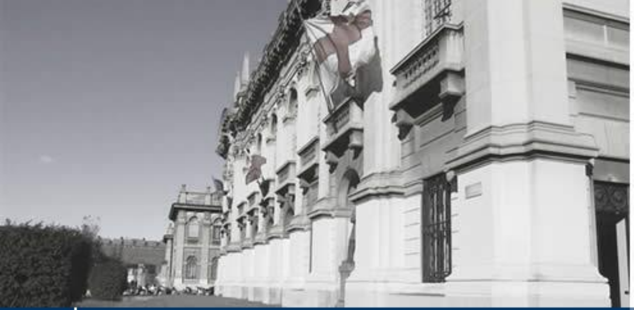


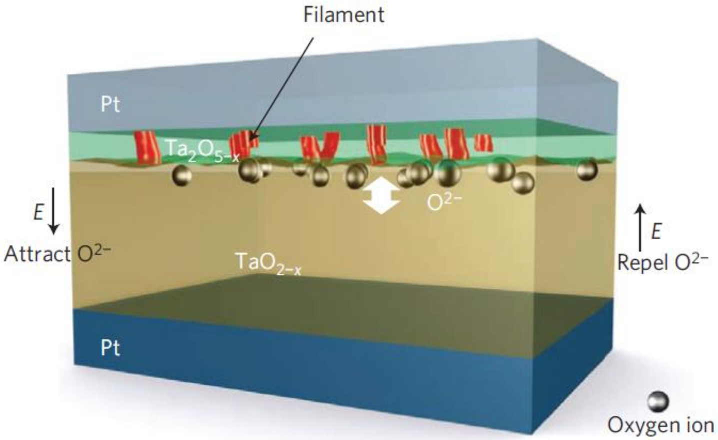
POLITECNICO DI MILANO

In-memory computing with emerging memory devices

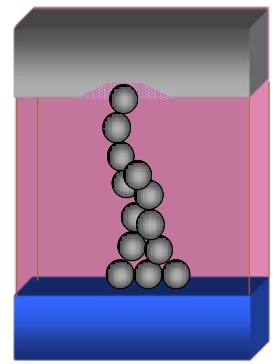
Daniele Ielmini

Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
daniele.ielmini@polimi.it

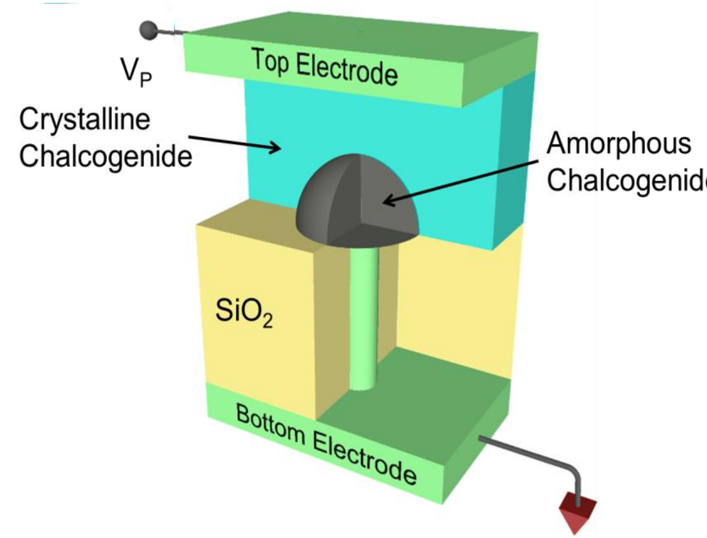
Emerging memory devices



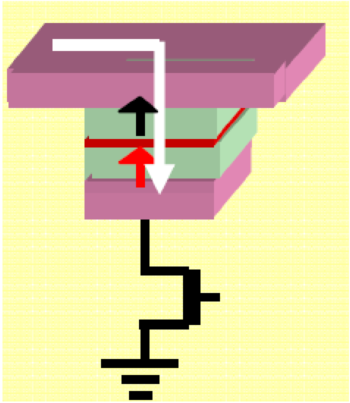
Resistive switching memory (RRAM)



Conductive bridge memory (CBRAM)

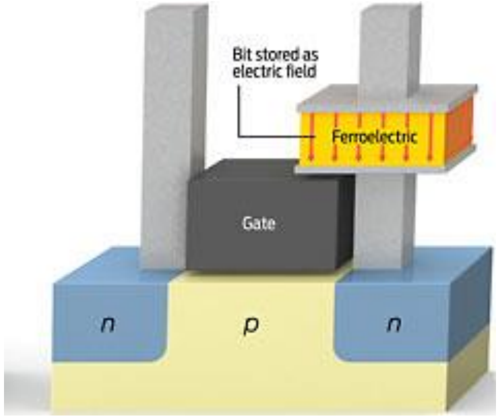


Phase change memory (PCM)

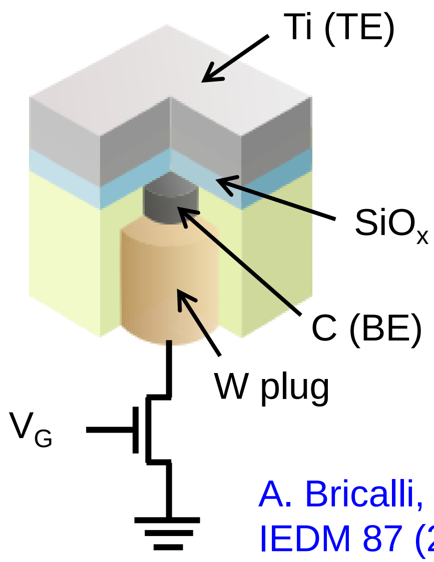


Spin-transfer torque memory (STT-RAM)

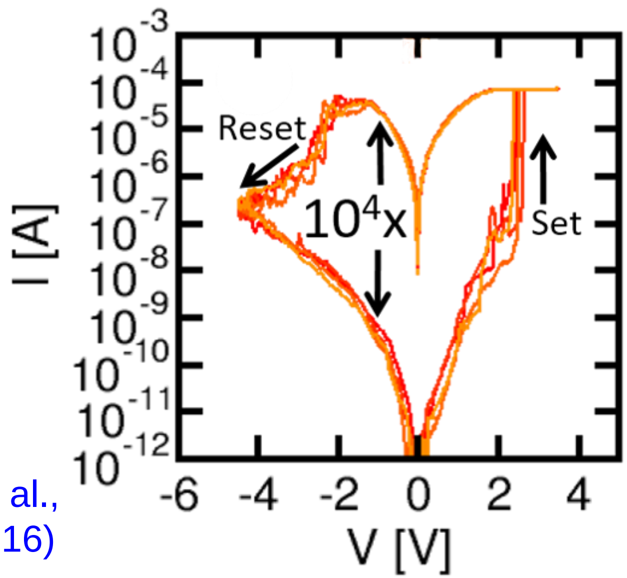
Ferroelectric memory (FeRAM)

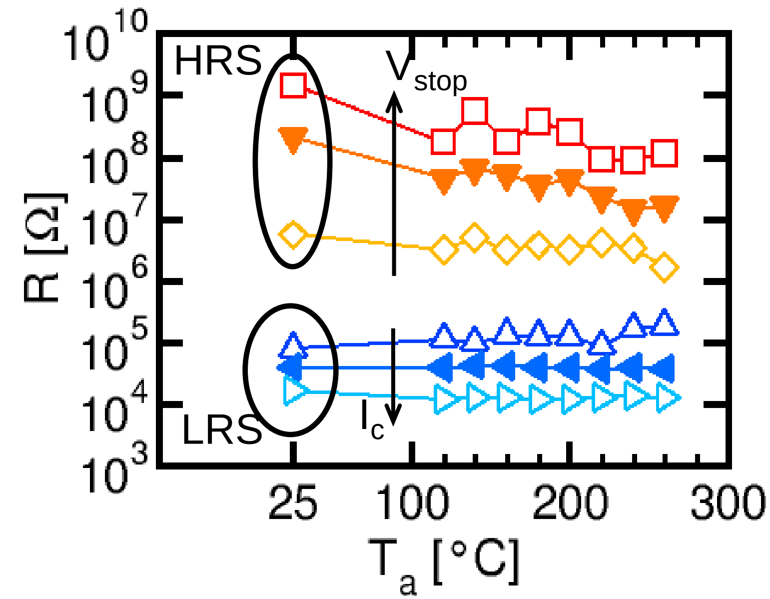


SiO_x-based RRAM devices @ Polimi



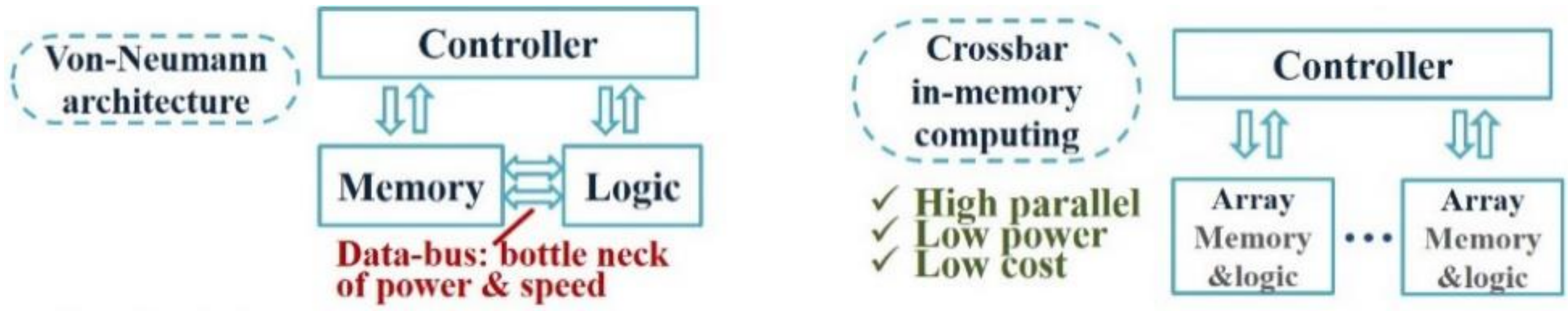
A. Bricalli, et al., IEDM 87 (2016)





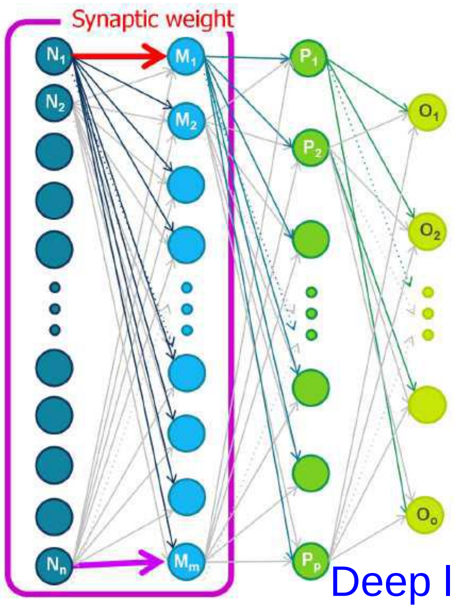
- Ti/SiO_x/C RRAM with 1T1R structure, 10^4 on-off ratio, 10^8 cycles, 1h @ 260°C retention
- Currently used for in-memory and neuro-computing projects:
  - RESCUE erc
  - DEEPEN FARE RICERCA IN ITALIA FRAMEWORK PER L'ATTRAZIONE E IL RAFFORZAMENTO DELLE ECCELLENZE PER LA RICERCA IN ITALIA

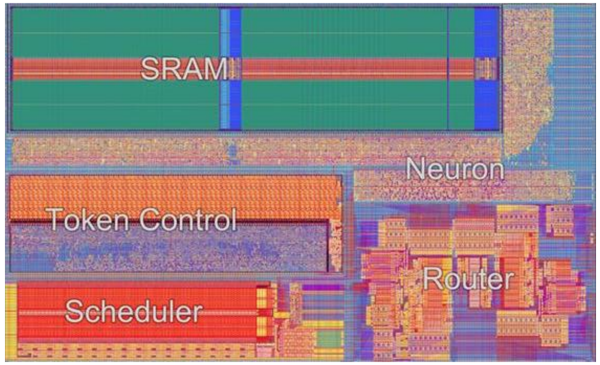
Why in-memory computing?



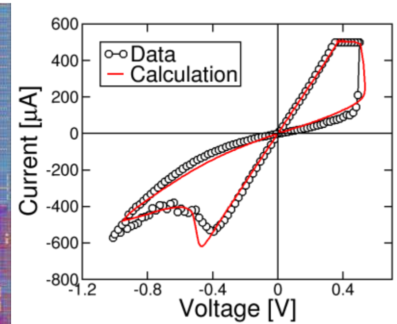
- Advantages:
  - In-memory computing: no distinction between memory and logic to overcome von Neumann bottleneck
  - Nonvolatile state → zero off-state power
  - Crossbar array → high gate/synapse density + physical computing
  - Back-end, 3D → easy integration with CMOS and high density
- Drawbacks:
  - High current, high voltage → high dynamic power
  - Long switching time → limited speed
  - Limited endurance

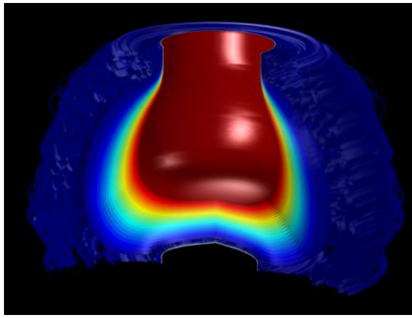
In-memory computing with emerging memory devices



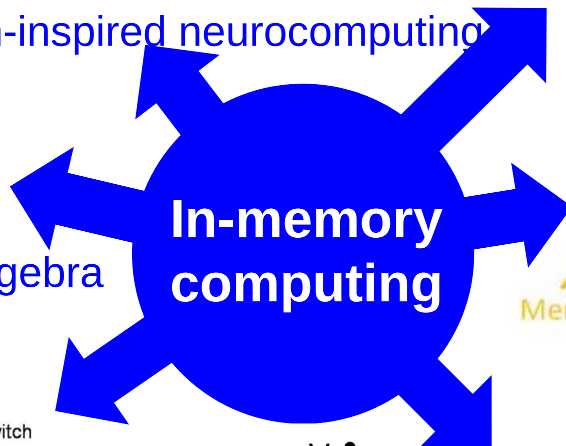


Brain-inspired neurocomputing

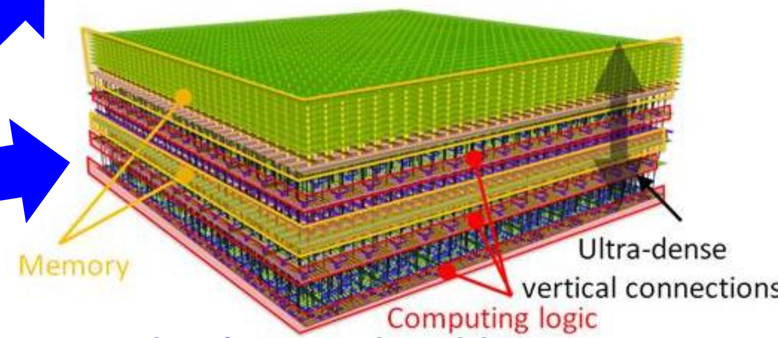




Device modeling

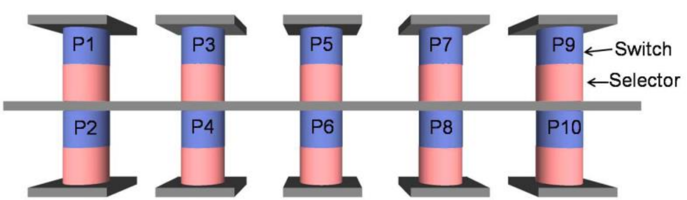


Deep learning
Approximate algebra

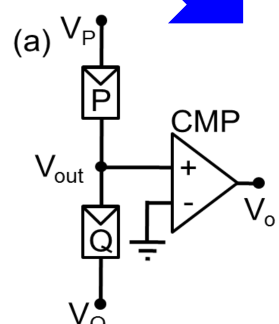


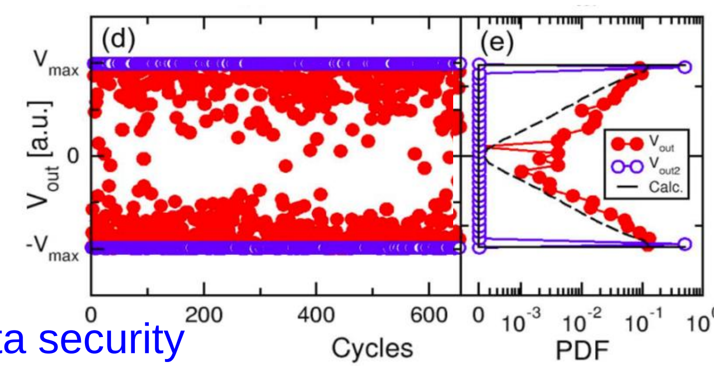
Technology and architecture

In-memory logic



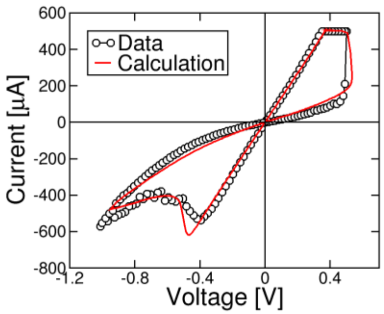
Clock=0	1	2	3	4	5
P1 = 0*	P2→ P1			P1 · P6	
P2 = A					
P3 = 0*		P4→ P3	P3 · P6		
P4 = B					
P5 = 0*					P6→ P5
P6 = 1			P3 · P6	P1 · P6	

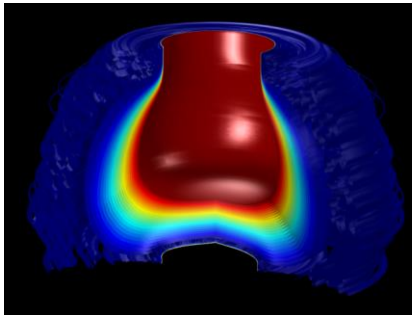




Chip/data security

Device modeling





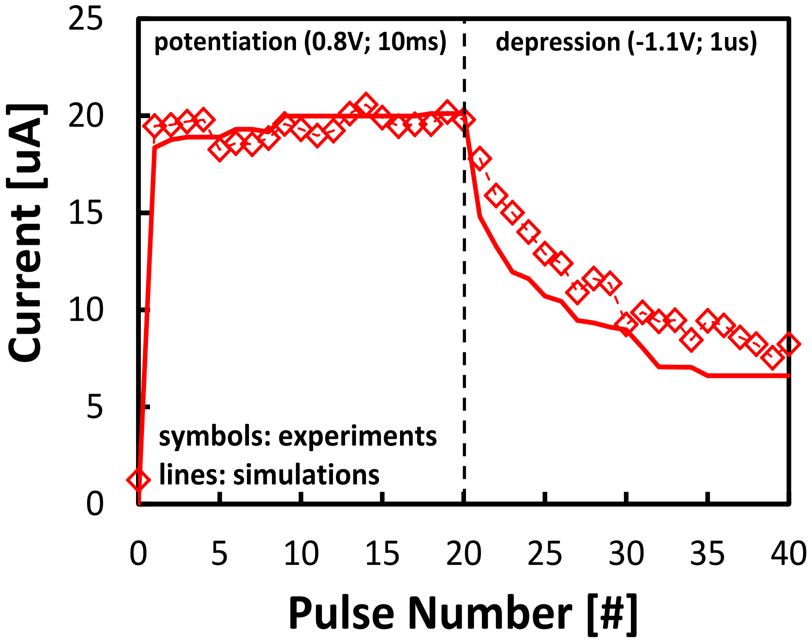
Device modeling

In-memory computing

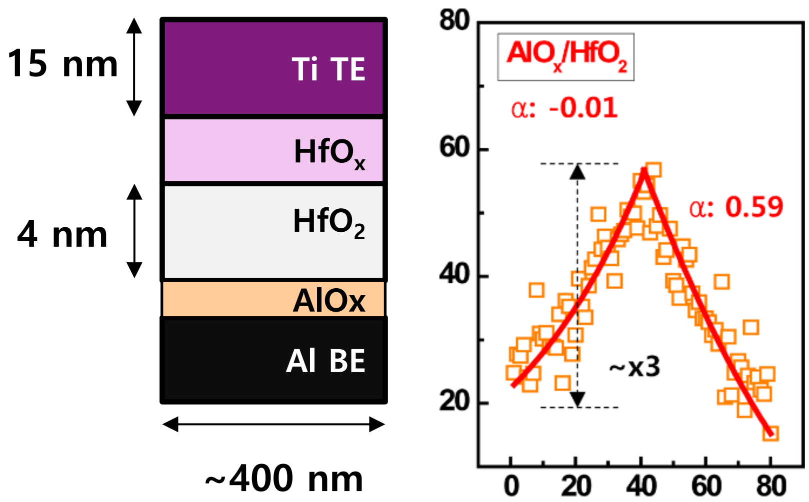
Unimore: neuromorphic computing device

- Investigation of material/device properties for analog switching
- Development of a multiscale modeling and simulation platform supporting design on novel devices and measurement interpretation
- Technologies targeted: RRAM, FeRAM

Abrupt switching in HfO_x RRAM in pulse regime

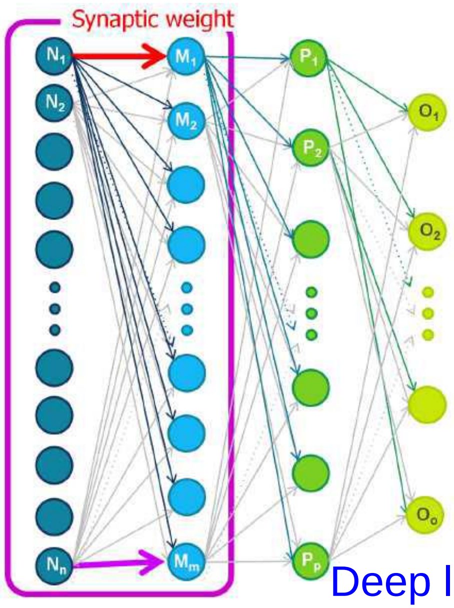


Analog switching in AlO_x-HfO_x stacks



L. Larcher et al, IEDM 2017

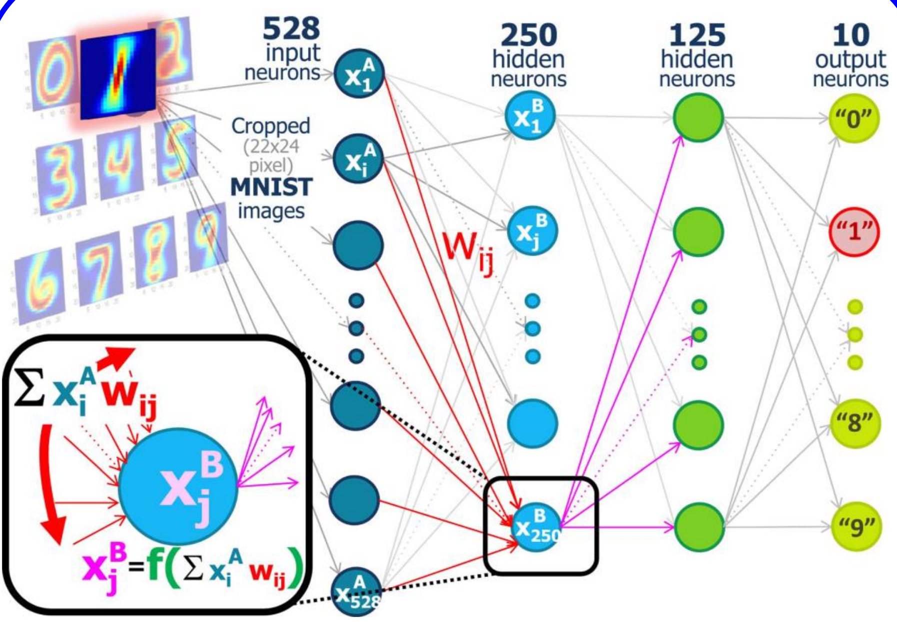
Deep learning



Deep learning
Approximate algebra

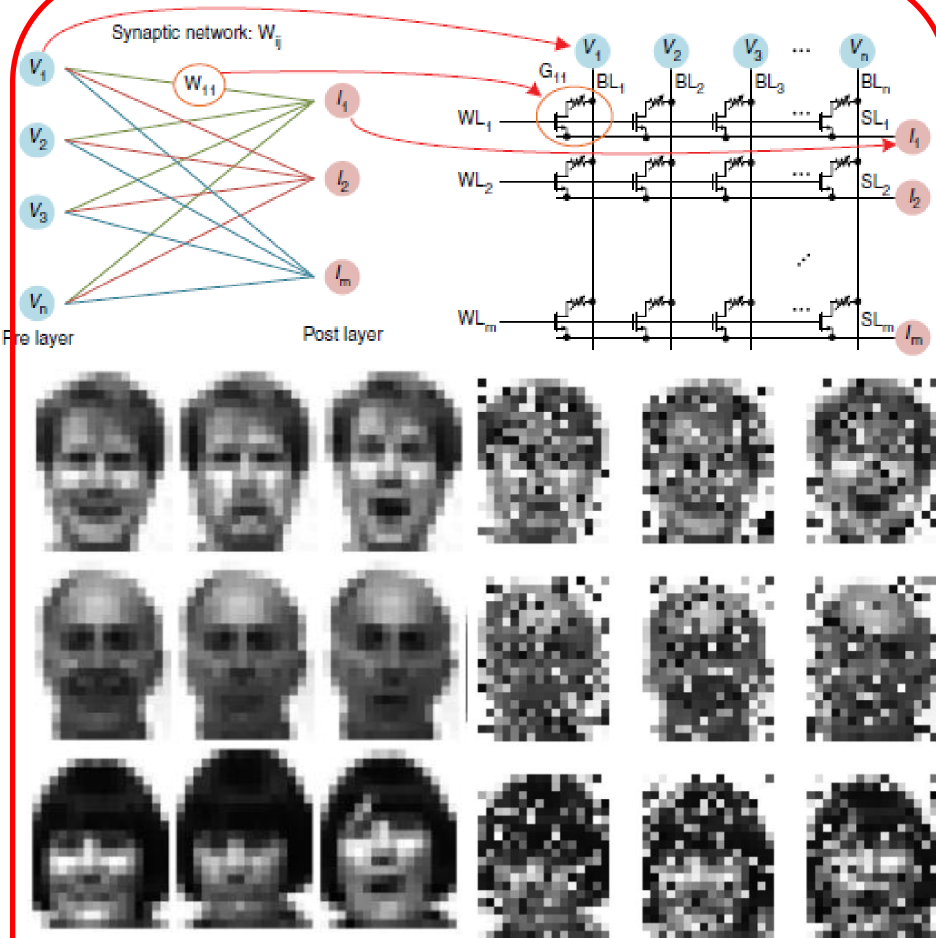
In-memory computing

Deep neural networks (DNN) with emerging memory



Learning and recognition of 60,000 handwritten digits by PCM array

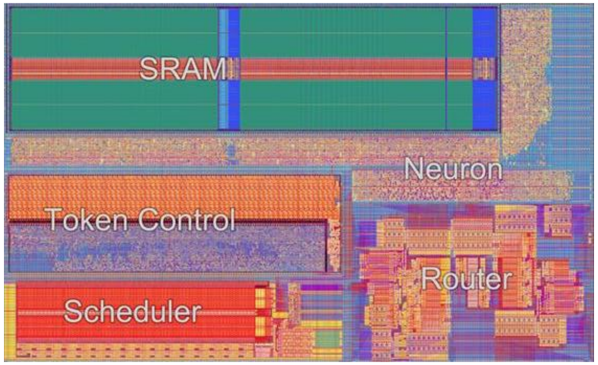
G. W. Burr, et al., IEDM Tech. Dig. (2014)



Learning and recognition of 9 faces by RRAM array

P. Yao, et al., Nat. Comm. (2017)

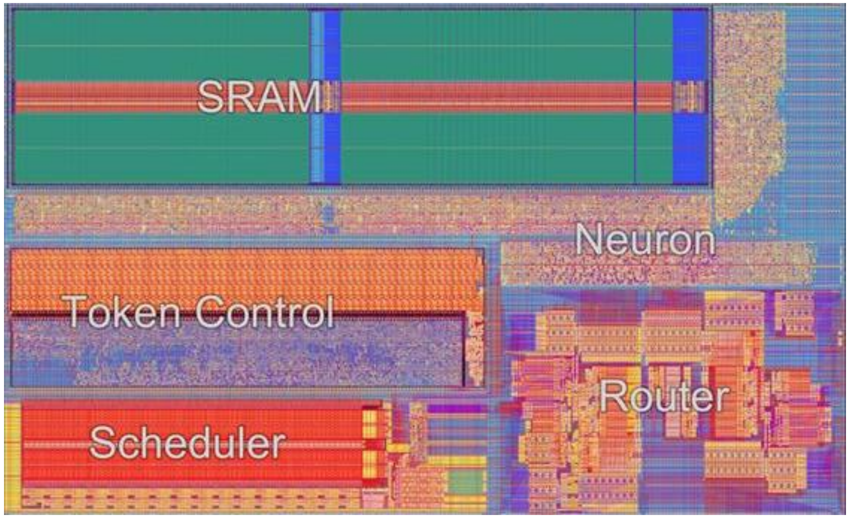
Brain-inspired neurocomputing



Brain-inspired neurocomputing

In-memory computing

A different approach: brain-inspired computing



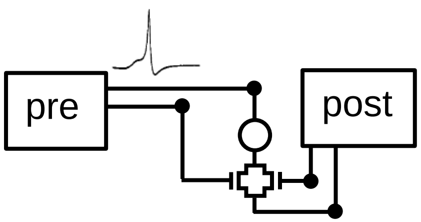
P. A. Merolla, et al., Science 345 (2014)

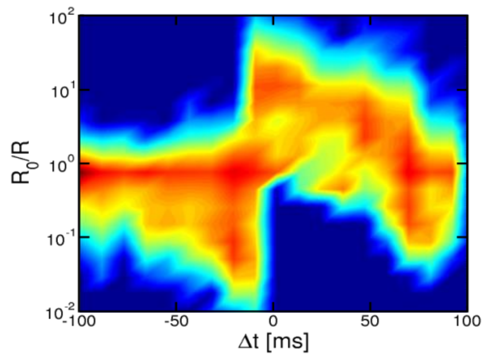
	IBM TrueNorth	Human brain	Gap
Number of cores	4096		
Number of neurons	10^6	0.86×10^{11}	$\sim 10^5$
Number of synapses	2.5×10^8	1.5×10^{14}	$\sim 10^6$
Power	63 mW	20 W	$\sim 10^3$

- A major breakthrough is needed to overcome the current limitation of brain-inspired computing technology

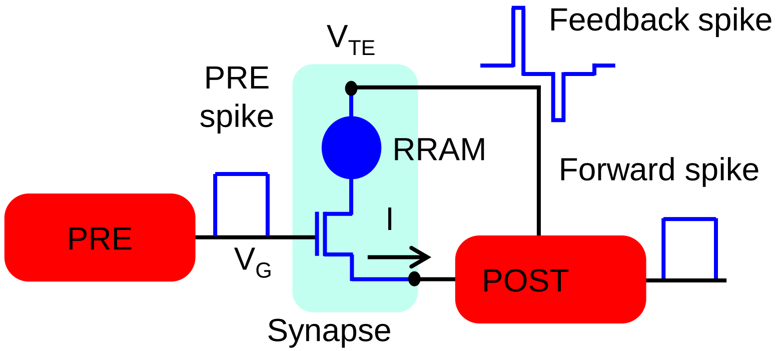
Neuromorphic computing @ Polimi

Hybrid CMOS/memristive synapses



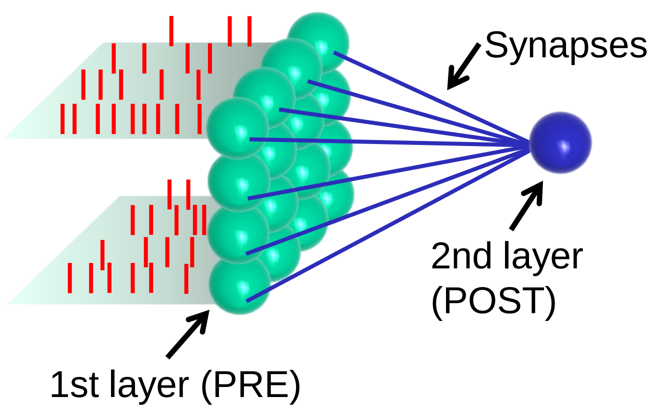


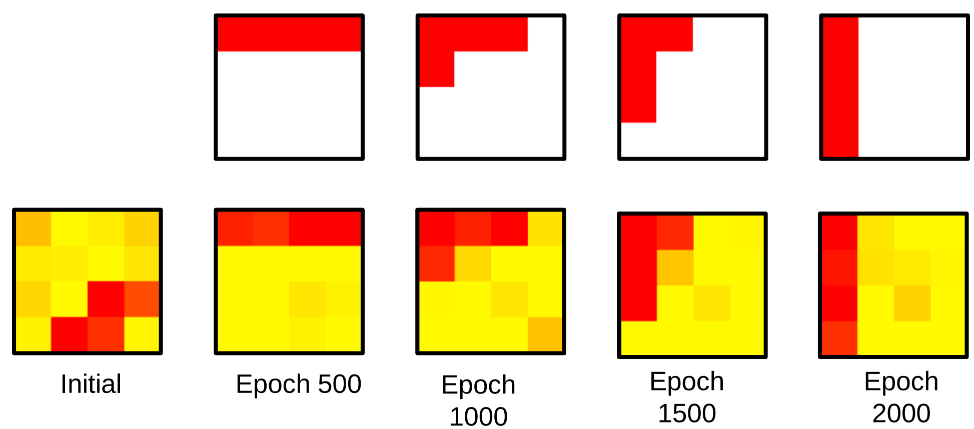
Z.-Q. Wang, et al., Front. Neurosci. 8, 438 (2015)



S. Ambrogio, et al., IEEE TED 63 (2016)

Demonstration of unsupervised learning





G. Pedretti, et al., Sci. Rep. 7:5288 (2017)

Projects on this topic

- Within Polimi:
  - ERC RESCUE (REsistive Switch CompUting beyond CMOS): develop resistive-switch computing devices, circuits and architectures for in-memory logic and brain-inspired neuro-computing (2015-2020)
  - DEEPEN (DEvicE-Physics Enabled Neuro-computing): develop neuromorphic circuits based on physical computing (2017-2020)
- Within EU:
  - NEURAM3 (Neural computing architectures in advanced monolithic 3D-VLSI nano-technologies): neuromorphic architecture with spiking neurons, FDSOI 28 nm, RRAM synapses (CEA Leti, CEA List, STMicro, imec, IBM, CNR, ETH, IMSE, JACU)
  - MNEMOSENE (Computation-in-memory architecture based on resistive devices)

Possible project outline

Devices	• Synaptic devices for spike timing dependent plasticity (STDP) • In-memory logic gates • Physical models • Compact models for circuit design and simulation • Algorithms for multilevel programming and physical computing
Circuits	• Design, simulation and fabrication of integrated neuro-circuits • Crossbar arrays for matrix-vector product and deep/convolutional neural networks • Integration of hybrid CMOS/memristive circuits
Systems	• Artificial vision • Robot/drone navigation • E-health • Traffic management
Neuroscience	• Interdisciplinary breath • Brain-inspired concepts
Neuroinformatics	• Simulation of spike processing and plasticity in neural networks