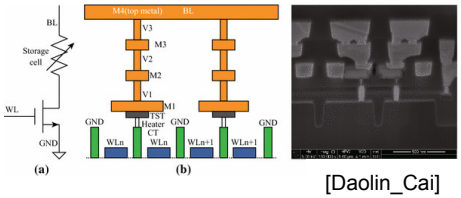
- Published integration result on OxRAM, CBRAM, MRAM and PCM.
- Integration of selector and Cross point architecture under going

➔MAD operational with most state of the art back-end memory available



Memory design flavors

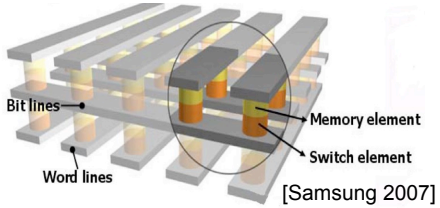
1T-1R



[Daolin_Cai]

- Easy writing and reading.
  No sneak-path issues.
  Simple peripheral circuits.
- High programming currents require big transistors.
  → Suitable for small to medium memories : embedded memories, caches…

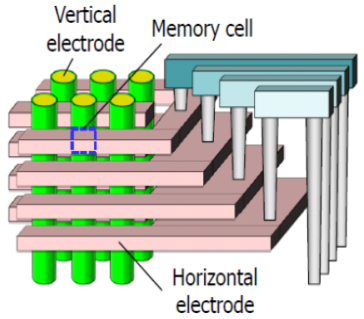
Crosspoint



[Samsung 2007]

- Density.
  → Very dense architecture suitable for high capacity memories.
- Current in narrow and long bit/ word lines is limited by electromigration and voltage drop.
  → Limitation of the size of the elementary array.

VRRAM



[I.G. Baek, Samsung, 2011 IEDM]

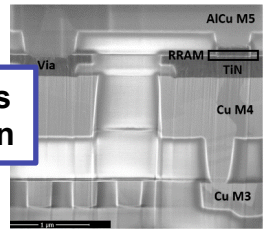
- Reduced number of masks w.r.t. crosspoint.
- Density limited by [metal line + horizontal RRAM device] pitch and voltage drop in the plate and pillar.
- But higher number of layers achievable for the same cost.

Leti's experience in memory designs

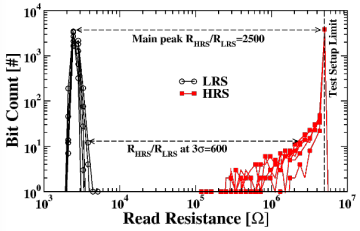
1T-1R: Tape out

Variability limits characterization

4Kb 1T-1R OxRAM



Description of the integration flow and TEM cross section of the integrated TiN/HfO_2/Ti/TiN OxRAM.

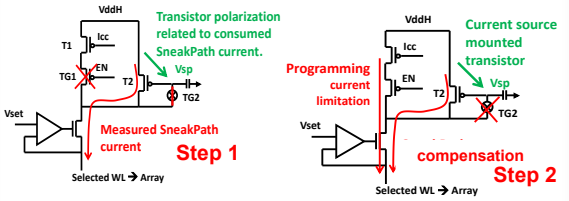


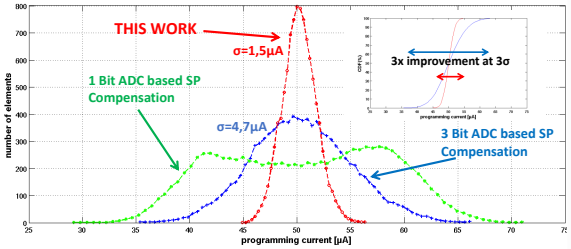
LRS and HRS distributions measured on 4kbits array with Set and Reset conditions allowing to obtain large resistance window

[Grossi IEDM2016]

Crossbar and compensation techniques

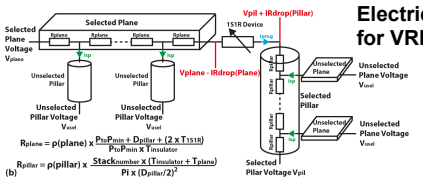
Proposed capacitor-based analog sampling approach





[Levisse – Nanoarch 2016]

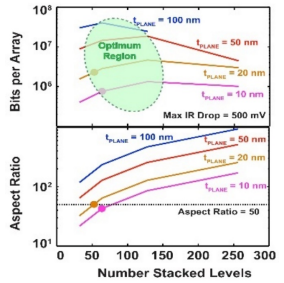
Models for vertical RRAM



Electrical Model for VRRAM

Array size estimation (top) and impact on the pillar aspect ratio, i.e. process cost (bot.)

Array size trade-off: IRdrop / pillar aspect ratio / process cost

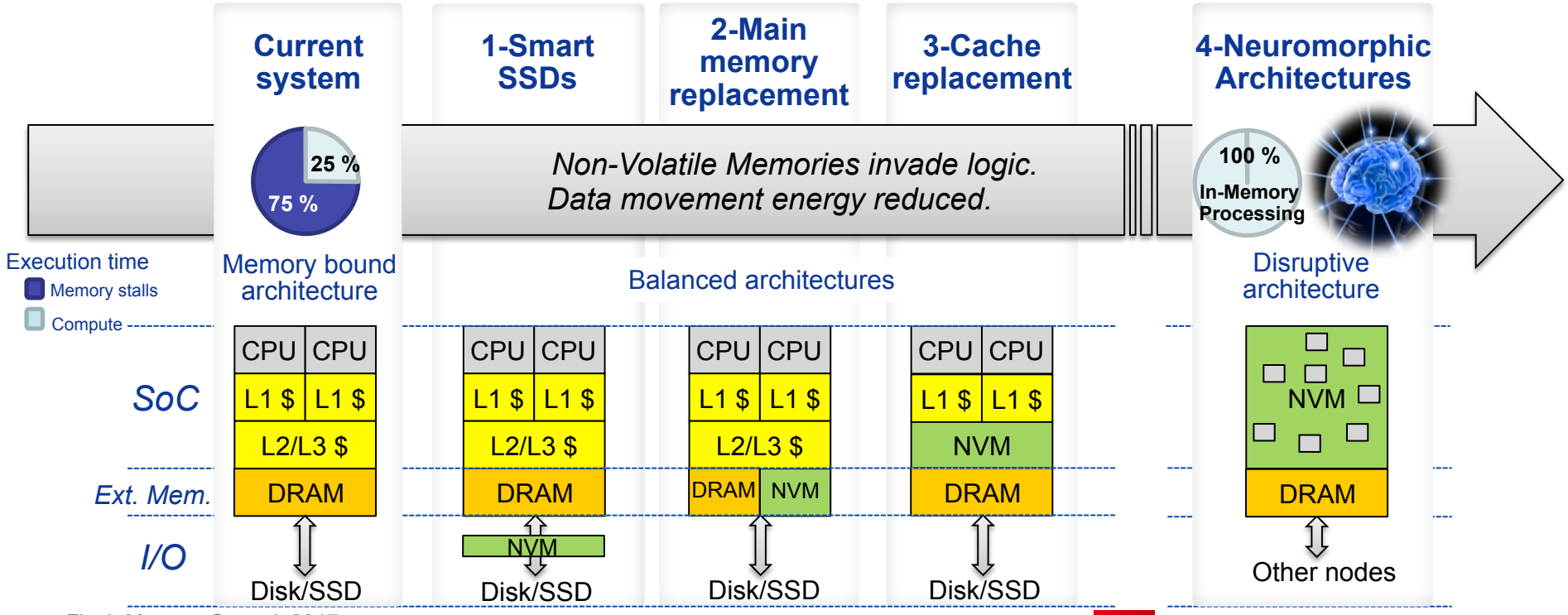


[Piccolboni IMW 2016]
[Levisse LASCAS 2017]

Outline

- Highlights on Leti's technology portfolio and design activities
- How new Non Volatile Memories can help to solve current computing system limitations ?
  - Storage,
  - Main memory,
  - Embedded caches,
  - Neuromorphic architectures.

New architecture paradigms with new NVM



Storage: Smart SSDs

1-Smart SSDs

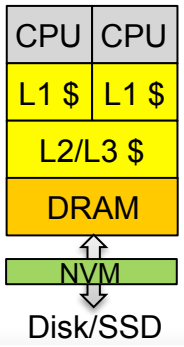
New NVM to reduce latency between SSD and main memory.

Target specifications:

Latency	Endurance	Retention	Cost/b	Power consumption
1 – 5µs	> 10^6	10 years @ 85°C	< 2 x flash cost	< flash

- The cost can be higher than flash because it is so much faster.

→ RRAM or PCM in crossbar or preferably VRRAM architecture



- Impacts on the computing system to take full benefit of smart SSDs:
  - The network bandwidth and latency must be improved,
  - H/W support for page management.
- Some computing capability may be inserted in these ultra fast SSD ("Processing In Memory"):
  - Key/value access,
  - Map functions of MapReduce algorithms,
  - …

Main memory

2-Main memory replacement

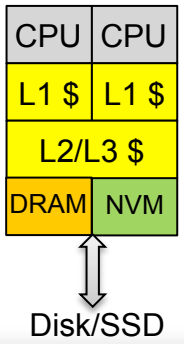
New NVM to improve main memory density and to reduce power consumption

Target specifications:

Latency	Endurance	Retention	Cost/b	Power consumption
< 50ns (< 200ns if mix of NVM and DRAM)	> 10^9 if smart wear-leveling is applied	> 5 days @ 85°C	< ½ DRAM cost	< DRAM

- Retention should be high enough to:
  - Prevent refresh to cause additional wear,
  - Recover power supply after a power loss.
- Cost must be sufficiently lower than DRAM to make this architecture competitive.
- But no improvement in latency is forecast.

→ RRAM or PCM in crossbar architecture



- System opportunities:
  - Simplifies data integrity management (reduces check-pointing traffic),
  - Allows greater quantities of working memory.

Cache replacement

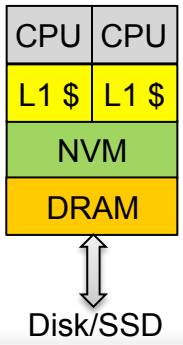
New NVM to improve cache density.

3-Cache replacement

Target specifications:

Latency	Endurance	Retention	Cost/b	Power consumption
< 30ns	> 10^{16}	> 5 days	< eDRAM cost	< eDRAM

- Improves last level cache density.
- May improve data recovery if jointly used with NVM main memory.
- It can be envisaged to use mixed SRAM/NVM caches to improve endurance:
  - Most often written data located in SRAM, others in NVM.



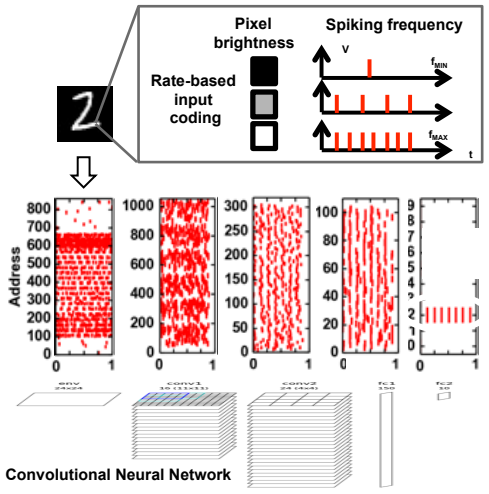
- Main requirements are speed and endurance.

→ **SST-MRAM**

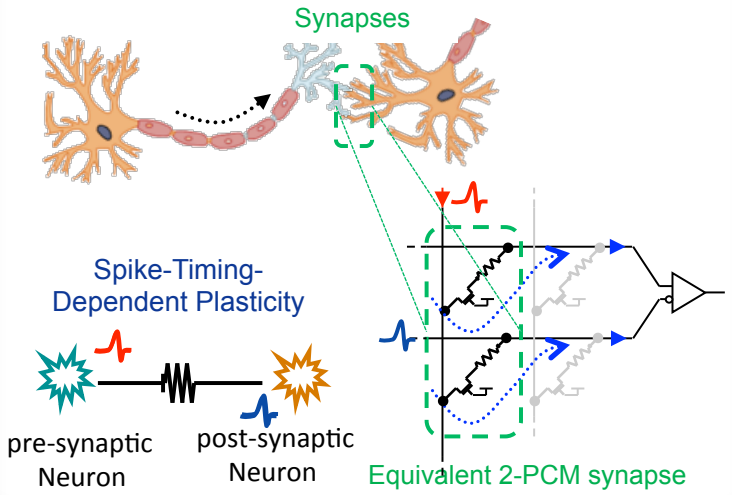
Neuromorphic Architectures

Merging compute, memory, storage and training

Spike-coding for Deep Neural Networks



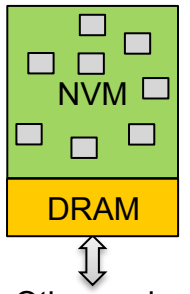
RRAM as synapse



[Vianello IMW2017]

4-Neuromorphic Architectures

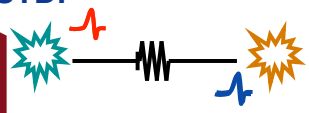
→ **RRAM**



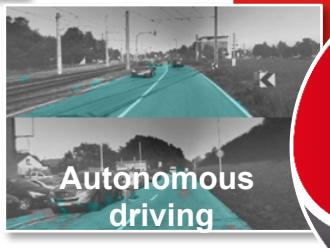
CEA tech assets in neuromorphic architectures

ImageClef2015

STDP



Real-time faces detection with gender & emotion



Application portfolio

Deep learning research

Software frameworks

Advanced implementa -tions

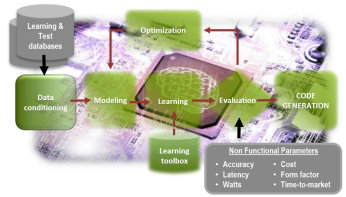
Hardware Accelerator

CO-DESIGN

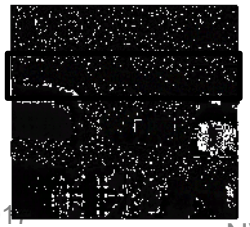
Exploration, NN generator, spike coding, co-design

N2D2

[Open Source]

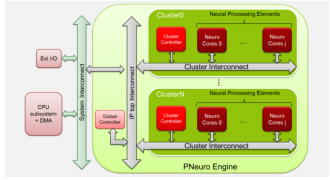


2-PCMs synapse

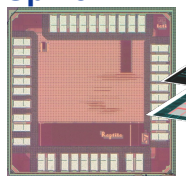


NeuRAM3

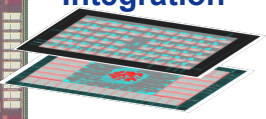
DSP



Spike



3D Integration



Conclusion: How new Non Volatile Memories Are Enabling Disruptive Computing Architectures ?

- By filling the gap between compute and storage in memory hierarchy:
  - Smart SSDs,
  - Main memory replacement,
  - Caches replacement.

➡ *No winner, the right memory type at the right place in the memory hierarchy.*

- By merging compute, memory, storage and training with Neuromorphic architectures.

➡ *Brings Artificial Intelligence to the edge.*

Thank You