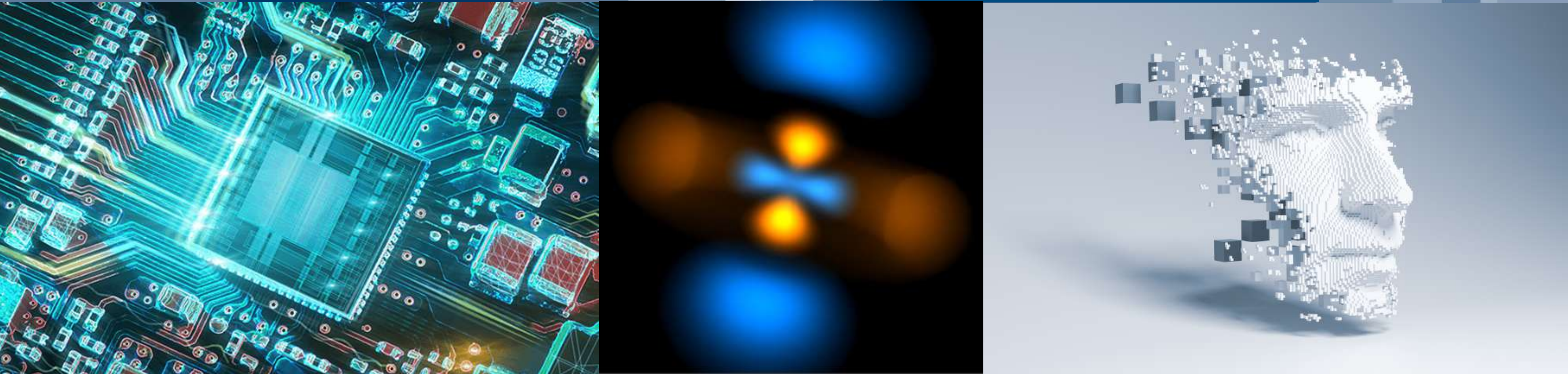
POLITECNICO
MILANO 1863

L'elettronica del futuro: bit, qubit e intelligenza artificiale



Sommario

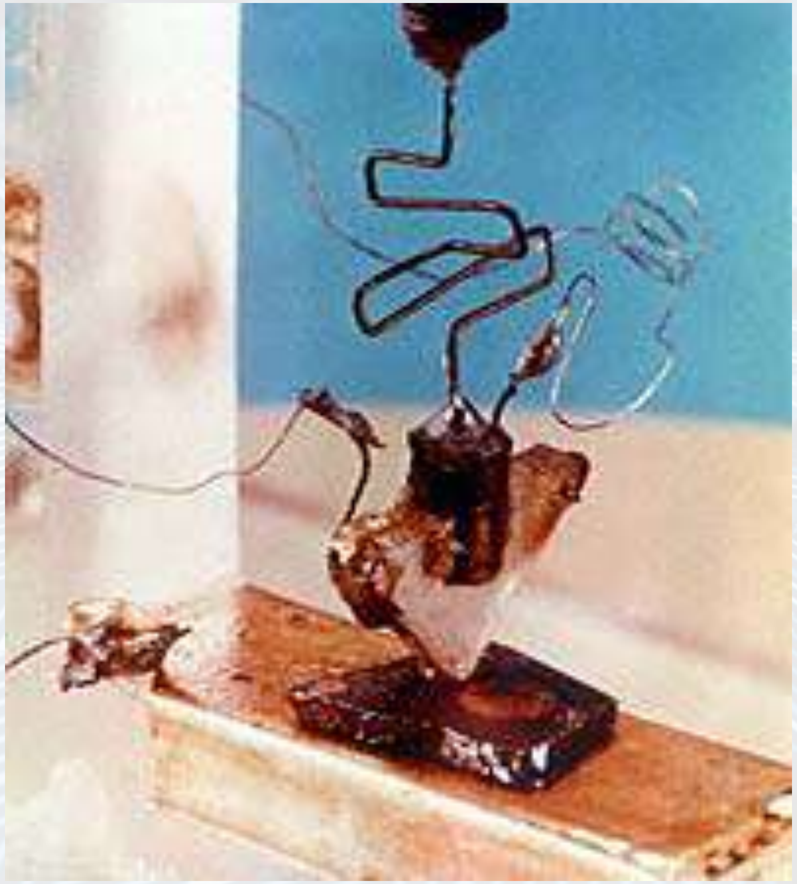
- Introduzione
- Breve storia dell’elettronica
- Evoluzione dell’elettronica: lo scaling
- L’elettronica del futuro: il quantum computing
- Intelligenza artificiale
- In-memory computing
- Conclusioni

Sommario

- Introduzione
- Breve storia dell’elettronica
- Evoluzione dell’elettronica: lo scaling
- L’elettronica del futuro: il quantum computing
- Intelligenza artificiale
- In-memory computing
- Conclusioni

1947: il primo transistore

J. Bardeen W. Shockley W. Brattain

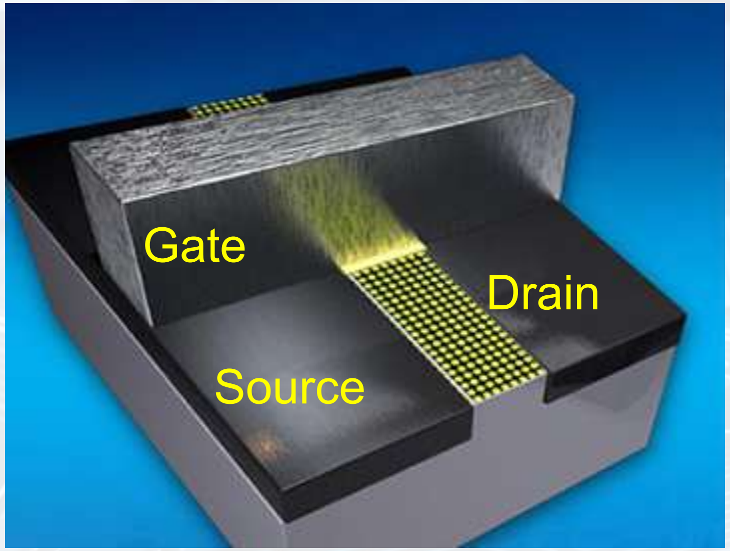


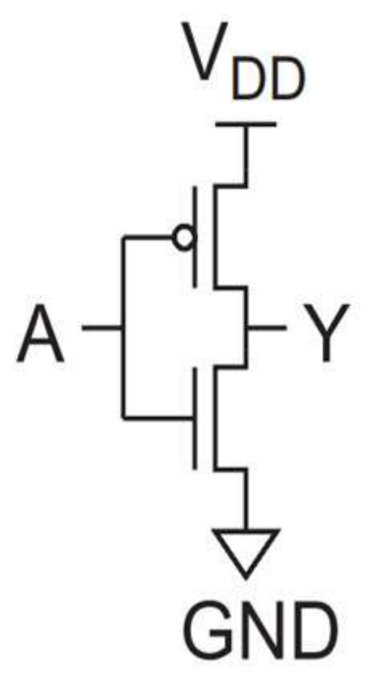
ALFR NOBEL
NAT MDCCC XXXIII OB MDCCC XCVI

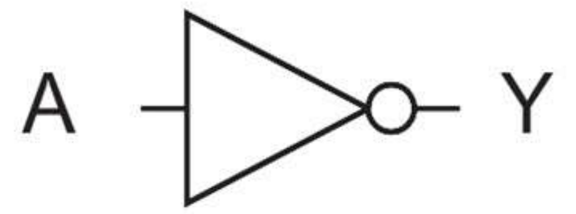
1956

- Dicembre 1947: Shockley, Brattain e Bardeen sviluppano un transistor a contatto di punto usando Ge

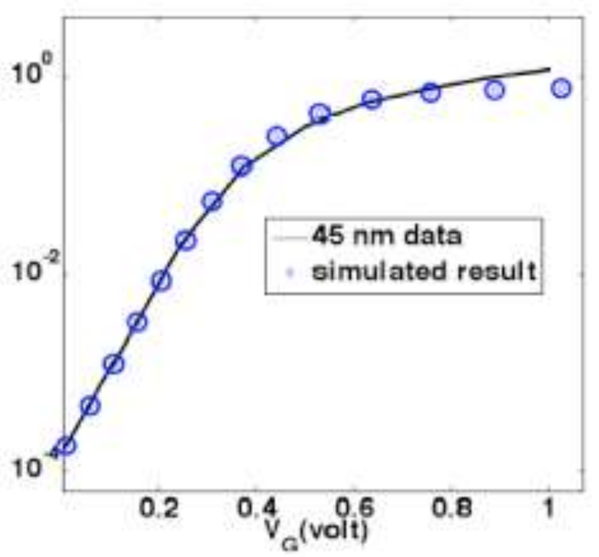
Il transistor e le porte logiche CMOS





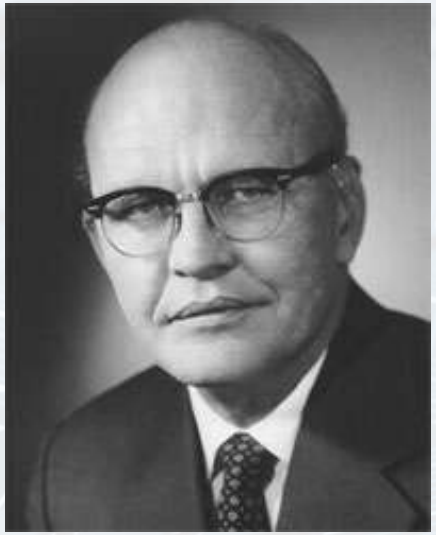


A	Y
0	1
1	0

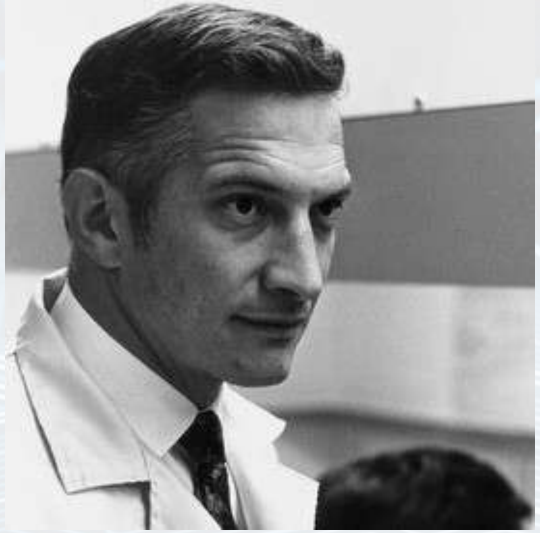


- Il transistore può essere visto come un interruttore di corrente pilotato dalla tensione di gate
- La tecnologia CMOS (complementary metal oxide semiconductor) permette di realizzare la porta logica NOT e tutte le altre operazioni logiche

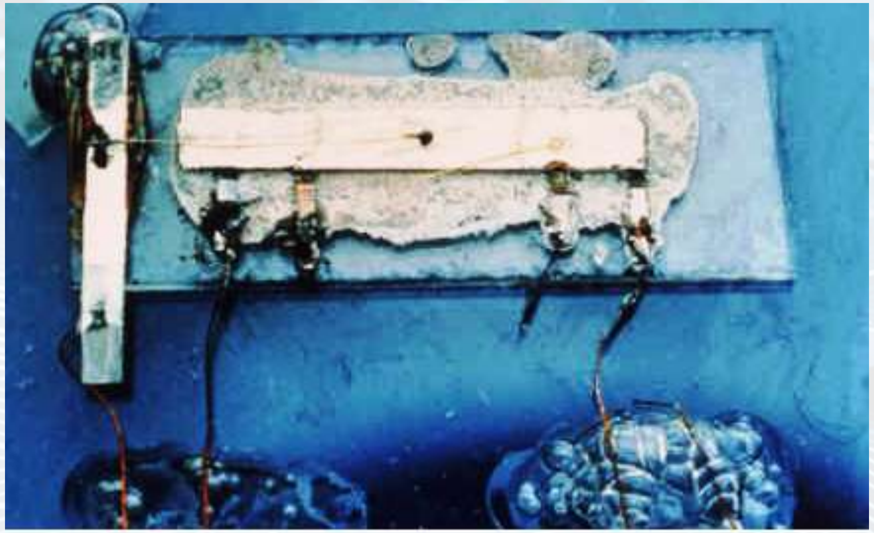
1958: il primo circuito integrato



J. Kilby
(1916 – 2001)



R. Noyce
(1927 – 1990)

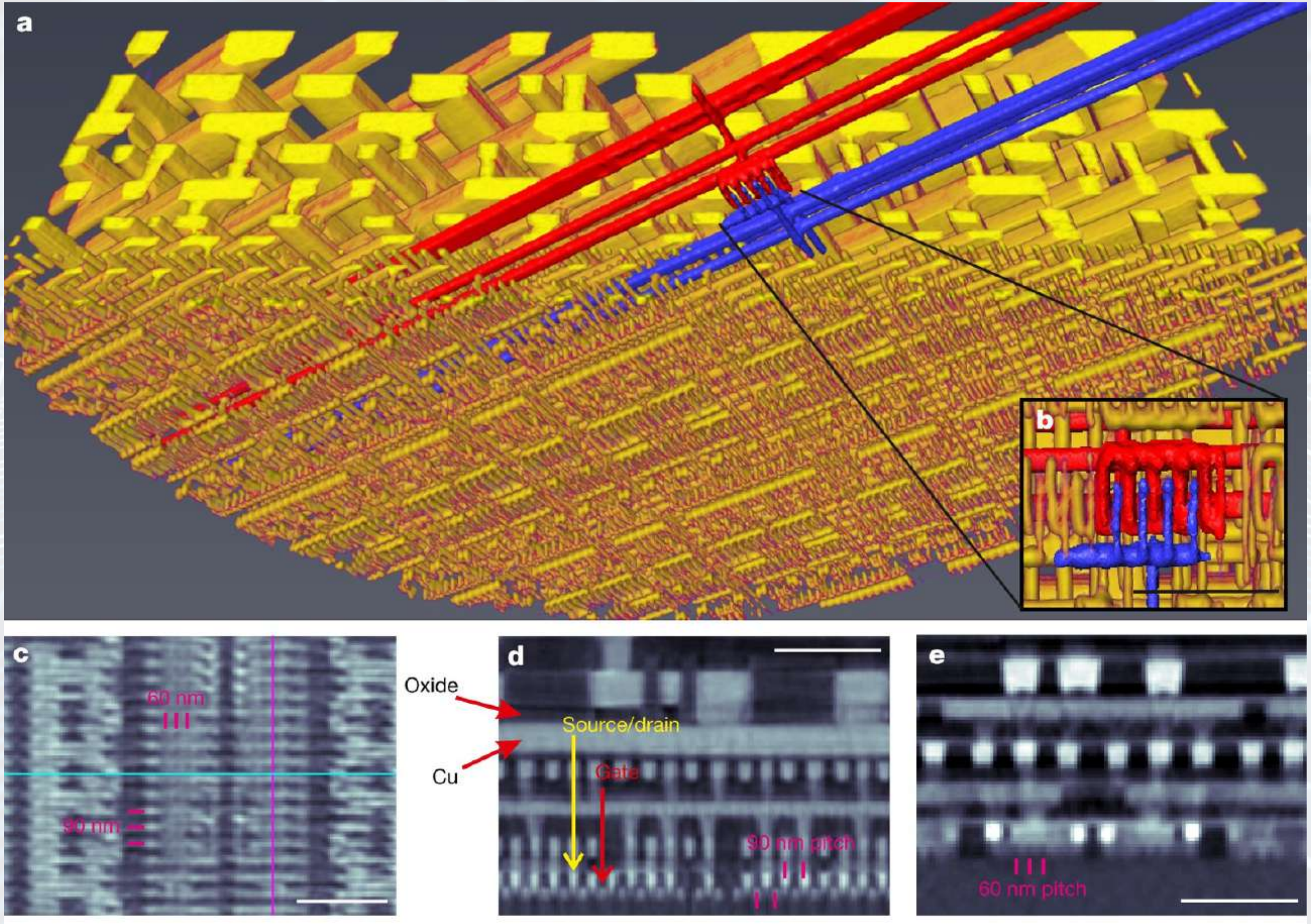


NAT
MDCCC
XXXIII
OB
MDCCC
XCVI
ALFR.
NOBEL

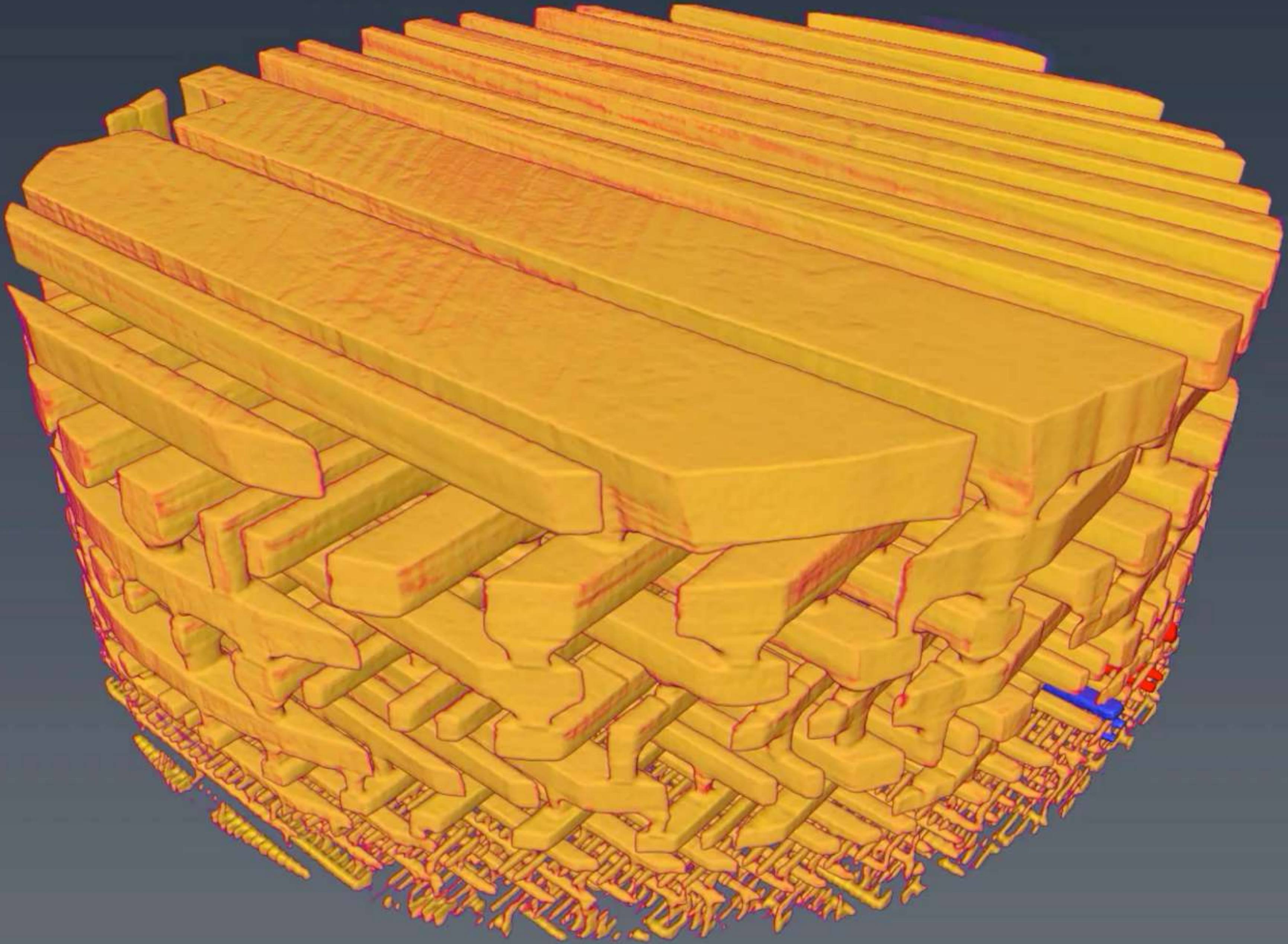
2000

- 1958: J. Kilby (Texas Instruments) inventa una tecnica per integrare tutti i dispositivi (transistor, R, C) in un solo cristallo di Ge
- Più tardi R. Noyce (Fairchild) migliora l’idea integrando anche le interconnessioni, preparando la strada per i circuiti integrati moderni

Un’architettura nanometrica


a
b
c
60 nm
90 nm
d
Oxide
Cu
Source/drain
Gate
90 nm pitch
e
60 nm pitch

M. Holler, et al., Nature 543, 402–406 (2017)

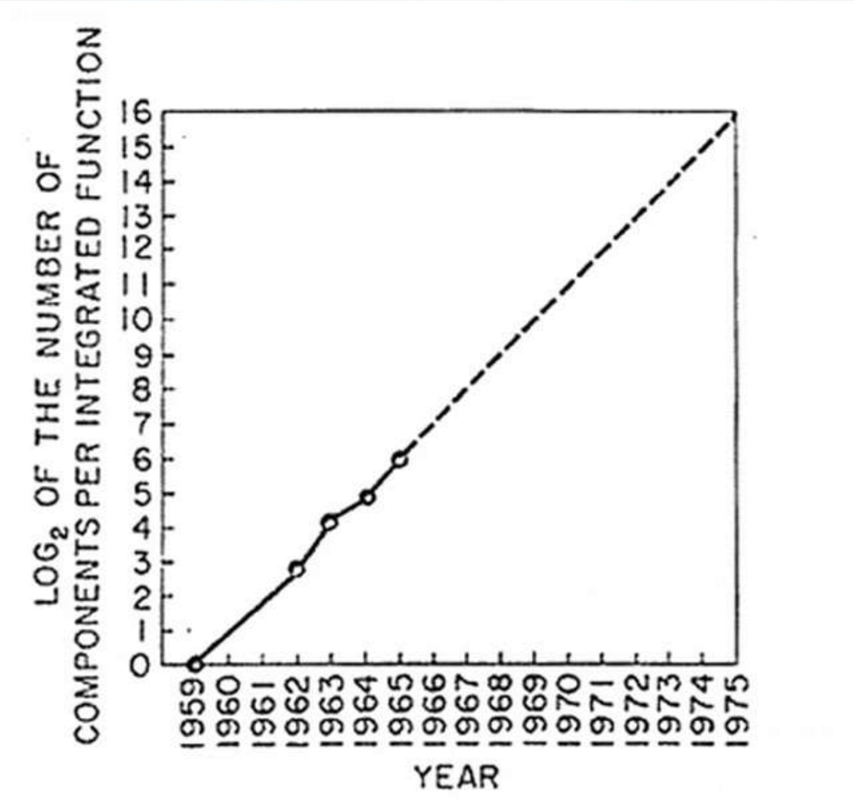


Sommario

- Introduzione
- Breve storia dell’elettronica
- Evoluzione dell’elettronica: lo scaling
- L’elettronica del futuro: il quantum computing
- Intelligenza artificiale
- In-memory computing
- Conclusioni

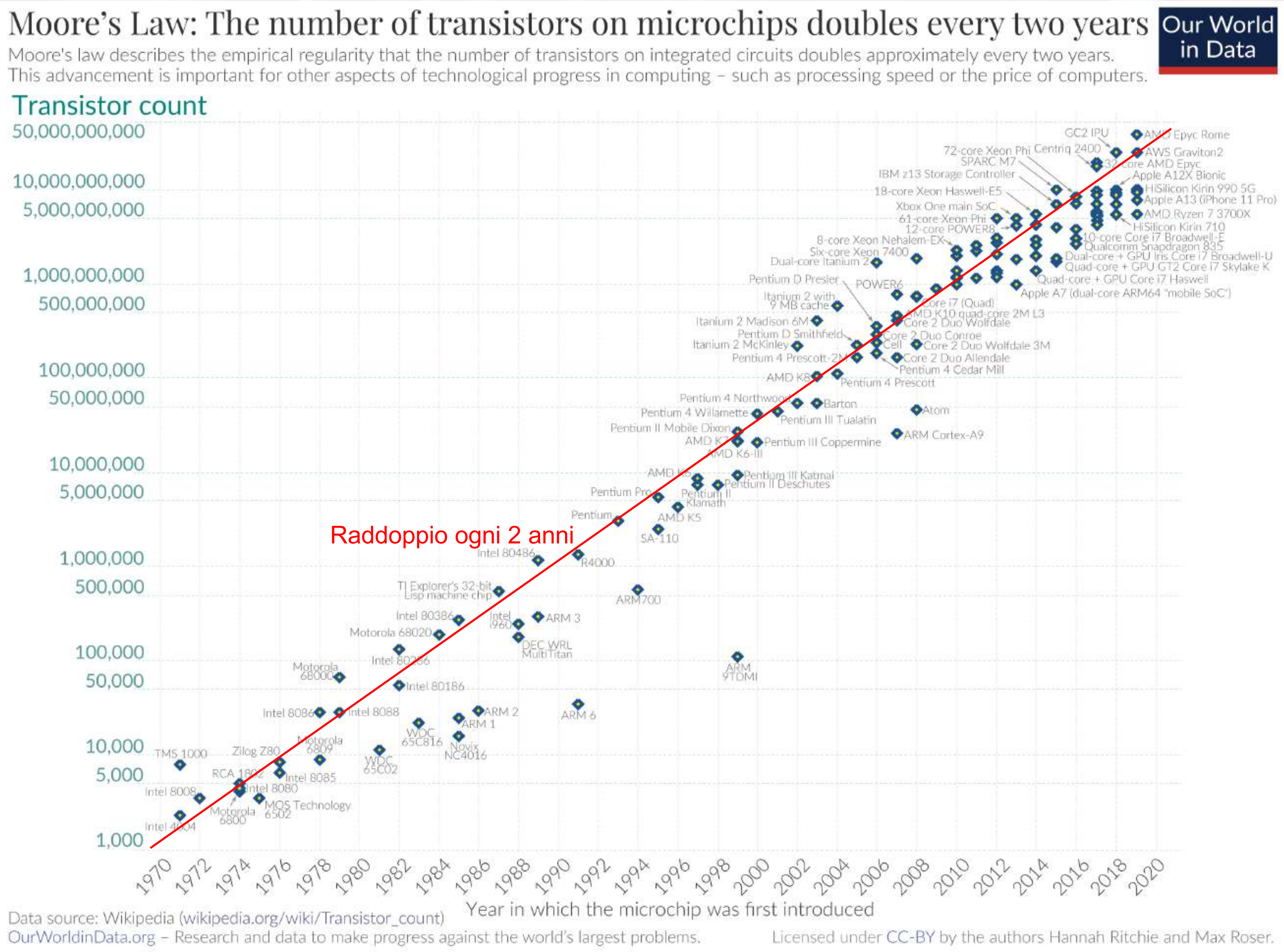
1965: la legge di Moore

Gordon Moore
1929-2023



- Dopo appena 4 anni dal primo circuito integrato commerciale, G. Moore osserva che il numero di transistori integrati raddoppia ogni anno
- Il costo per bit diminuisce grazie al miglioramento della tecnologia litografica

L’infallibile predizione di Moore



Moore’s Law: The number of transistors on microchips doubles every two years

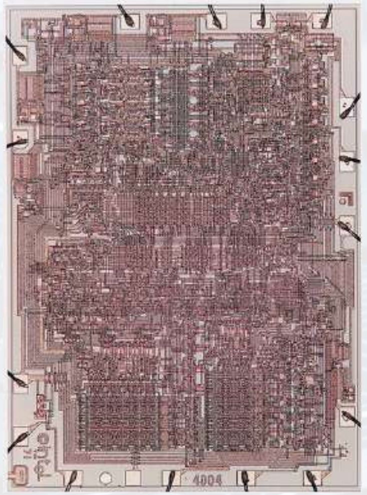
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

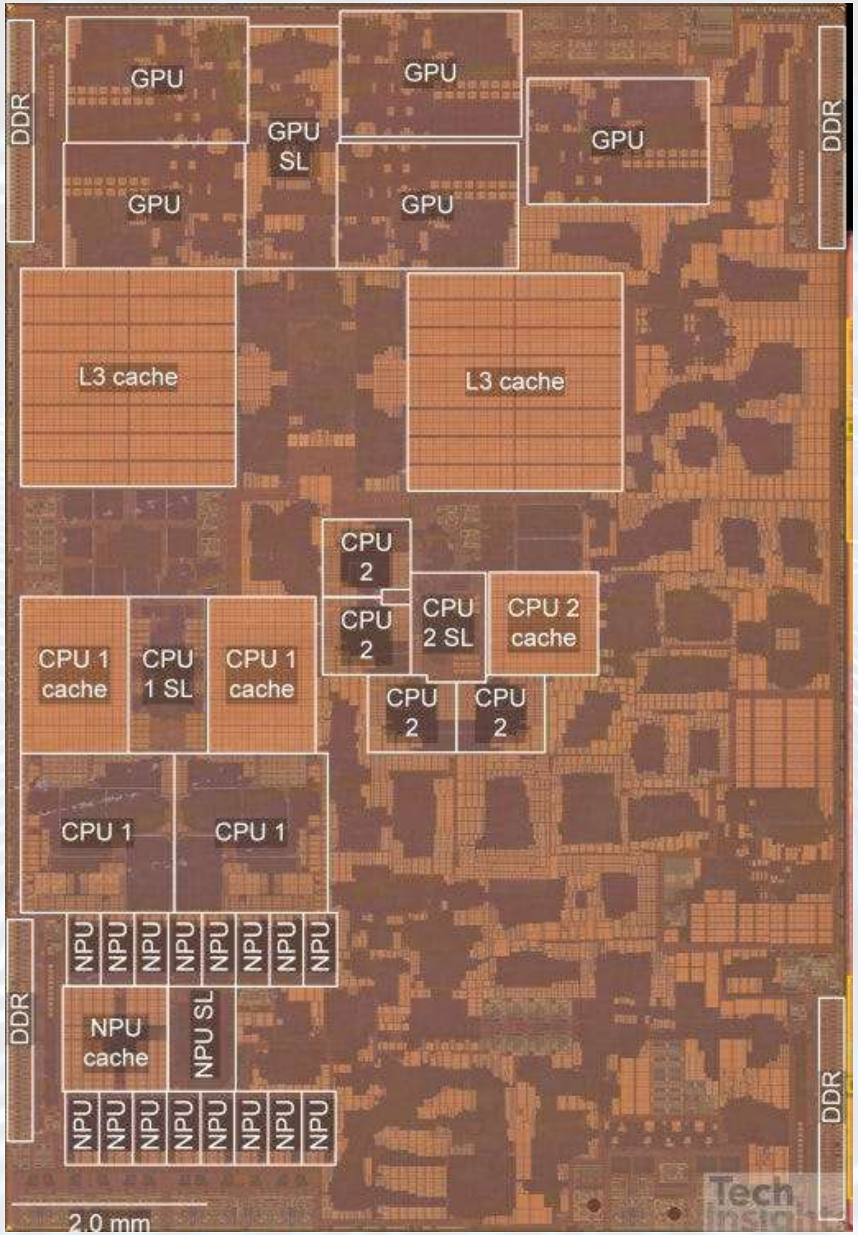
OurWorldinData.org – Research and data to make progress against the world’s largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Scaling della central processing unit (CPU)



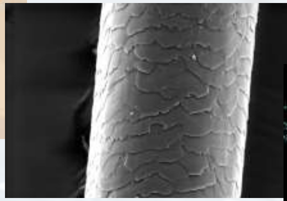
Intel 4004 (1971)
- N = 2300
- 12 mm^2
- f_{CK} = 0.74 MHz
- F = 10 μm


DDR
GPU
GPU
GPU SL
GPU
GPU
GPU
DDR
L3 cache
L3 cache
CPU 2
CPU 2
CPU 2 SL
CPU 2 cache
CPU 1 cache
CPU 1 SL
CPU 1 cache
CPU 2
CPU 2
CPU 1
CPU 1
NPU
NPU
NPU
NPU
NPU
NPU
NPU
NPU
DDR
NPU cache
NPU SL
NPU
NPU
NPU
NPU
NPU
NPU
NPU
NPU
DDR
2.0 mm
TechInsights

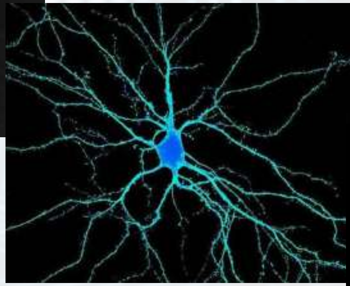
Apple A15 (2021)
- N = 15 miliardi
- 108 mm^2
- f_{CK} = 3.24 GHz
- F = 5 nm

Dal mm al nm

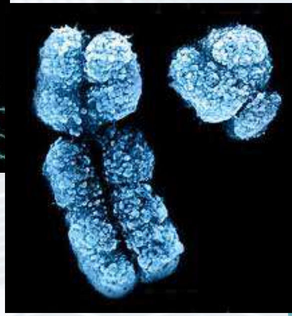
Pulce
(1 mm)



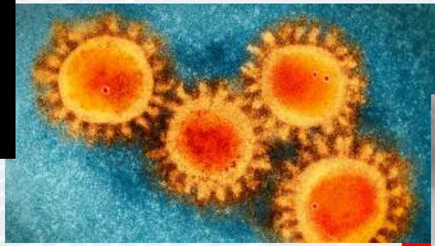
Capello
(50 μm)



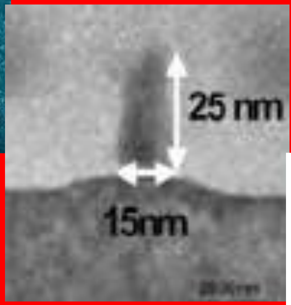
Neurone
(10 μm)



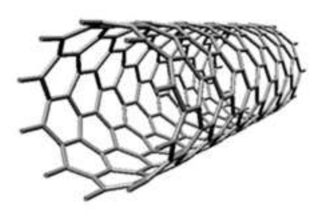
Cromosoma
(1 μm)



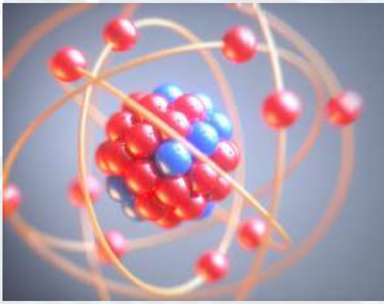
Virus COVID-19
(100 nm)



Transistor
(10 nm)

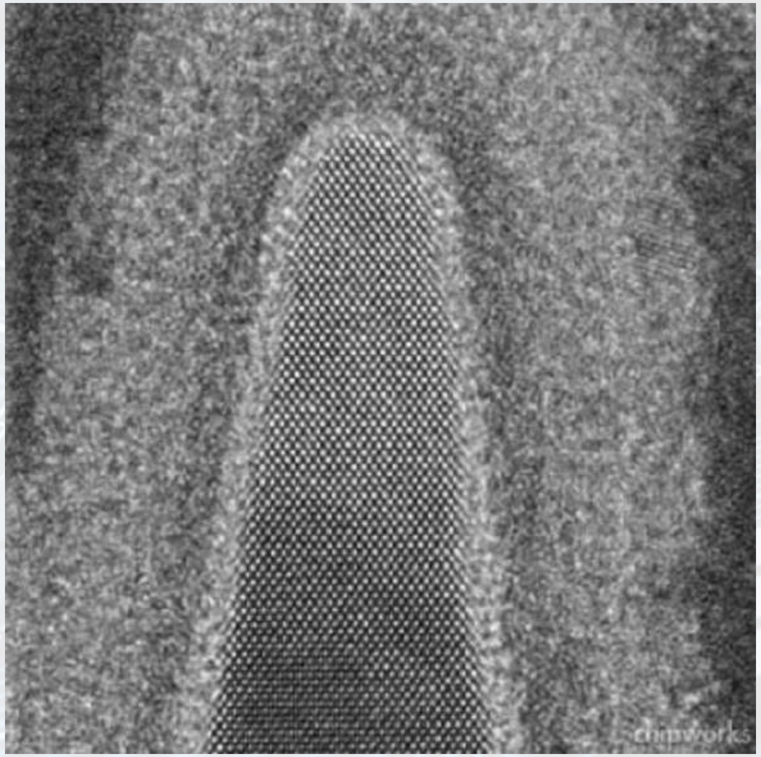


Carbon
nanotube
(1 nm)

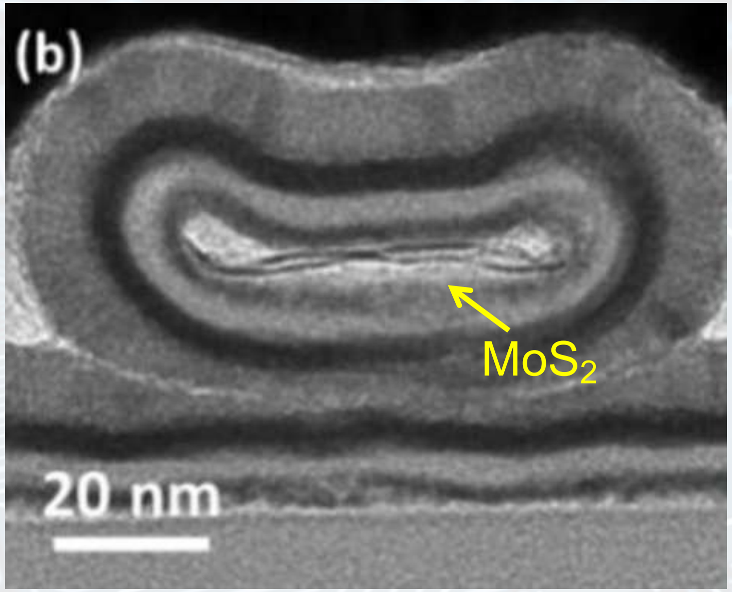


Atomo
(0.3 nm)

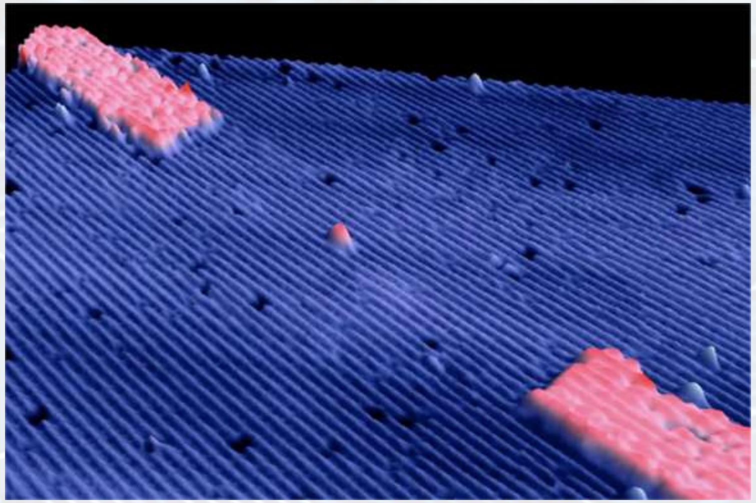
Scaling del transistor fino al limite del singolo atomo



30 strati atomici (finFET)



1 strato atomico (MoS$_2$)

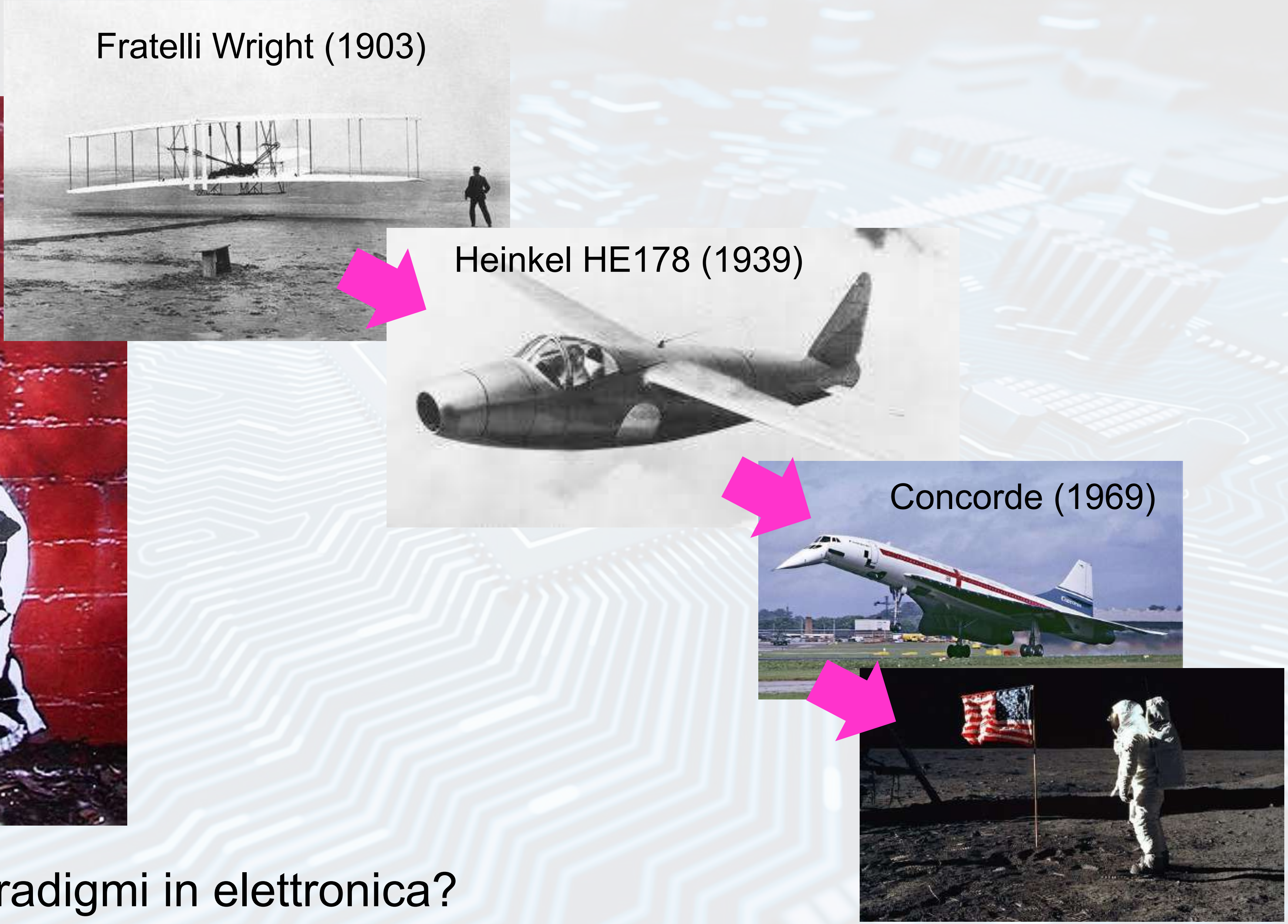


1 atomo (P:Si)

- Sono stati dimostrati transistori a singolo strato atomico o addirittura a singolo atomo
- La legge di Moore prima o poi fallirà!

Un cambio di paradigma

KEEP YOUR COINS,
I WANT CHANGE


Fratelli Wright (1903)

Heinkel HE178 (1939)

Concorde (1969)

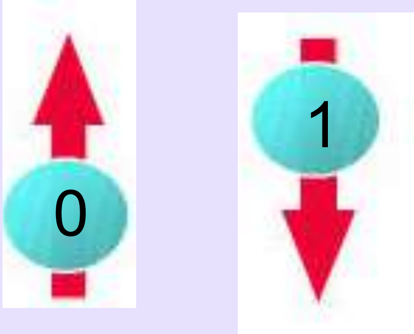
- Quali sono i nuovi paradigmi in elettronica?

Sommario

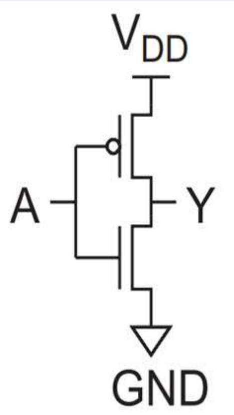
- Introduzione
- Breve storia dell’elettronica
- Evoluzione dell’elettronica: lo scaling
- L’elettronica del futuro: il quantum computing
- Intelligenza artificiale
- In-memory computing
- Conclusioni

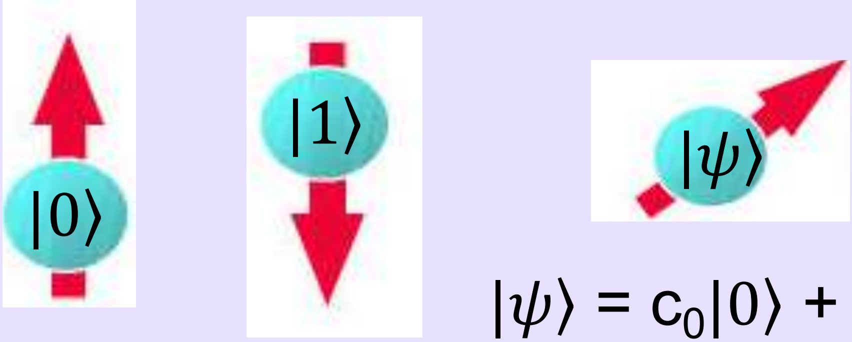
daniele.ielmini@polimi.it

Quantum computing



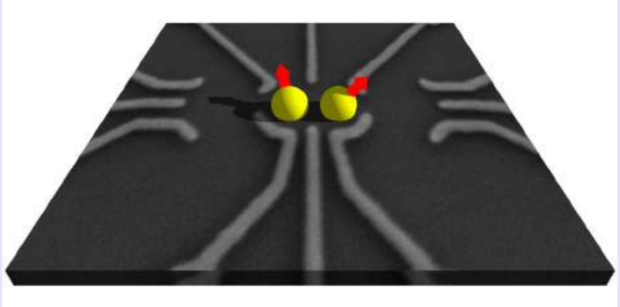
- **Bit**: mattoncino di informazione dell’algebra Booleana
- Può essere 0 o 1
- Rappresentato dal potenziale al nodo Y nell’inverter CMOS





$|\psi\rangle = c_0|0\rangle + c_1|1\rangle$

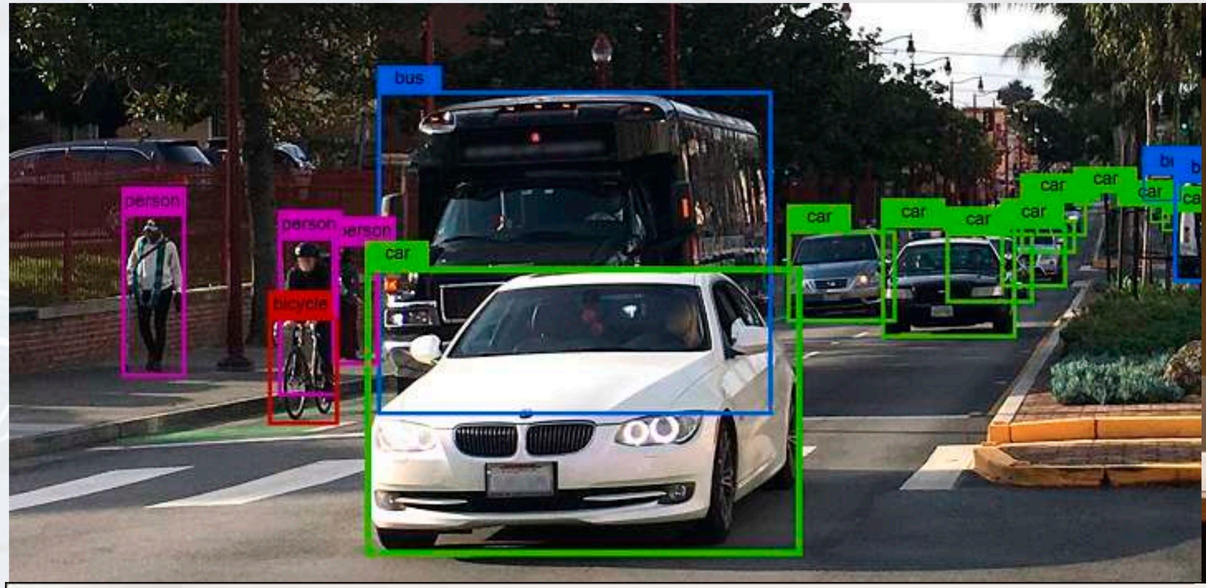
- **Qubit** (quantum bit)= mattoncino di informazione $|\psi>$ nel quantum computing (QC)
- Si trova in una **sovrapposizione** di 0 e 1
- Rappresentato dallo spin di un elettrone

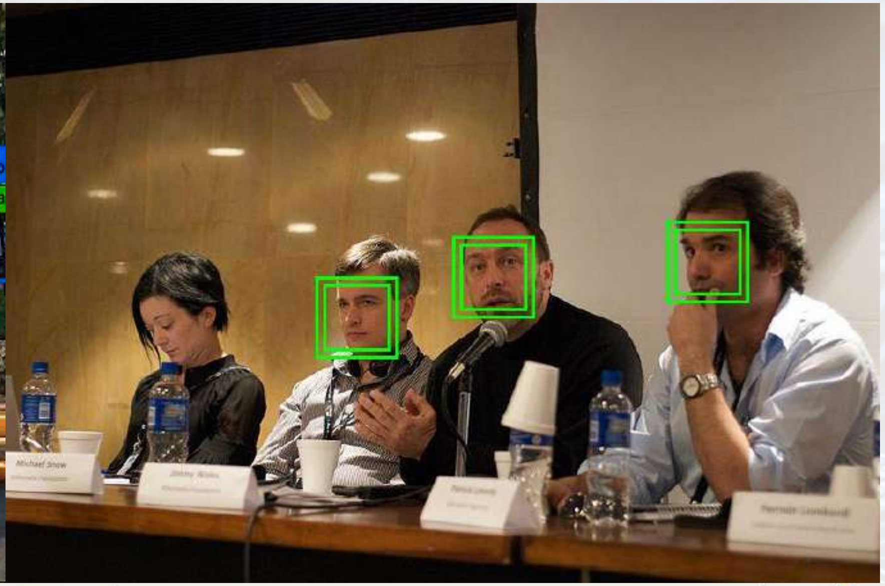


Sommario

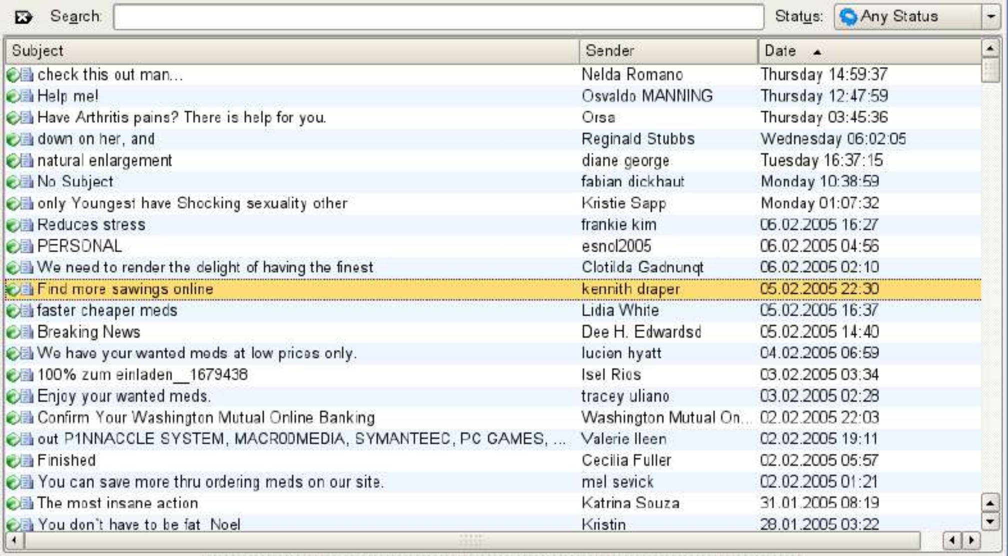
- Introduzione
- Breve storia dell’elettronica
- Evoluzione dell’elettronica: lo scaling
- L’elettronica del futuro: il quantum computing
- Intelligenza artificiale
- In-memory computing
- Conclusioni

Riconoscimento di oggetti, facce, messaggi vocali …



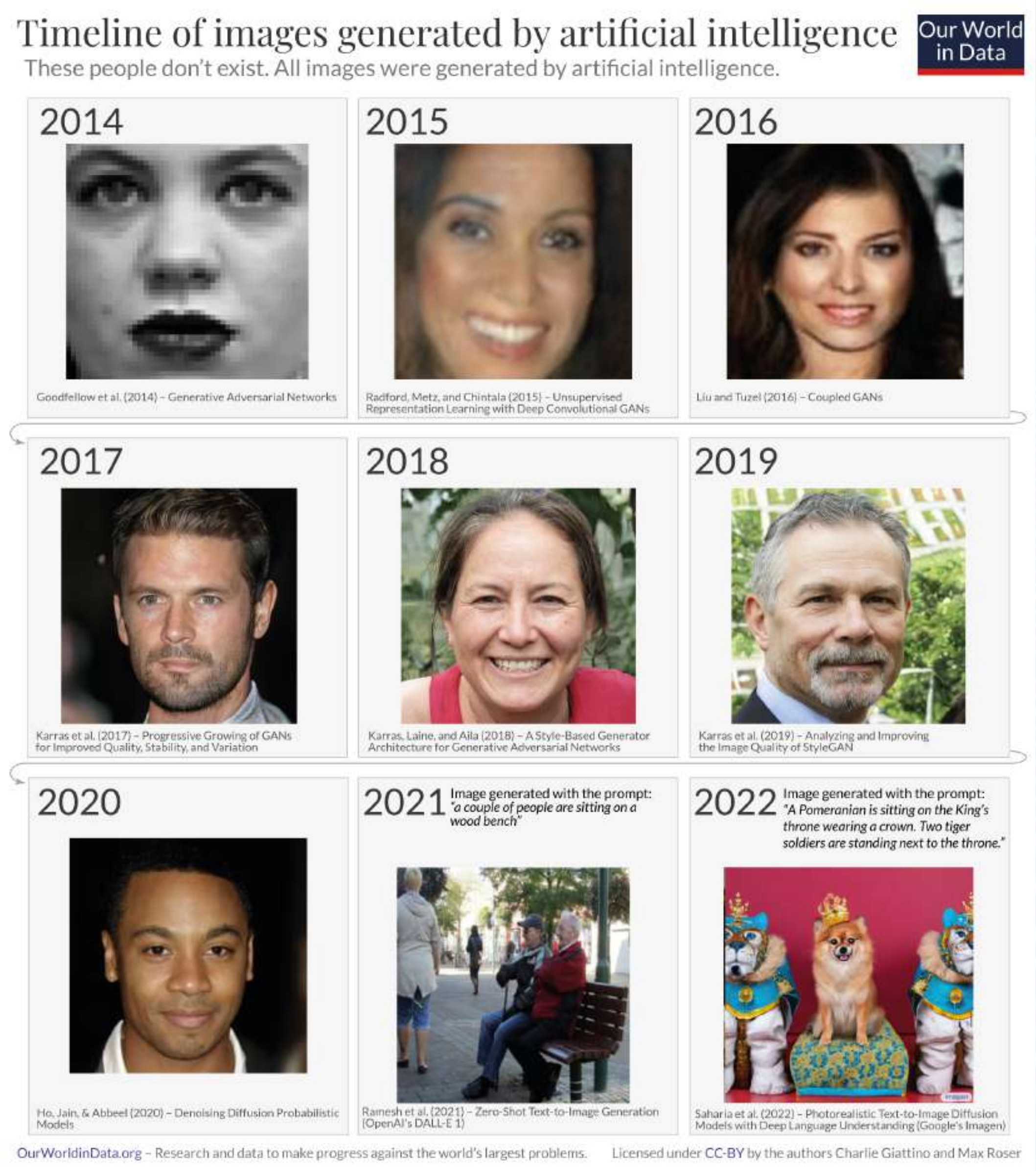


Did you say Alexa?



Subject	Sender	Date ▲
check this out man...	Nelda Romano	Thursday 14:59:37
Help me!	Osvaldo MANNING	Thursday 12:47:59
Have Arthritis pains? There is help for you.	Orsa	Thursday 03:45:36
down on her, and	Reginald Stubbs	Wednesday 06:02:05
natural enlargement	diane george	Tuesday 16:37:15
No Subject	fabian dickhaut	Monday 10:38:59
only Youngest have Shocking sexuality other	Kristie Sapp	Monday 01:07:32
Reduces stress	frankie kim	06.02.2005 16:27
PERSONAL	esnol2005	06.02.2005 04:56
We need to render the delight of having the finest	Clotilda Gadnunqt	06.02.2005 02:10
Find more sawings online	kennith draper	05.02.2005 22:30
faster cheaper meds	Lidia White	05.02.2005 16:37
Breaking News	Dee H. Edwardsd	05.02.2005 14:40
We have your wanted meds at low prices only.	lucien hyatt	04.02.2005 06:59
100% zum einladen__1679438	Isel Rios	03.02.2005 03:34
Enjoy your wanted meds.	tracey uliano	03.02.2005 02:28
Confirm Your Washington Mutual Online Banking	Washington Mutual On...	02.02.2005 22:03
out P1NNACCLE SYSTEM, MACR0DMEDIA, SYMANTEEC, PC GAMES, ...	Valerie Ileen	02.02.2005 19:11
Finished	Cecilia Fuller	02.02.2005 05:57
You can save more thru ordering meds on our site.	mel sevick	02.02.2005 01:21
The most insane action	Katrina Souza	31.01.2005 08:19
You don't have to be fat Noel	Kristin	28.01.2005 03:22

Generazione di facce, opere d’arte, testo …



Timeline of images generated by artificial intelligence

These people don’t exist. All images were generated by artificial intelligence.

2014 – Goodfellow et al. (2014) – Generative Adversarial Networks

2015 – Radford, Metz, and Chintala (2015) – Unsupervised Representation Learning with Deep Convolutional GANs

2016 – Liu and Tuzel (2016) – Coupled GANs

2017 – Karras et al. (2017) – Progressive Growing of GANs for Improved Quality, Stability, and Variation

2018 – Karras, Laine, and Aila (2018) – A Style-Based Generator Architecture for Generative Adversarial Networks

2019 – Karras et al. (2019) – Analyzing and Improving the Image Quality of StyleGAN

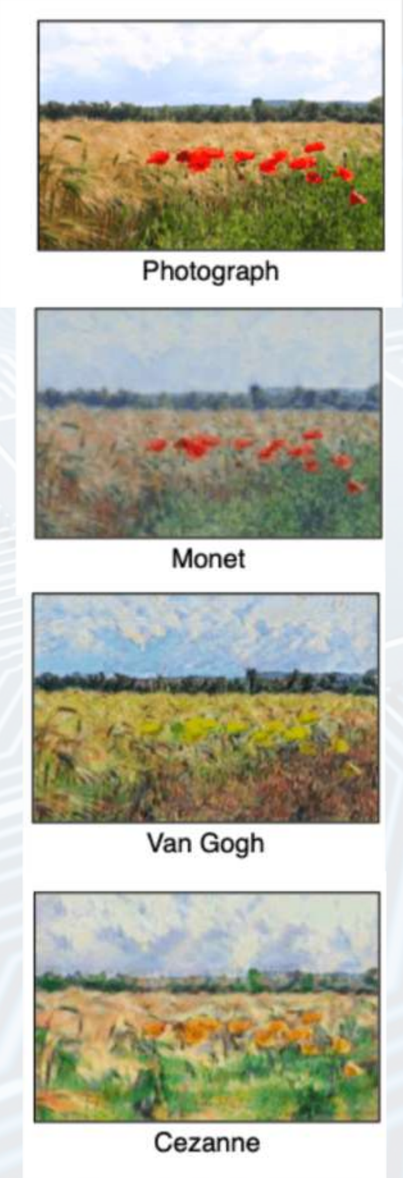
2020 – Ho, Jain, & Abbeel (2020) – Denoising Diffusion Probabilistic Models

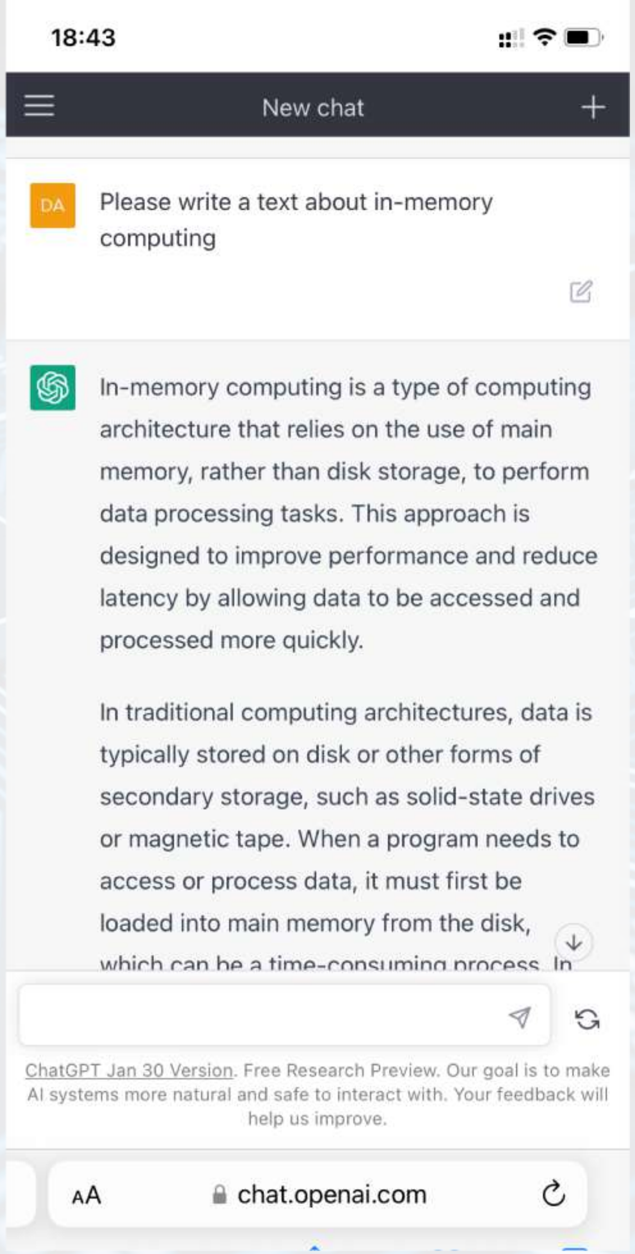
2021 Image generated with the prompt: *"a couple of people are sitting on a wood bench"* – Ramesh et al. (2021) – Zero-Shot Text-to-Image Generation (OpenAI's DALL-E 1)

2022 Image generated with the prompt: *"A Pomeranian is sitting on the King's throne wearing a crown. Two tiger soldiers are standing next to the throne."* – Saharia et al. (2022) – Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Google's Imagen)

OurWorldinData.org – Research and data to make progress against the world’s largest problems.

Licensed under CC-BY by the authors Charlie Giattino and Max Roser





Please write a text about in-memory computing

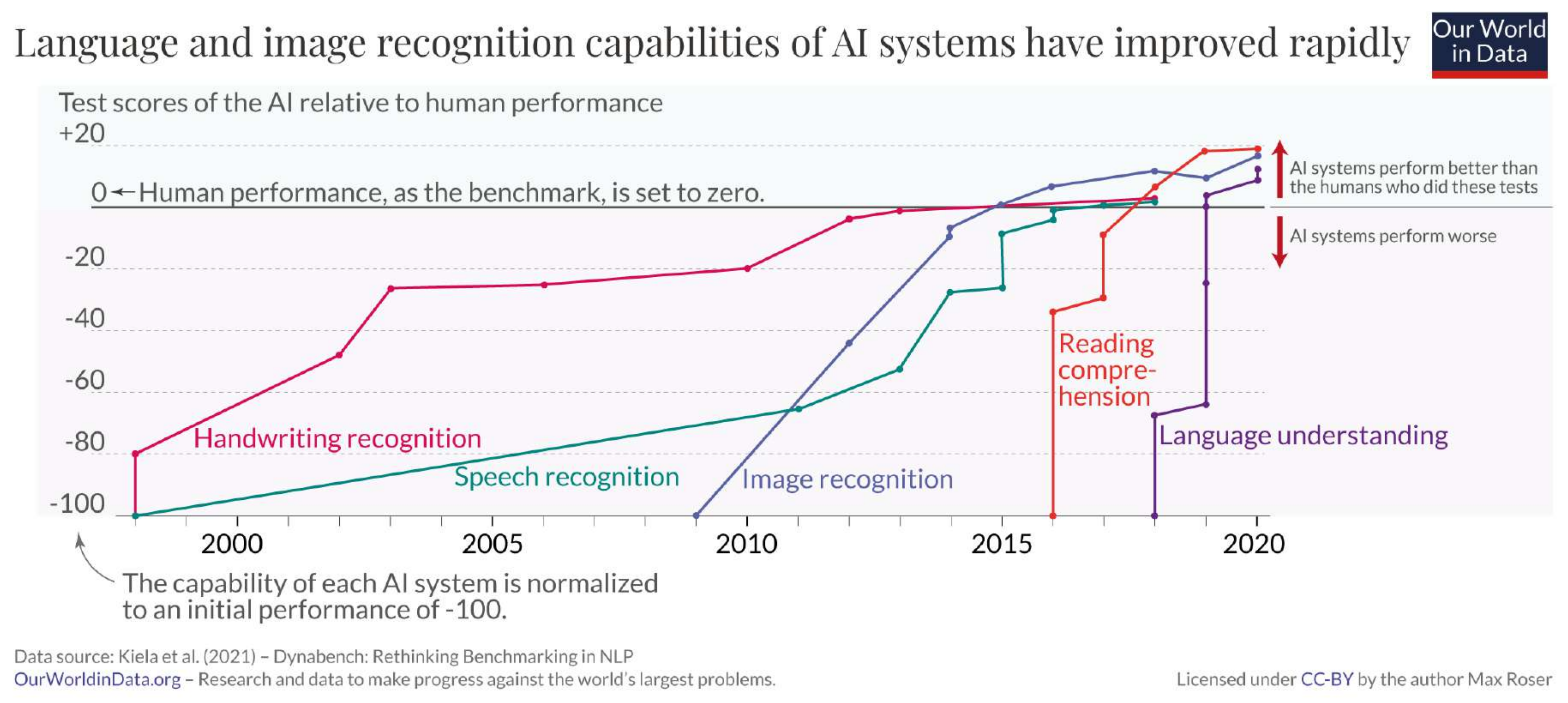
In-memory computing is a type of computing architecture that relies on the use of main memory, rather than disk storage, to perform data processing tasks. This approach is designed to improve performance and reduce latency by allowing data to be accessed and processed more quickly.

In traditional computing architectures, data is typically stored on disk or other forms of secondary storage, such as solid-state drives or magnetic tape. When a program needs to access or process data, it must first be loaded into main memory from the disk, which can be a time-consuming process. In

ChatGPT Jan 30 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

chat.openai.com

AI capabilities vs. the human brain

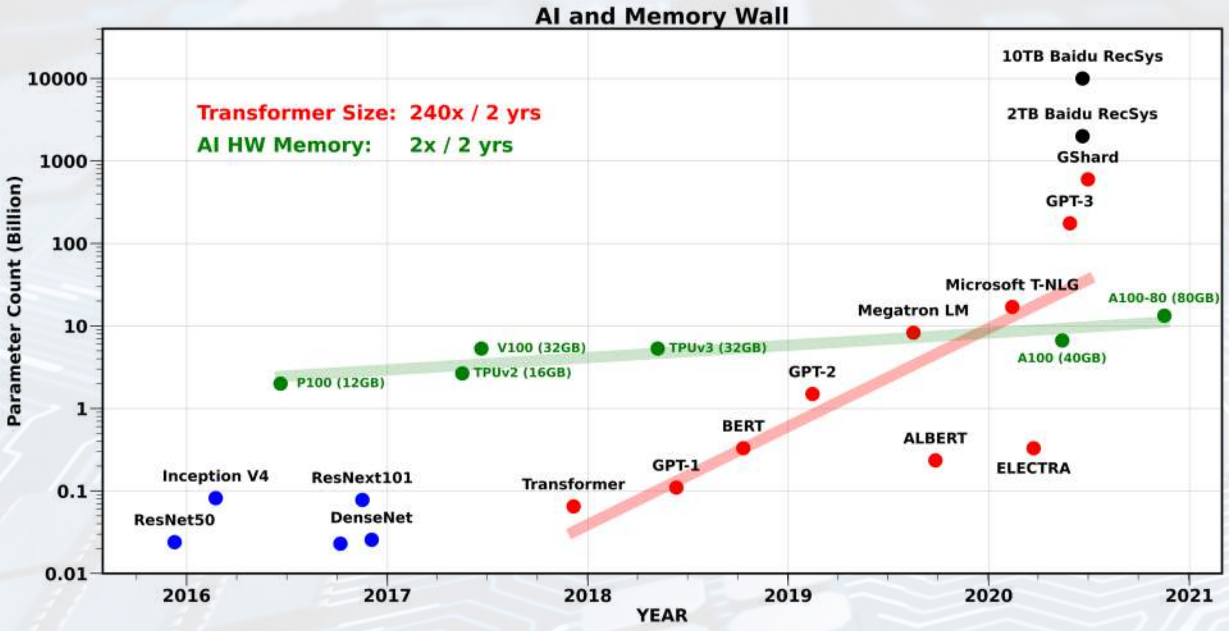


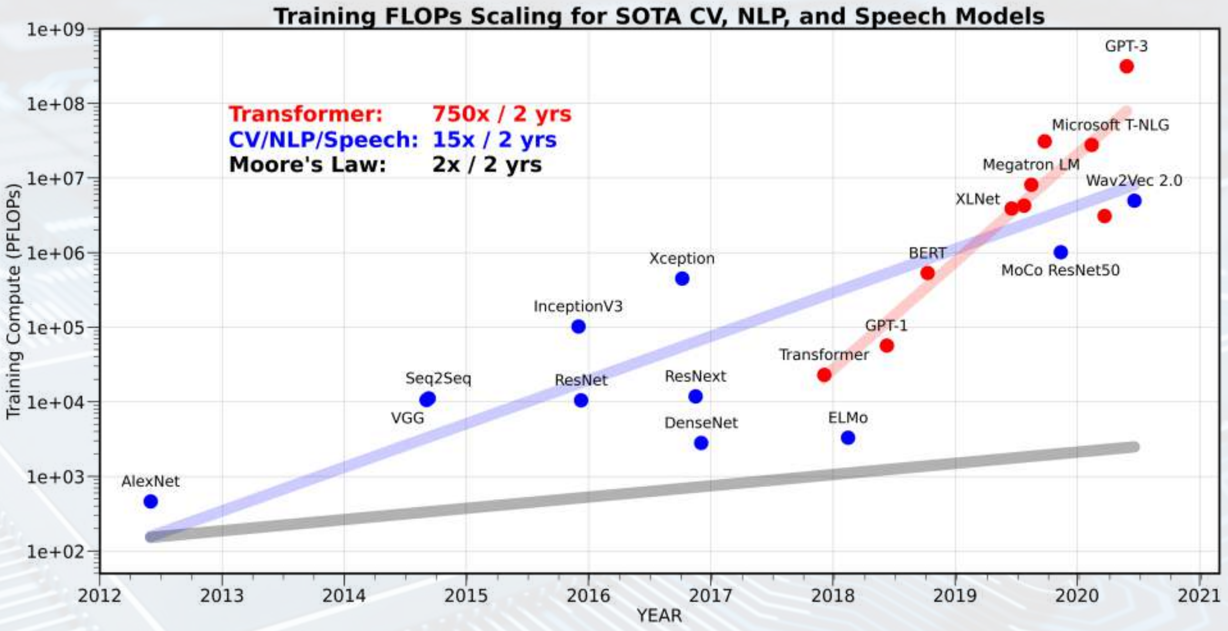
Language and image recognition capabilities of AI systems have improved rapidly

Data source: Kiela et al. (2021) – Dynabench: Rethinking Benchmarking in NLP
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the author Max Roser

AI scaling

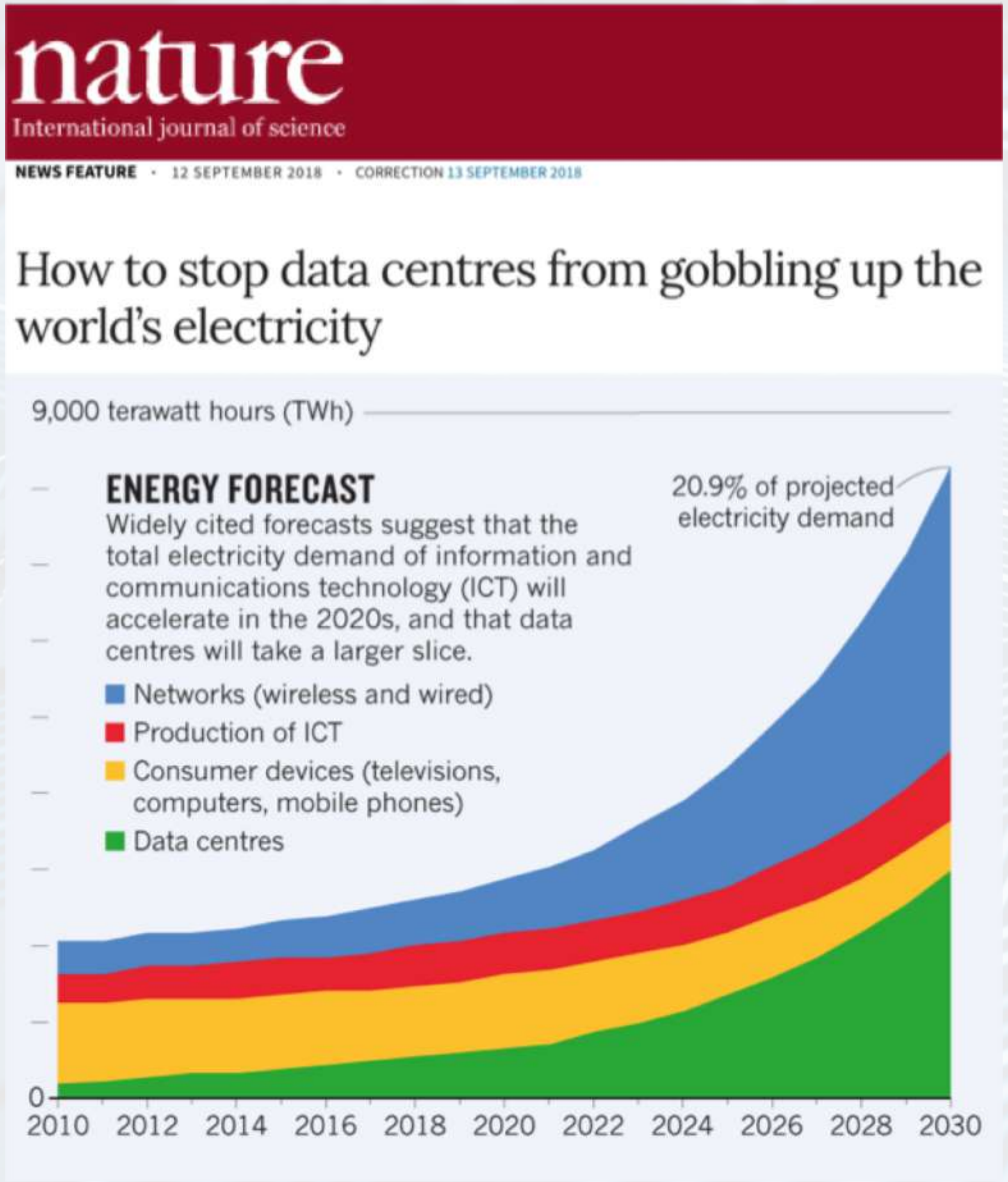




- Parameters = weights
- FLOPs = how many floating point operations are required to train a given model
- FLOPS = how many floating point operations can be executed per second in a given hardware, e.g., 82.6 TFLOPS for Nvidia's RTX 4090
- Note 1E9*1E15/1E14 = 1E10 s corresponding to 300 years

daniele.ielmini@polimi.it

Il carbon footprint dell’intelligenza artificiale



nature
International journal of science

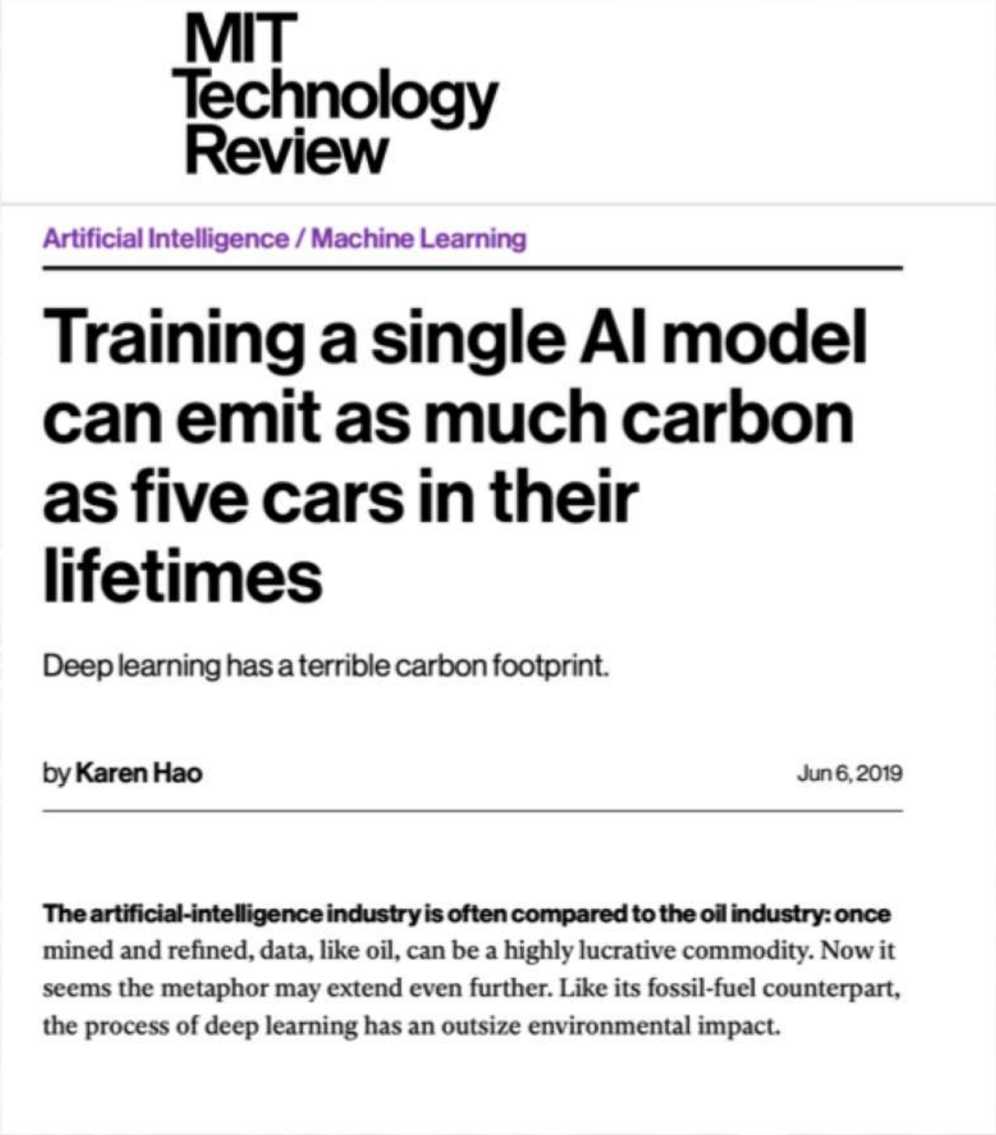
NEWS FEATURE · 12 SEPTEMBER 2018 · CORRECTION 13 SEPTEMBER 2018

How to stop data centres from gobbling up the world’s electricity

ENERGY FORECAST
Widely cited forecasts suggest that the total electricity demand of information and communications technology (ICT) will accelerate in the 2020s, and that data centres will take a larger slice.

- 🟦 Networks (wireless and wired)
- 🟥 Production of ICT
- 🟨 Consumer devices (televisions, computers, mobile phones)
- 🟩 Data centres

9,000 terawatt hours (TWh)
20.9% of projected electricity demand
0
2010
2012
2014
2016
2018
2020
2022
2024
2026
2028
2030



MIT Technology Review

Artificial Intelligence / Machine Learning

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

by **Karen Hao** Jun 6, 2019

The artificial-intelligence industry is often compared to the oil industry: once mined and refined, data, like oil, can be a highly lucrative commodity. Now it seems the metaphor may extend even further. Like its fossil-fuel counterpart, the process of deep learning has an outsize environmental impact.

AlphaGo vs. Lee Sedol

ALPHAGO
00:00:49
AlphaGo
Google DeepMind
LEE SEDOL
00:01:00

20 W
1250 cm^3
1.5 kg
10^{11} neurons
10^{15} synapses

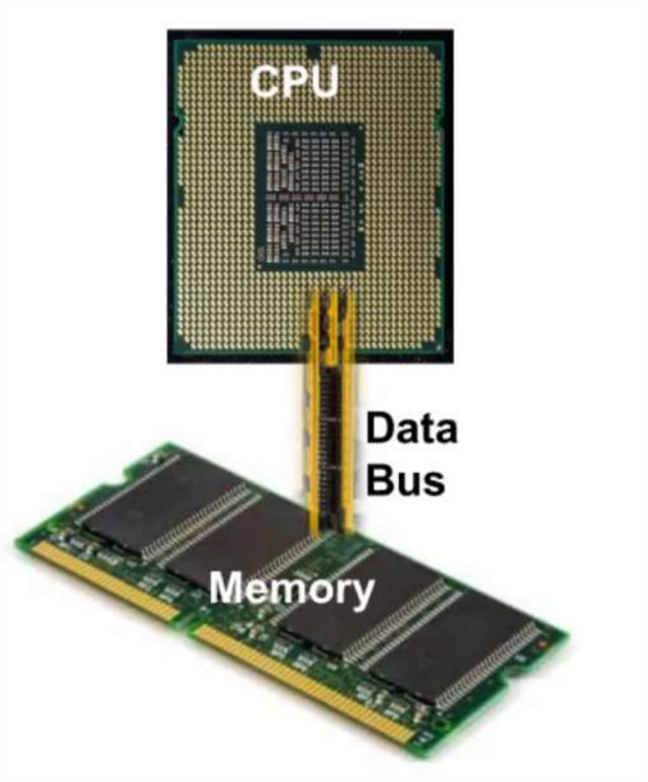
- 170 kW
- 1200 CPUs
- 176 GPU
- 30 x 10^{12} op/s

AlphaGo ha battuto il campione Lee Sedol, ma lo scontro è stato impari …

Architettura di von Neumann e architettura neuromorfica

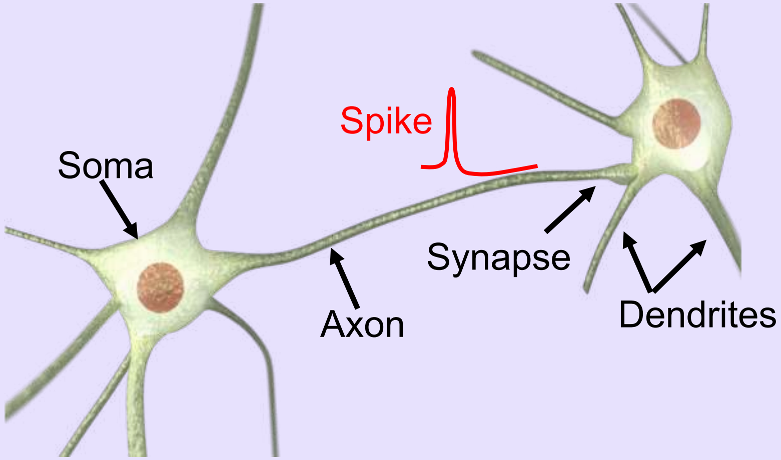
Architettura di Von Neumann

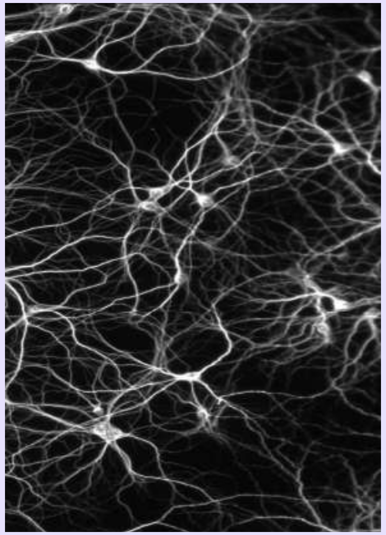
Paul von Neumann
(1903 – 1957)

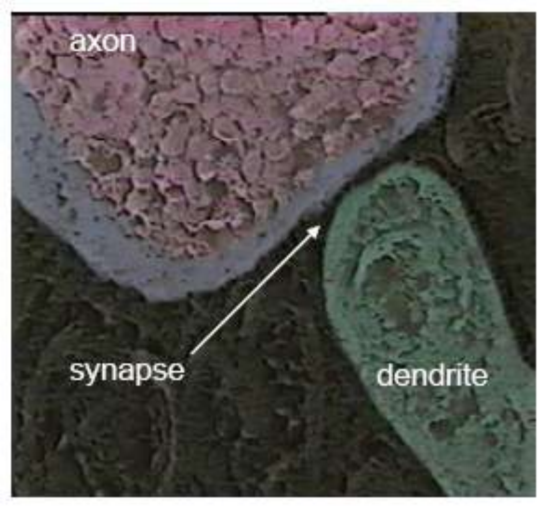

CPU
Data Bus
Memory

Il collo di bottiglia del calcolo è rappresentato non tanto dall’esecuzione delle operazioni, quanto dallo **spostamento dei dati**

Architettura neuromorfica

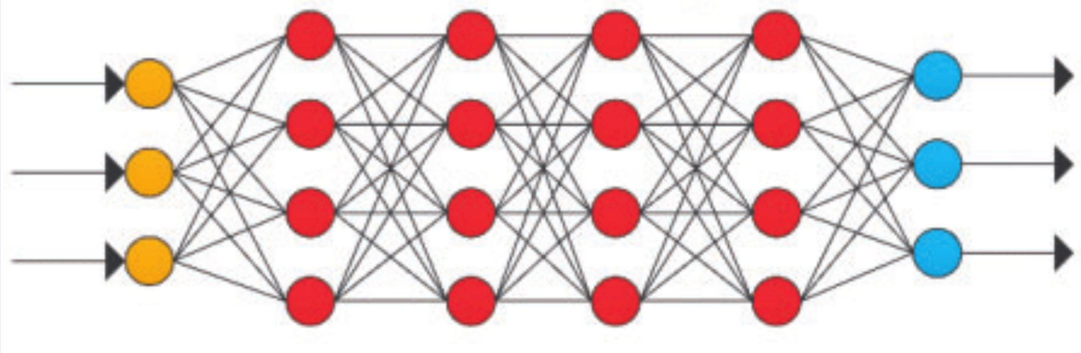

Spike
Soma
Synapse
Axon
Dendrites



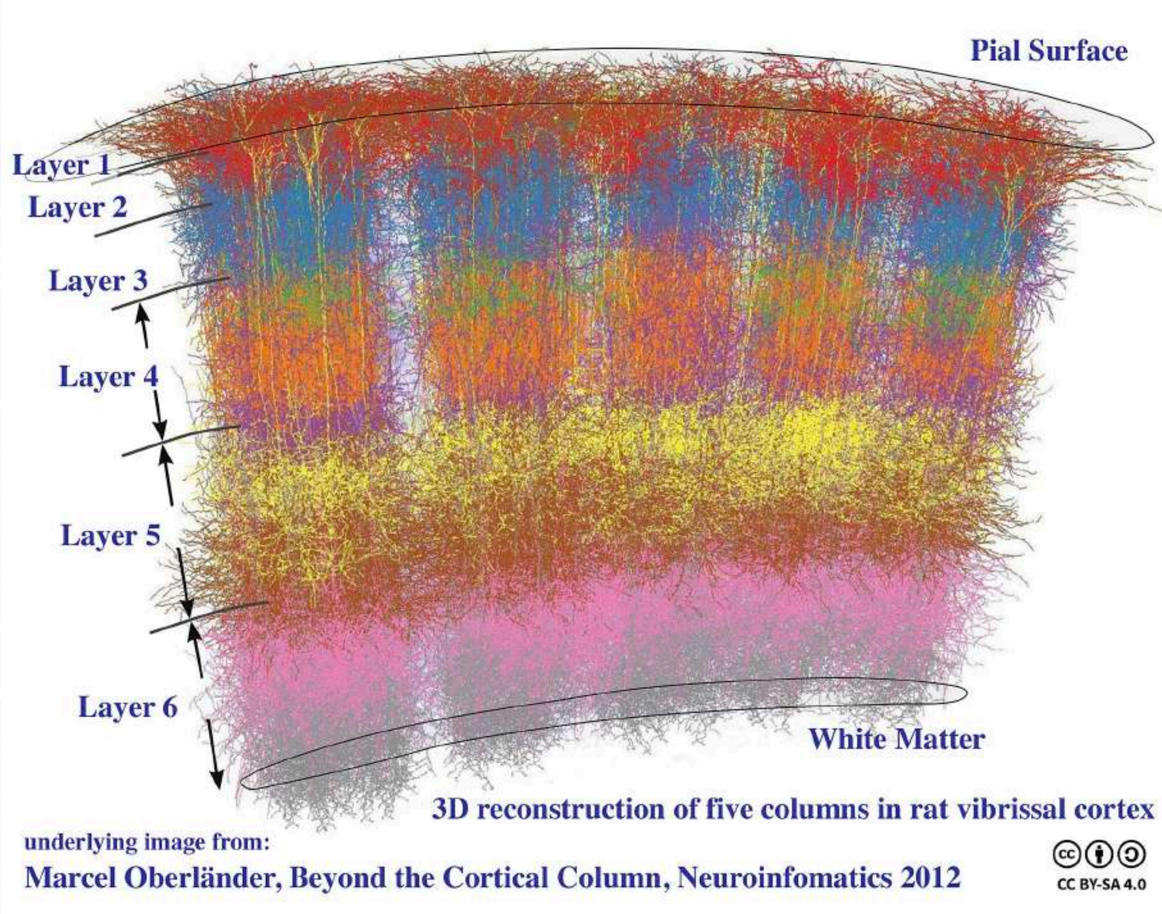

axon
synapse
dendrite

Le reti neurali

Frank Rosenblatt
(1928 – 1971)



Rete neurale profonda (deep learning)

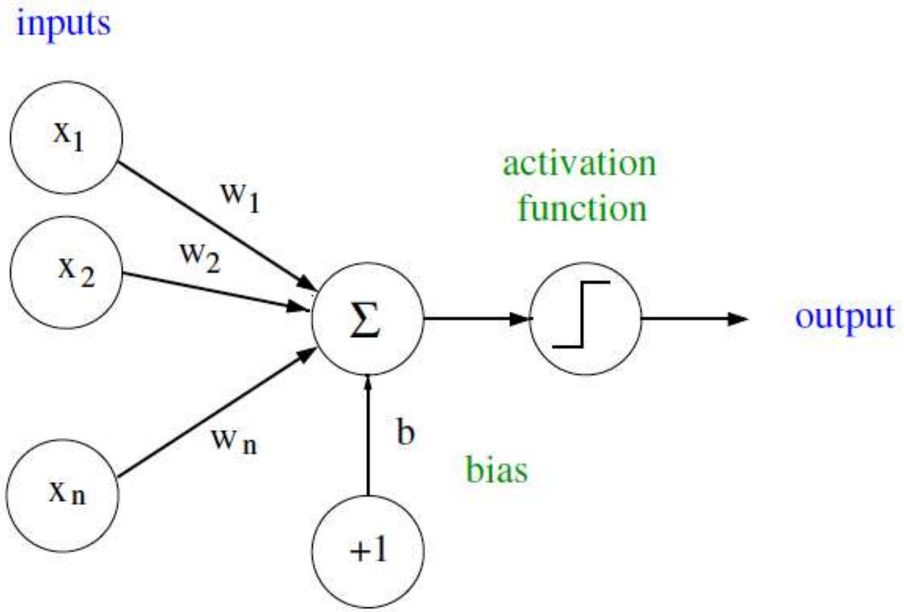

Pial Surface
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
White Matter
3D reconstruction of five columns in rat vibrissal cortex

underlying image from:
Marcel Oberländer, Beyond the Cortical Column, Neuroinfomatics 2012

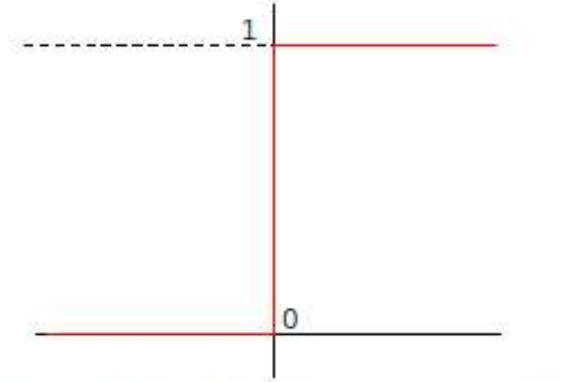
CC BY-SA 4.0

Strati di neuroni nella corteccia cerebrale

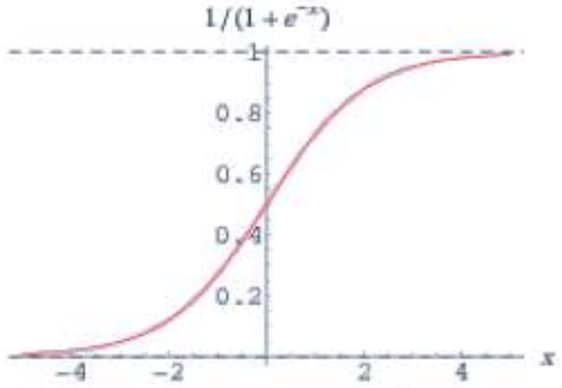
Somma, prodotto e attivazione non lineare


inputs
x_1
x_2
x_n
w_1
w_2
w_n
Σ
activation function
output
b
bias
+1

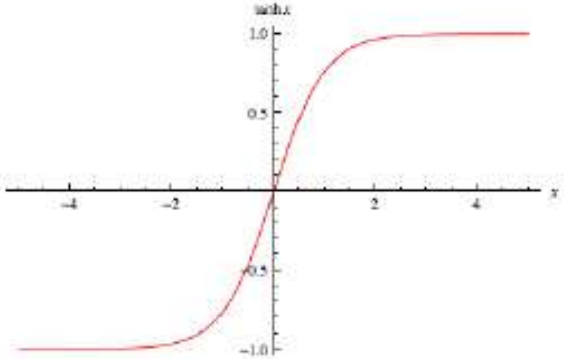
McCulloch and Pitts (1943)



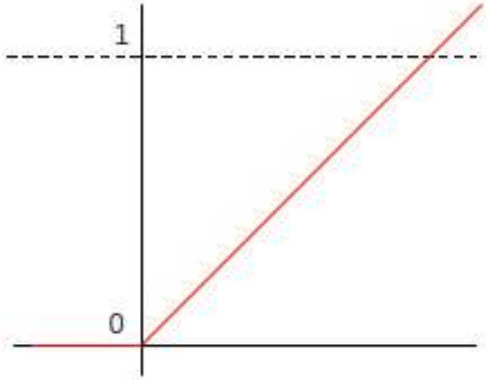
threshold: if $x > 0$ then 1 else 0



logistic function: $\frac{1}{1+e^{-x}}$

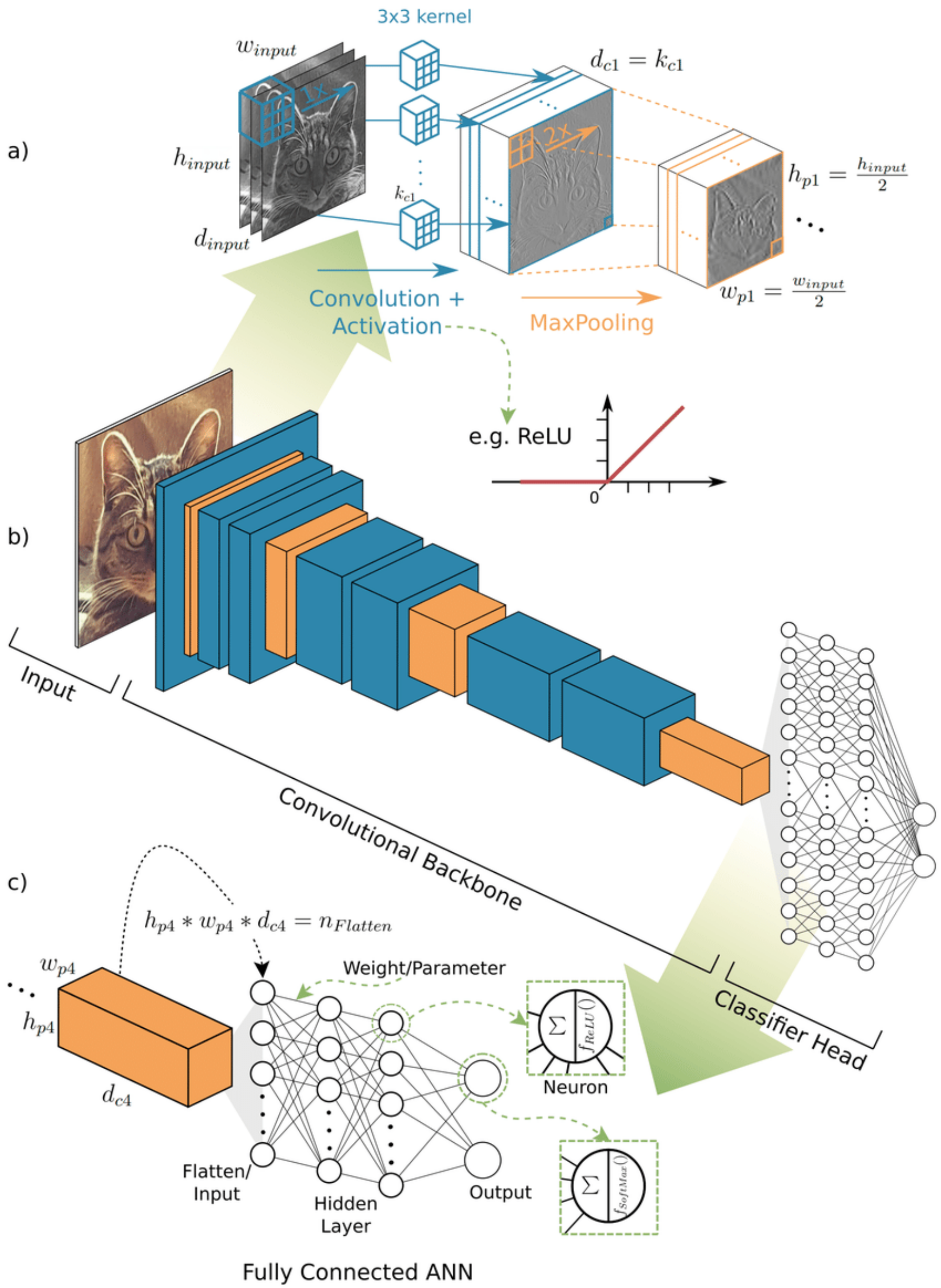


hyperbolic tangent: $\frac{e^x - e^{-x}}{e^x + e^{-x}}$



rectified linear (RELU): if $x > 0$ then x else 0

Una moderna rete neurale profonda (deep neural network)


a)
3x3 kernel
w_{input}
h_{input}
d_{input}
1x
k_{c1}
$d_{c1} = k_{c1}$
2x
$h_{p1} = \frac{h_{input}}{2}$
$w_{p1} = \frac{w_{input}}{2}$
Convolution +
Activation
MaxPooling
e.g. ReLU
0
b)
Input
Convolutional Backbone
Classifier Head
c)
$h_{p4} * w_{p4} * d_{c4} = n_{Flatten}$
w_{p4}
h_{p4}
d_{c4}
Weight/Parameter
Σ
$f_{ReLU}()$
Neuron
$f_{SoftMax}()$
Flatten/
Input
Hidden
Layer
Output
Fully Connected ANN

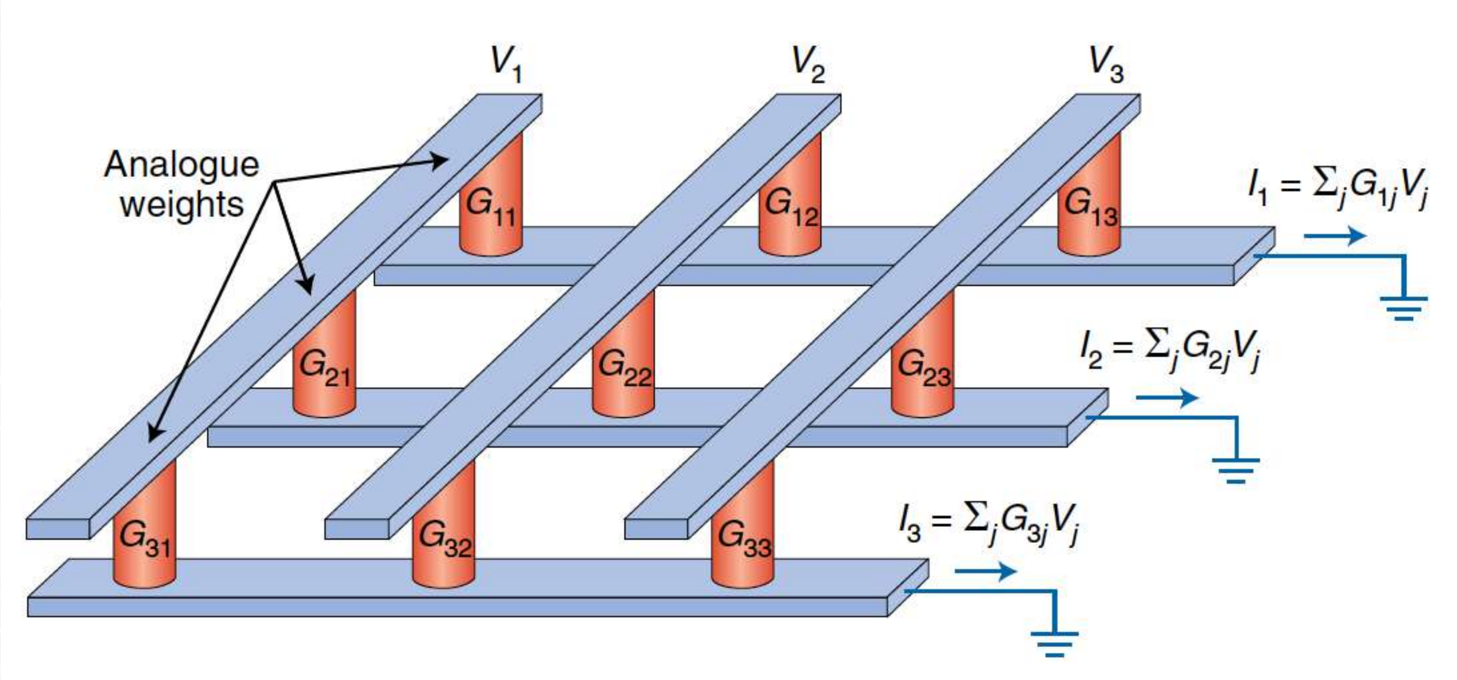
- Convolutional backbone = convolutional neural network (CNN) layers to capture features
- Classifier head = fully-connected neural network (FCNN) layers for the final classification
- Maxpooling = take the maximum within a NxN subregion

daniele.ielmini@polimi.it

Sommario

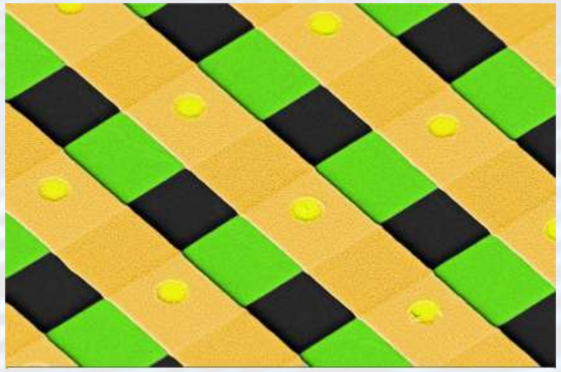
- Introduzione
- Breve storia dell’elettronica
- Evoluzione dell’elettronica: lo scaling
- L’elettronica del futuro: il quantum computing
- Intelligenza artificiale
- In-memory computing
- Conclusioni

Prodotto matrice-vettore con leggi fisiche



V_1
V_2
V_3
Analogue weights
G_{11}
G_{12}
G_{13}
G_{21}
G_{22}
G_{23}
G_{31}
G_{32}
G_{33}
$I_1 = \Sigma_j G_{1j} V_j$
$I_2 = \Sigma_j G_{2j} V_j$
$I_3 = \Sigma_j G_{3j} V_j$

D. Ielmini and H.-S. P. Wong, Nature Electronics 1, 333 (2018)

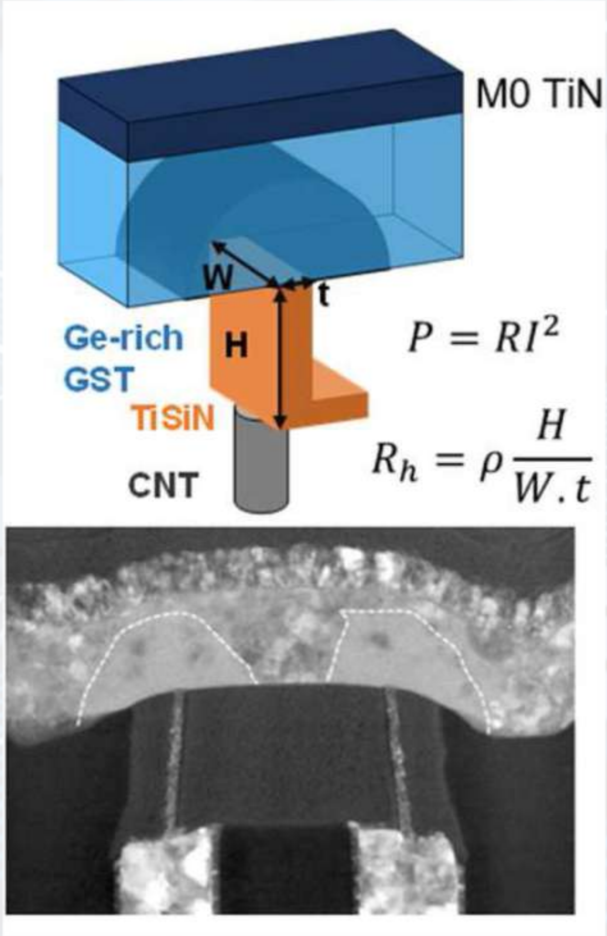


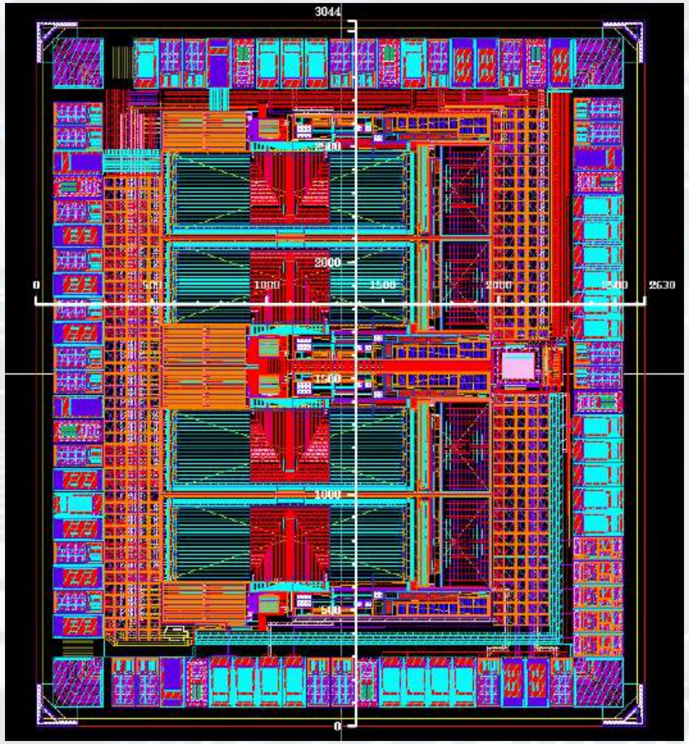
Kirchhoff’s law

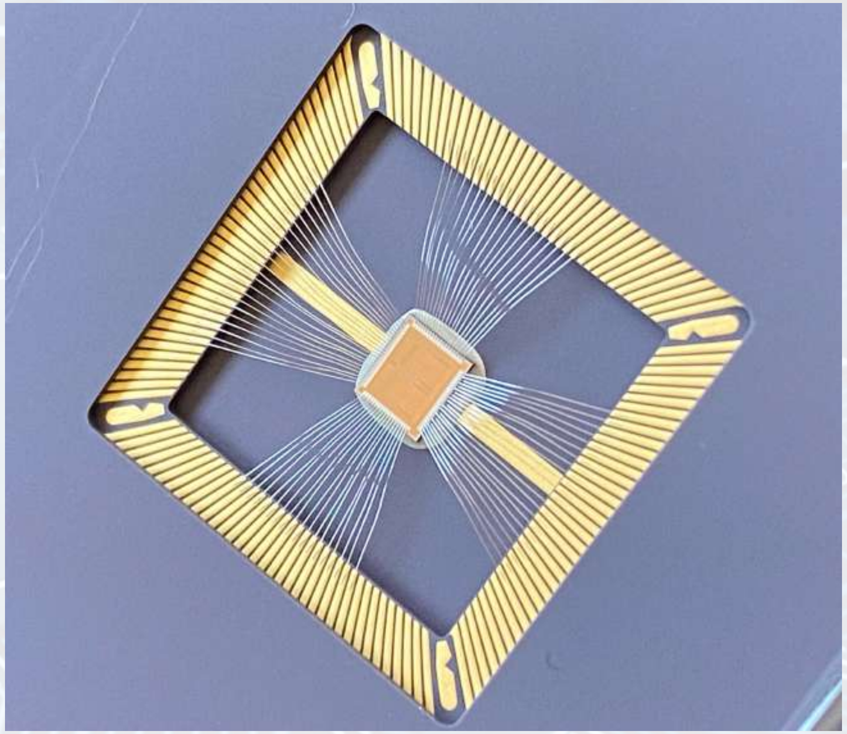
$$I_j = \sum_i G_{ij} V_i$$

Ohm’s law

Design di circuiti integrati per gli acceleratori AI


M0 TiN
W
t
Ge-rich GST
H
TiSiN
CNT
$P = RI^2$
$R_h = \rho \frac{H}{W.t}$


3044
0
500
1000
1500
2000
2500
2630

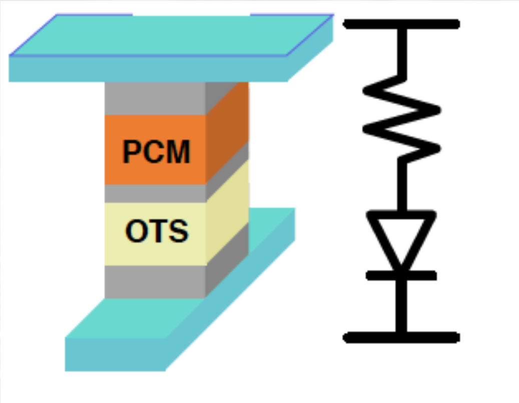


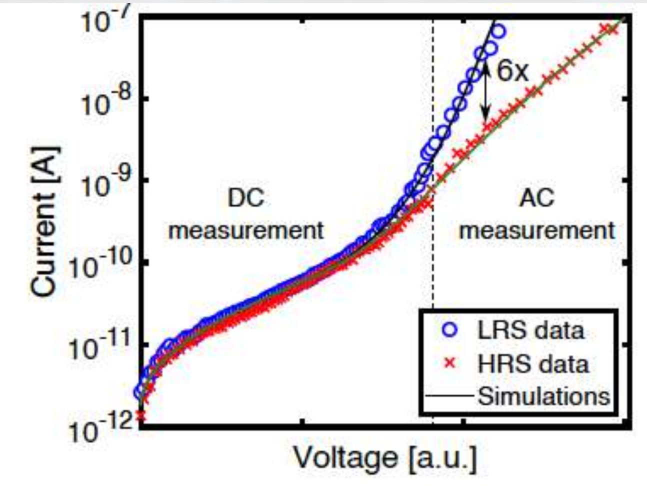
StorAIge

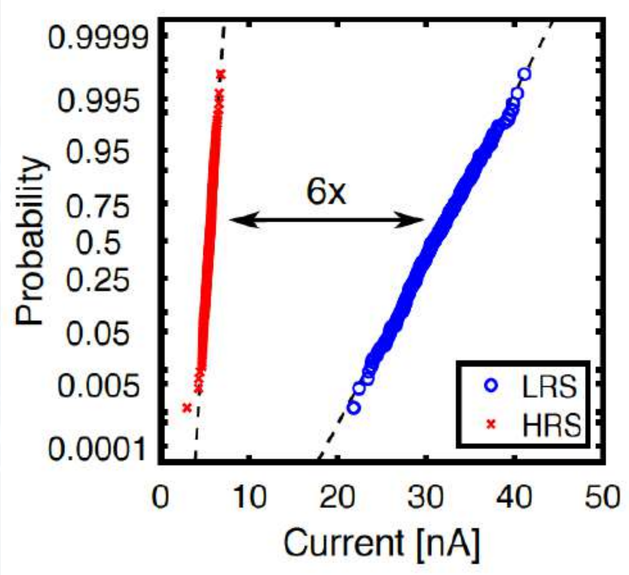
Horizon 2020
European Union Funding
for Research & Innovation

- The first chip was designed in CMOS 90 nm technology
- Currently we are designing a new chip in 28 nm FD-SOI technology

Nuovi dispositivi per gli acceleratori AI

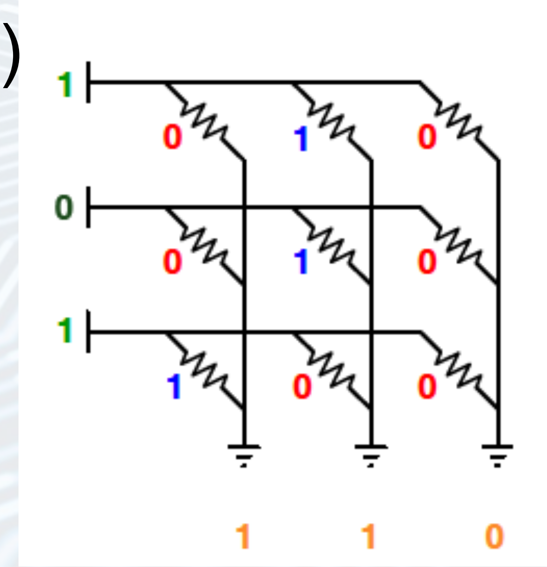


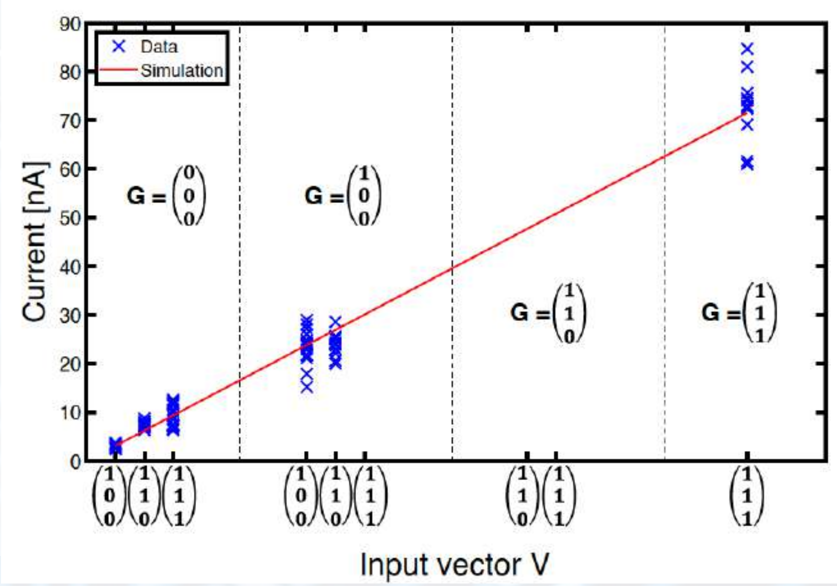




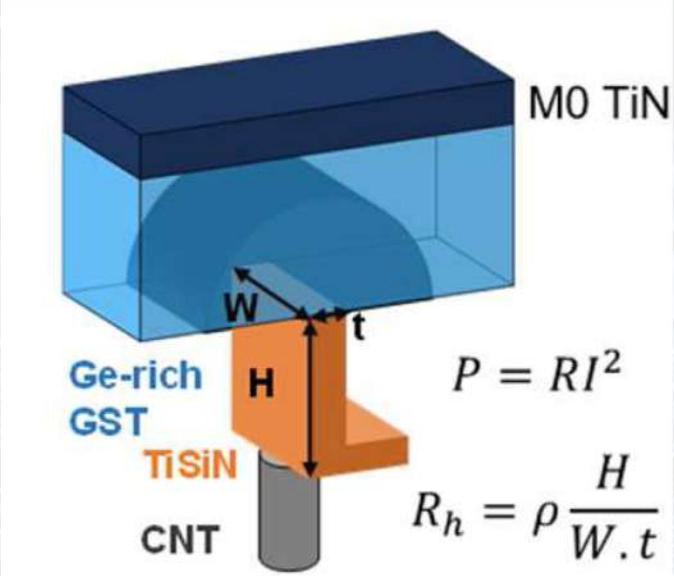
- 2-terminal structure (no transistor) has several advantages:
  - Less lines
  - Smaller footprint
  - Smaller current due to subthreshold operation
  - Higher accuracy due to lower IR drop

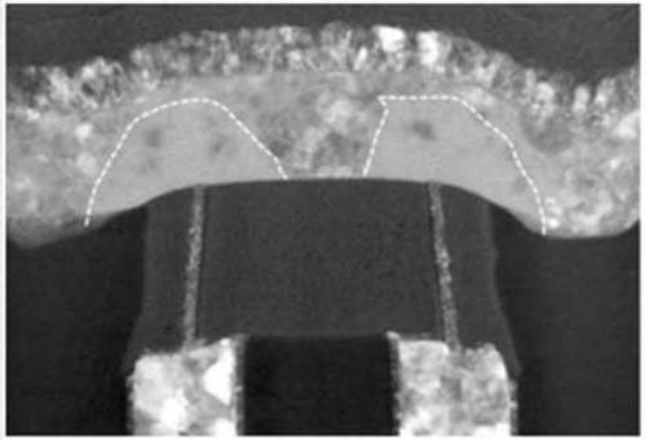
N. Lepri, et al., IMW (2023)

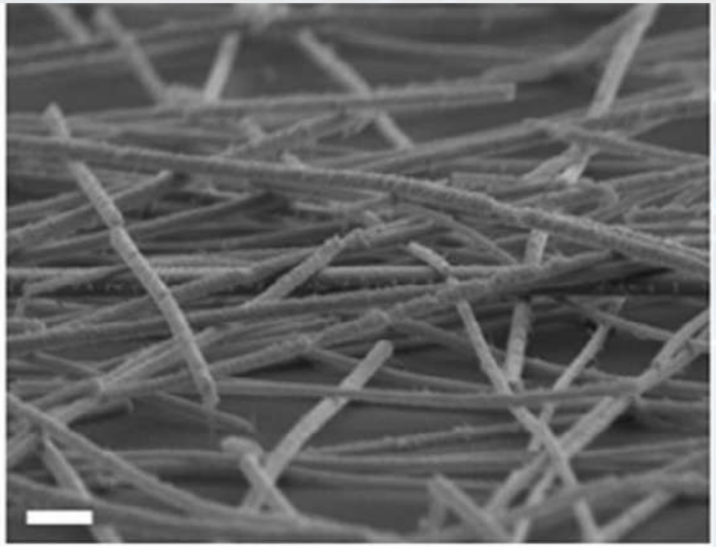


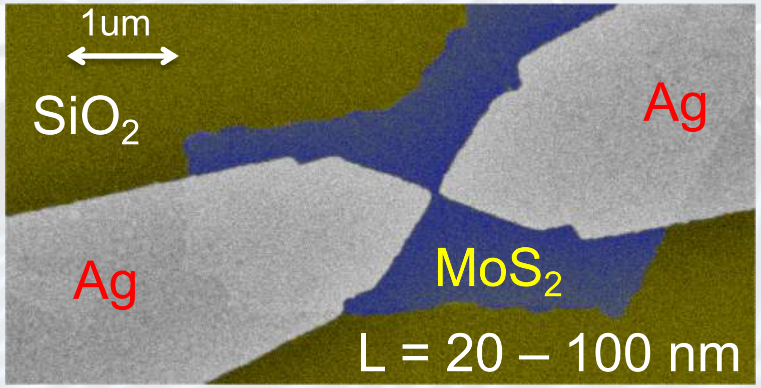


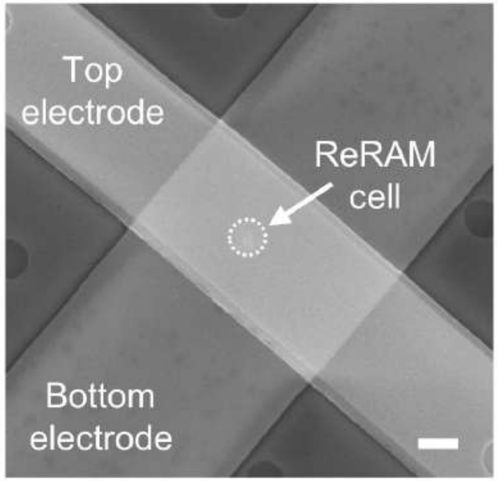
Hardware neuromorfico: dispositivi

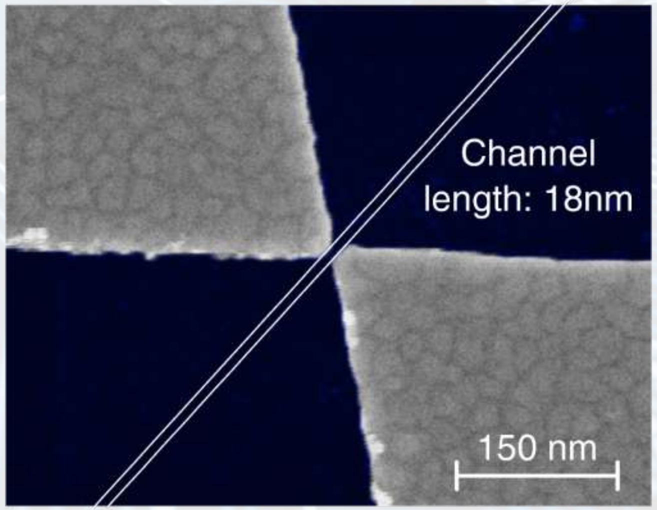

M0 TiN
W
t
H
Ge-rich
GST
TiSiN
CNT
$P = RI^2$
$R_h = \rho \frac{H}{W.t}$






1um
SiO_2
Ag
Ag
MoS_2
L = 20 – 100 nm

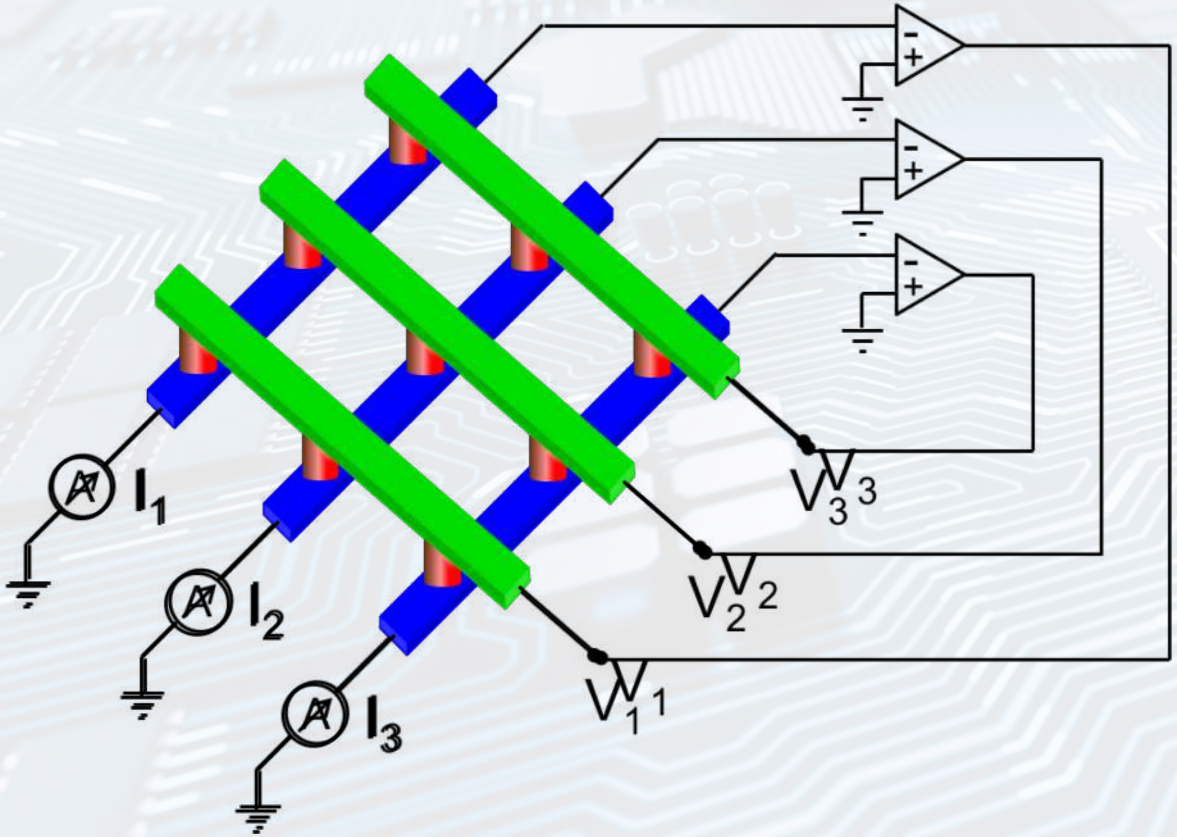

Top
electrode
ReRAM
cell
Bottom
electrode

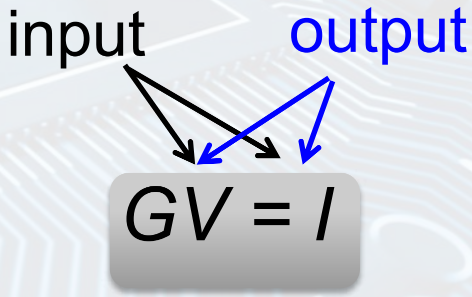

Channel
length: 18nm
150 nm

Nuovi dispositivi per riprodurre fisicamente i processi cognitivi

polifab
POLITECNICO DI MILANO

In-memory computing for matrix inversion


I_1
I_2
I_3
V_1
V_2
V_3
−
+

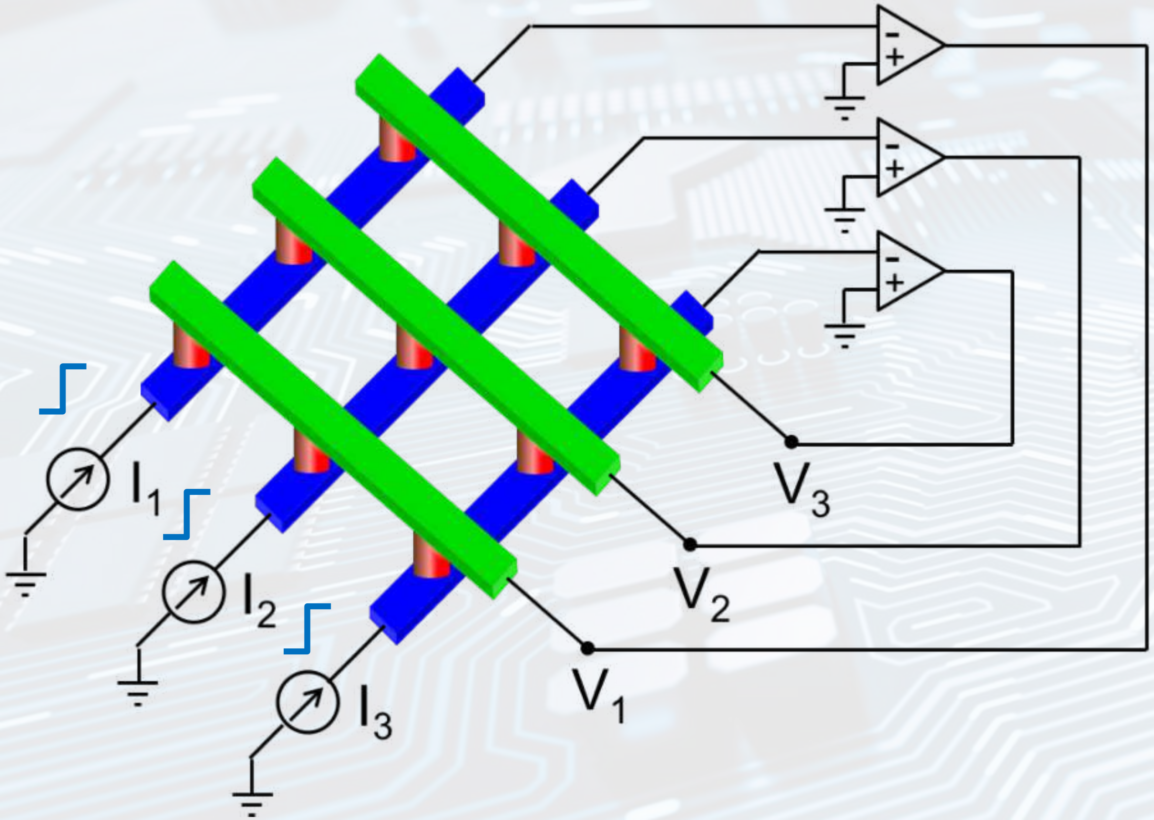

input
output

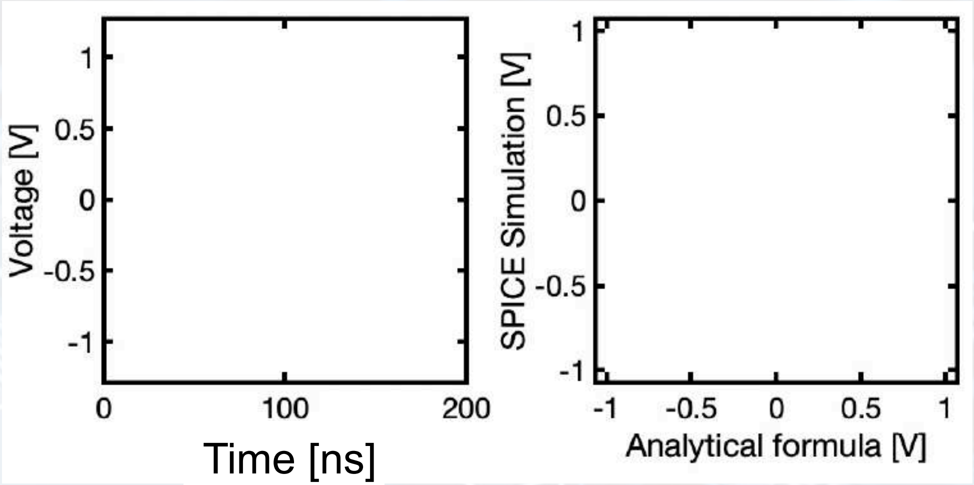
$GV = I$

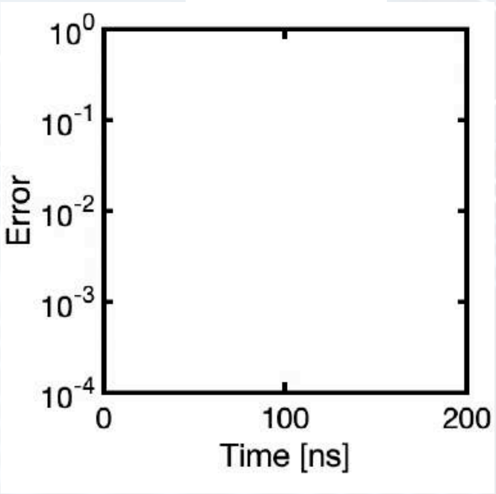
$Ax = b$

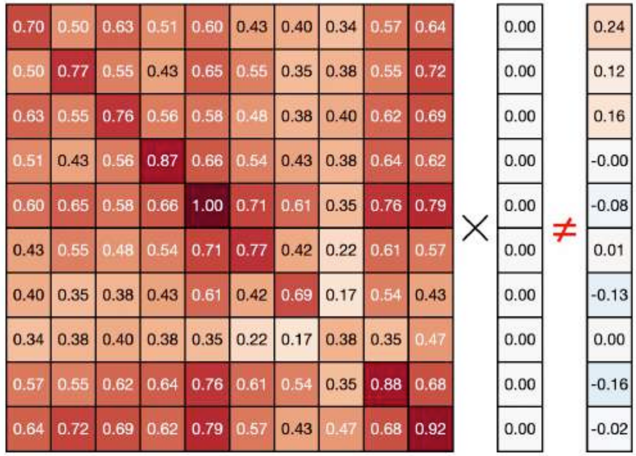
Z. Sun, *et al.*, PNAS 116, 4123 (2019)

Simulazione circuitale





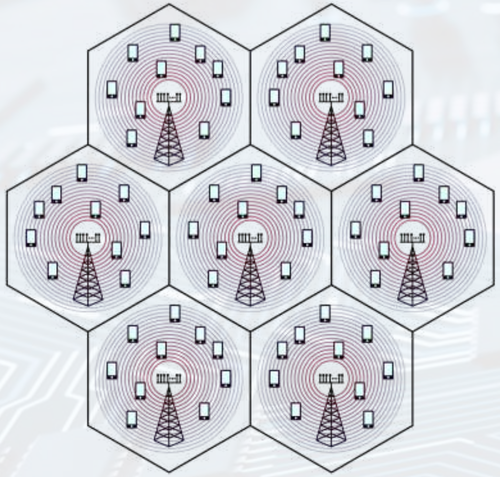


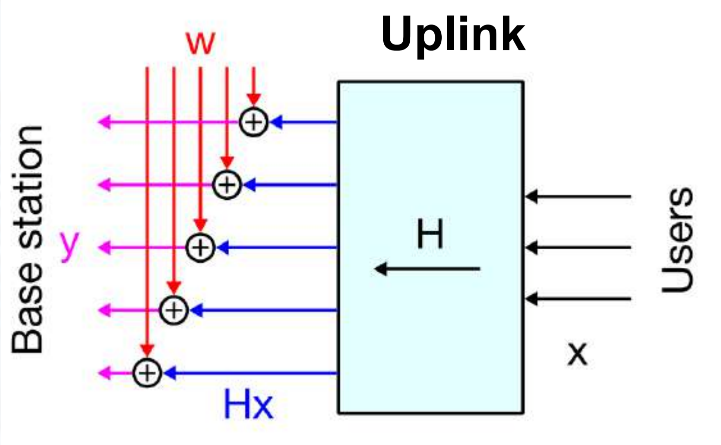


$$\begin{bmatrix} 0.70 & 0.50 & 0.63 & 0.51 & 0.60 & 0.43 & 0.40 & 0.34 & 0.57 & 0.64 \\ 0.50 & 0.77 & 0.55 & 0.43 & 0.65 & 0.55 & 0.35 & 0.38 & 0.55 & 0.72 \\ 0.63 & 0.55 & 0.76 & 0.56 & 0.58 & 0.48 & 0.38 & 0.40 & 0.62 & 0.69 \\ 0.51 & 0.43 & 0.56 & 0.87 & 0.66 & 0.54 & 0.43 & 0.38 & 0.64 & 0.62 \\ 0.60 & 0.65 & 0.58 & 0.66 & 1.00 & 0.71 & 0.61 & 0.35 & 0.76 & 0.79 \\ 0.43 & 0.55 & 0.48 & 0.54 & 0.71 & 0.77 & 0.42 & 0.22 & 0.61 & 0.57 \\ 0.40 & 0.35 & 0.38 & 0.43 & 0.61 & 0.42 & 0.69 & 0.17 & 0.54 & 0.43 \\ 0.34 & 0.38 & 0.40 & 0.38 & 0.35 & 0.22 & 0.17 & 0.38 & 0.35 & 0.47 \\ 0.57 & 0.55 & 0.62 & 0.64 & 0.76 & 0.61 & 0.54 & 0.35 & 0.88 & 0.68 \\ 0.64 & 0.72 & 0.69 & 0.62 & 0.79 & 0.57 & 0.43 & 0.47 & 0.68 & 0.92 \end{bmatrix} \times \begin{bmatrix} 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix} \neq \begin{bmatrix} 0.24 \\ 0.12 \\ 0.16 \\ -0.00 \\ -0.08 \\ 0.01 \\ -0.13 \\ 0.00 \\ -0.16 \\ -0.02 \end{bmatrix}$$

- Nuovo concetto di calcolo sviluppato presso PoliMI è in grado di risolvere sistemi lineari di equazioni, inversioni di matrici e regressioni lineari/logistiche in una sola operazione → complessità O(1)

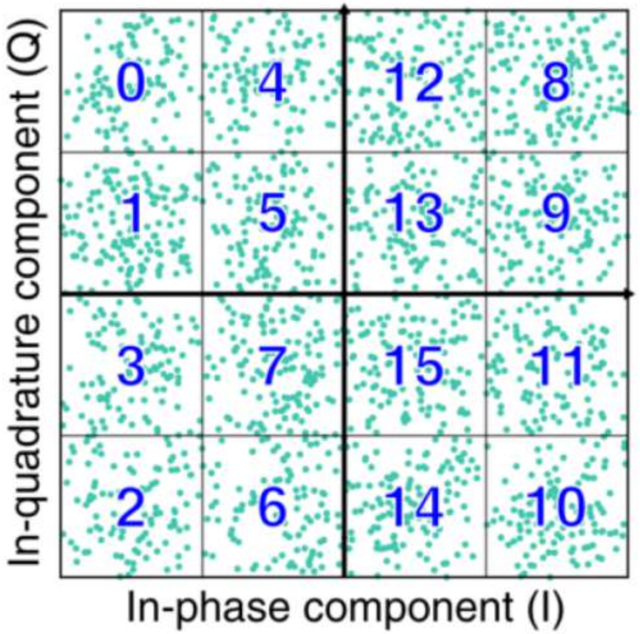
Massive multiple-in multiple-out (MIMO)





Zero Forcing

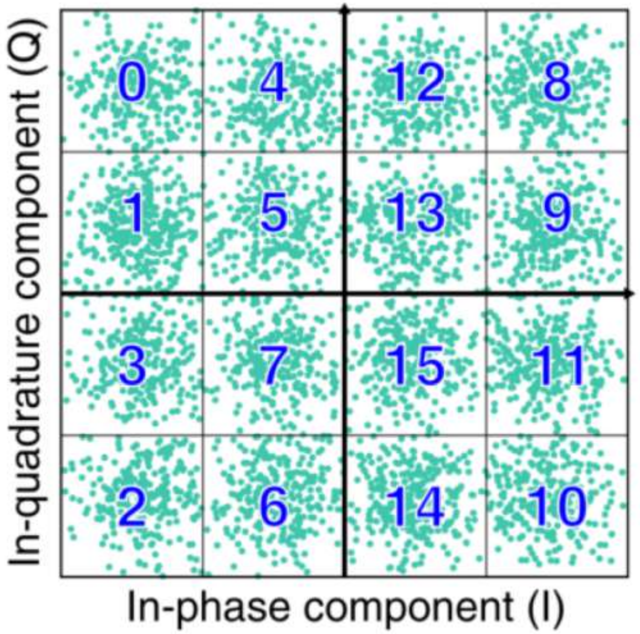
$$\tilde{x} = (H^* H)^{-1} H^* y$$



- × Amplifies noise
- ✓ Requires less information

Regularized Zero-Forcing

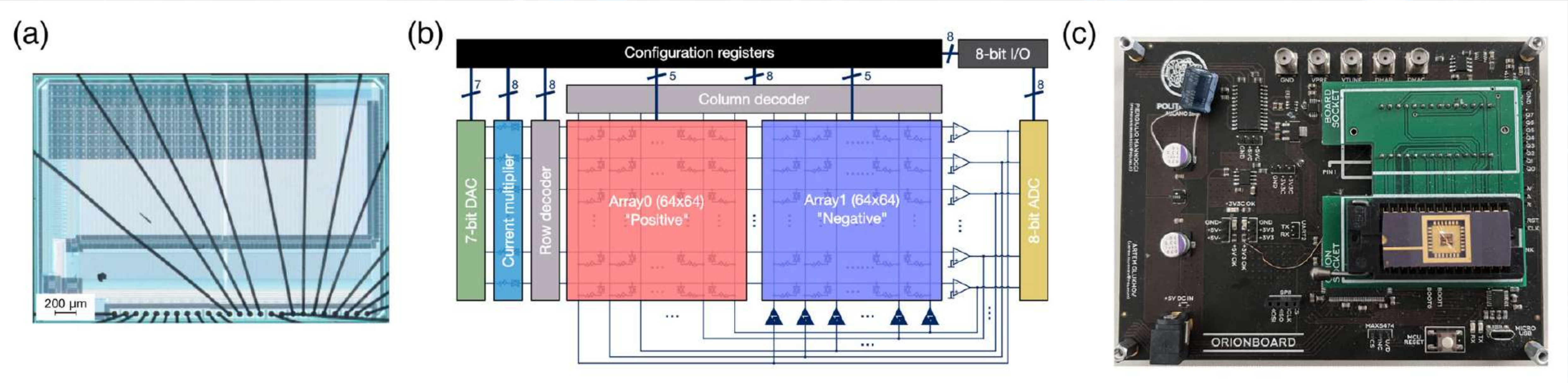
$$\tilde{x} = (H^* H + \sigma_n^2 I)^{-1} H^* y$$

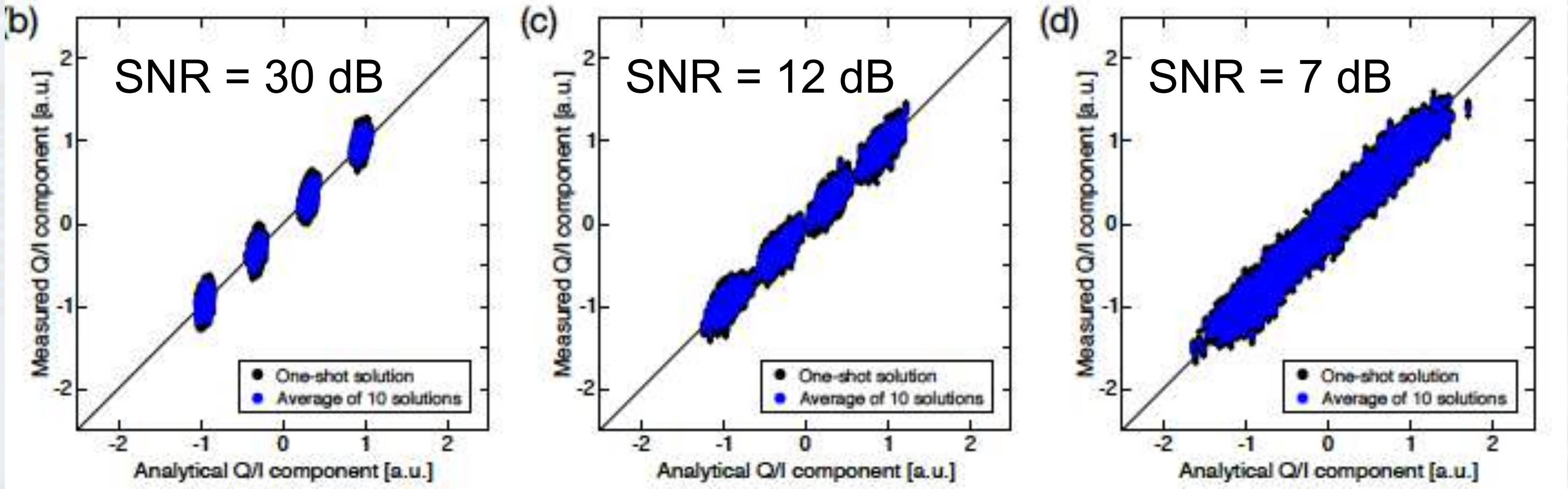


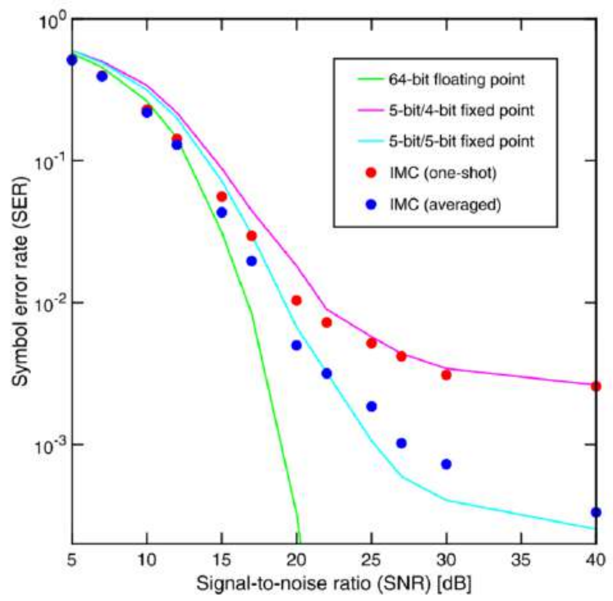
- ✓ Balances signal and noise
- × Requires channel noise estimate

P. Mannocci, et al., IEEE ISCAS (2023)

Circuito integrato riconfigurabile per la decodifica MIMO

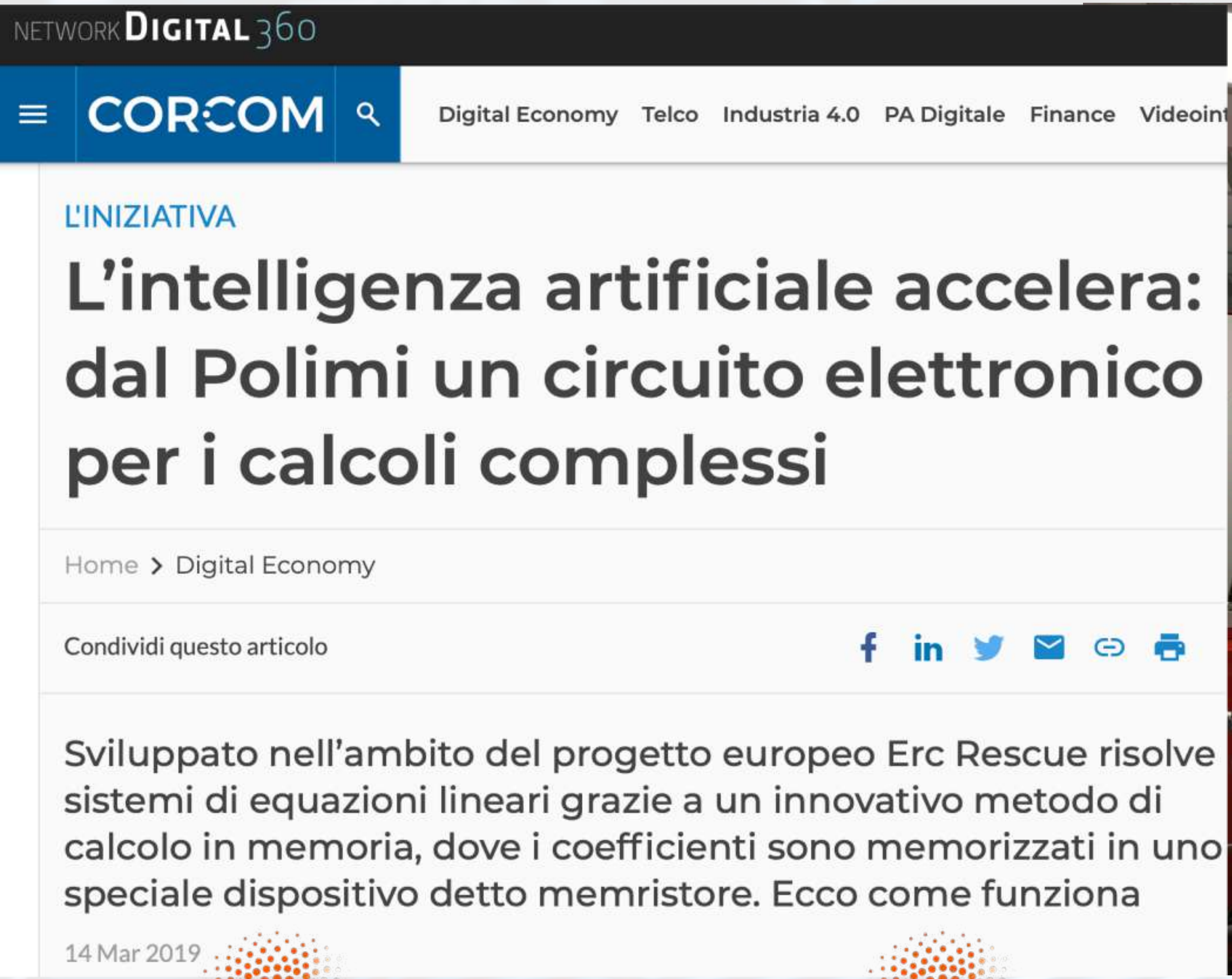

(a)
200 μm
(b)
Configuration registers
8-bit I/O
Column decoder
7-bit DAC
Current multiplier
Row decoder
Array0 (64x64) "Positive"
Array1 (64x64) "Negative"
8-bit ADC
7
8
8
5
8
5
8
8
(c)
ORIONBOARD





P. Mannocci, et al., IEEE ISCAS (2023)

Il contributo del Politecnico di Milano


NETWORK DIGITAL 360
CORCOM
Digital Economy
Telco
Industria 4.0
PA Digitale
Finance
Videoint

L'INIZIATIVA

L'intelligenza artificiale accelera: dal Polimi un circuito elettronico per i calcoli complessi

Home > Digital Economy

Condividi questo articolo

Sviluppato nell'ambito del progetto europeo Erc Rescue risolve sistemi di equazioni lineari grazie a un innovativo metodo di calcolo in memoria, dove i coefficienti sono memorizzati in uno speciale dispositivo detto memristore. Ecco come funziona

14 Mar 2019

erc RESCUE (2015-2020)

erc ANIMATE (2023-2028)

Look inside!

POLITECNICO

MILANO 1863