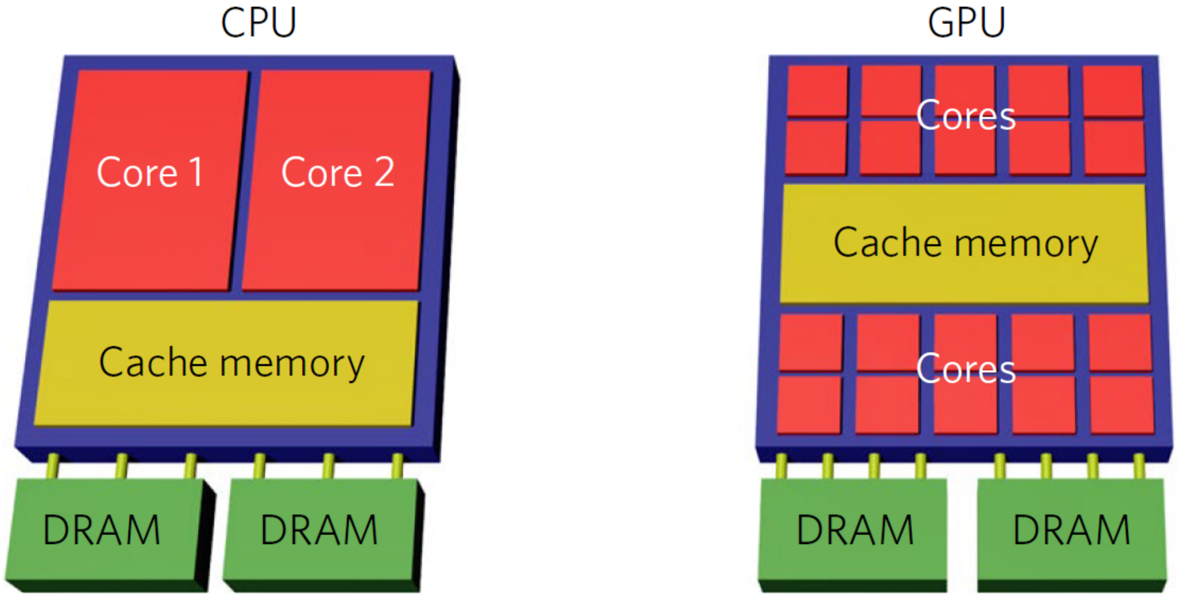
POLITECNICO
MILANO 1863

In-memory computing with emerging memory devices

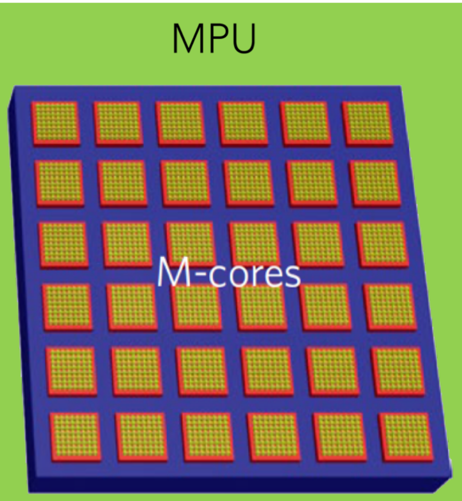
Daniele Ielmini

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

From von Neumann to in-memory computing



Von Neumann architecture
- Volatile memory (DRAM)
- Data movement
- Memory bottleneck



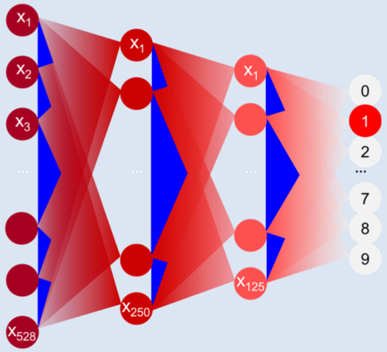
In-memory computing
- Resistive memory
- Compute *in situ*
- High parallelism

M A Zidan, *et al. Nat. Electron.* (2018)

Three types of in-memory computing

In-memory computing

Artificial neural networks



Algebra accelerators

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

Brain-inspired neural networks

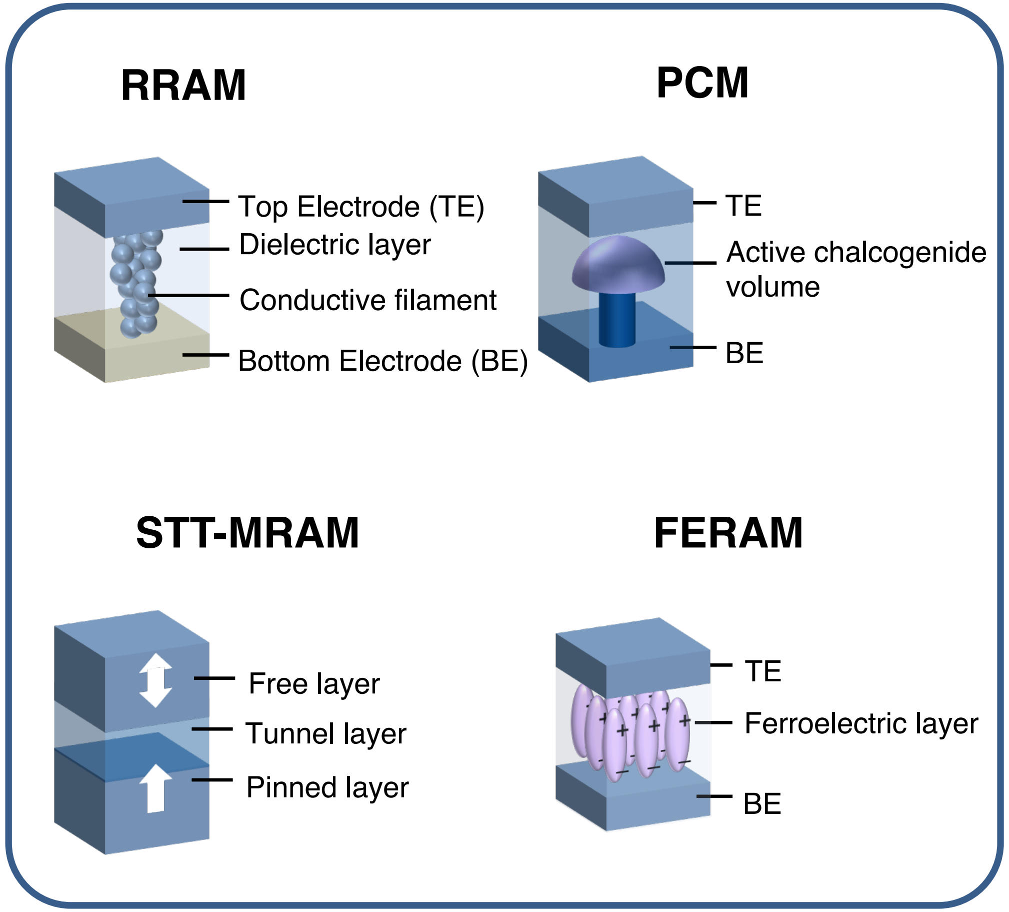
- Deep neural networks
- Convolutional neural networks
- LSTM
- Supervised training
- Backpropagation

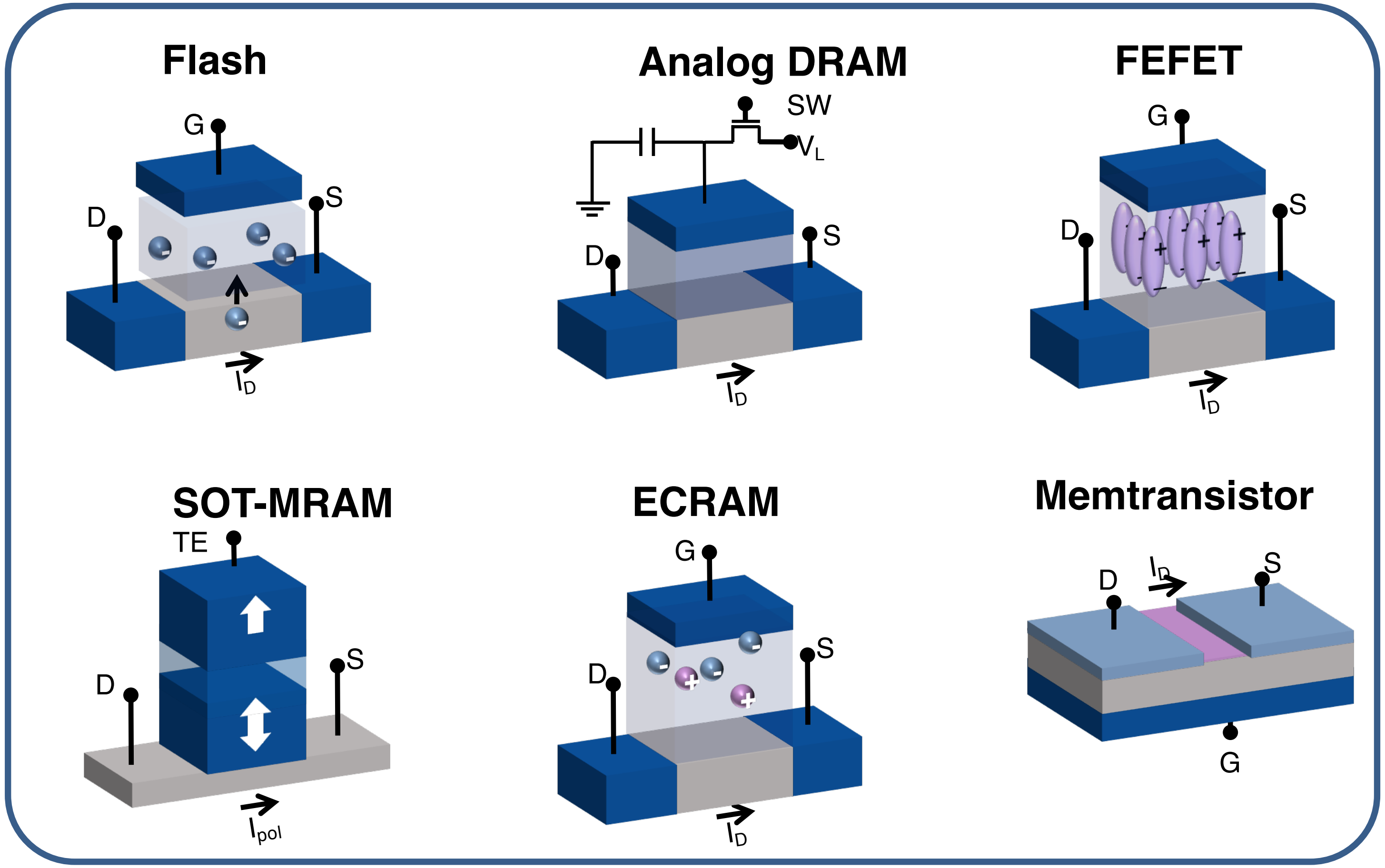
- Linear system solver
- Matrix inversion
- Regression
- Eigenvector calculation
- Differential equations

- Unsupervised learning
- Reinforcement learning
- Associative memories
- Spatiotemporal networks
- Reservoir computing

Adapted from D. Ielmini and S. Ambrogio, Nanotechnology 31, 092001 (2019)

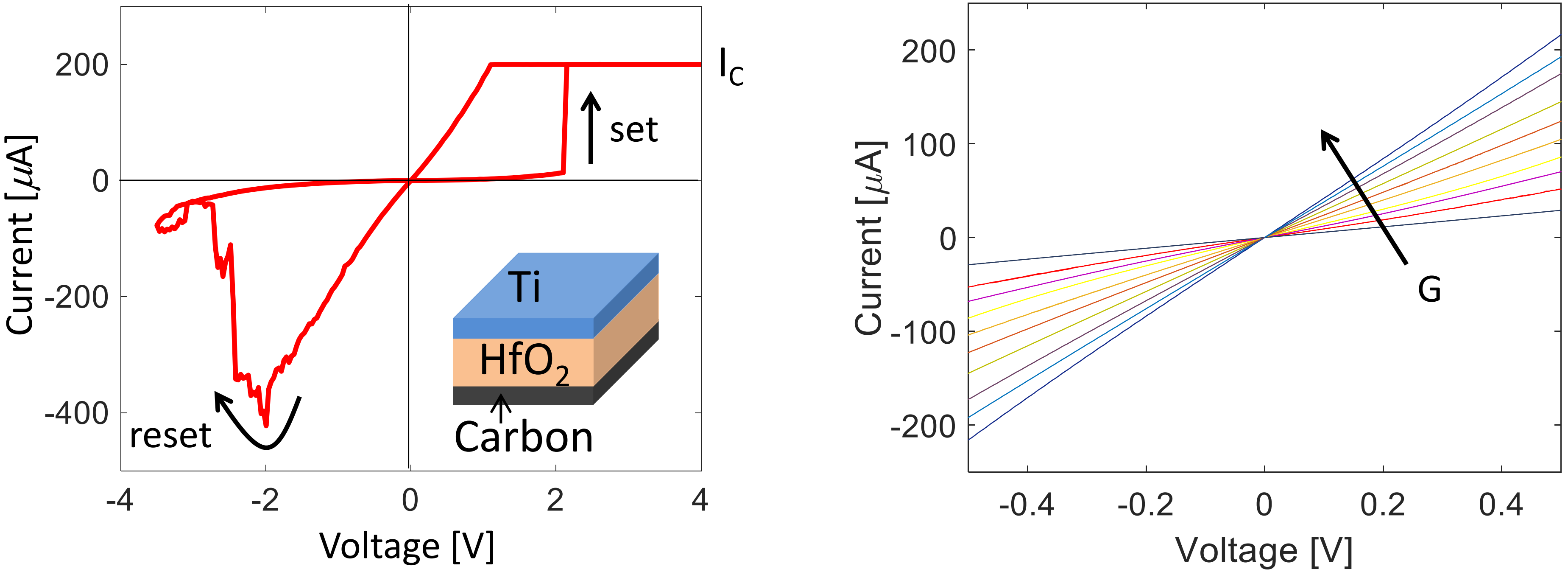
Memory devices for in-memory computing





D. Ielmini and G. Pedretti, Adv. Intell. Syst. 1, 2000040 (2020)

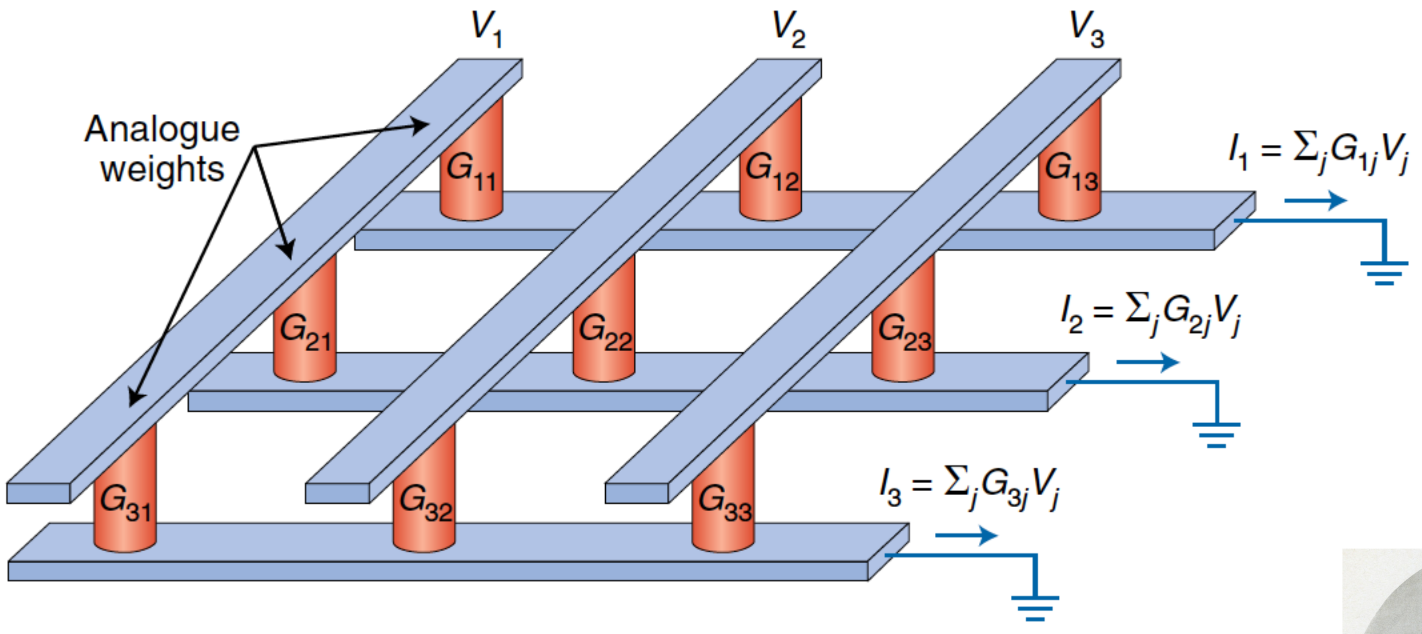
Analogue resistive memory



- Analogue conductance controlled by a compliance current I_C
- Good linearity below 0.5 V

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

Matrix-vector multiplication



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

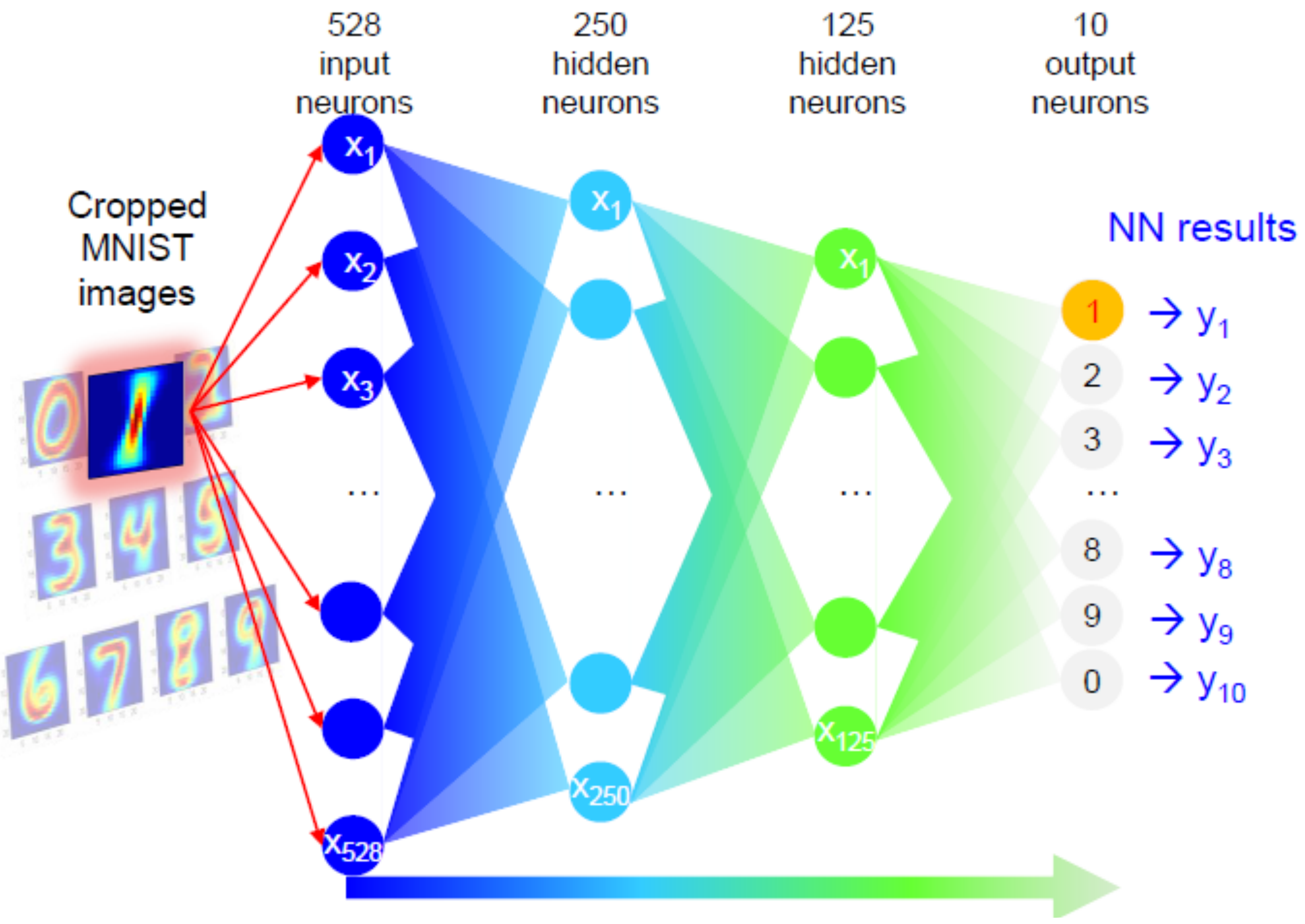
- Multiplying a matrix A and a vector x in a CPU requires individual products a_{ij}*x_j, and summation → multiply/accumulate (MAC) process
- In a crossbar, the operation is carried out <u>physically</u> by Kirchhoff's and Ohm's law, in just <u>one step</u>

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

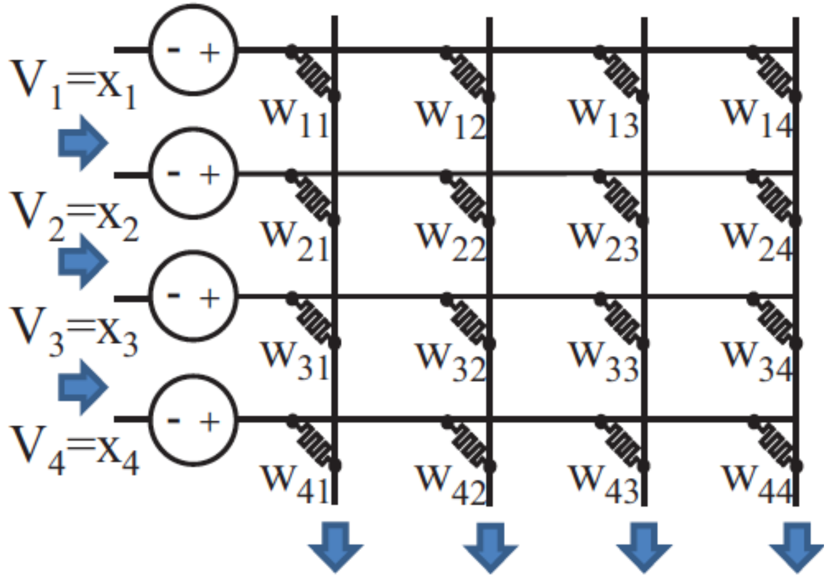
Kirchhoff's law

Ohm's law

1 – Forward propagation



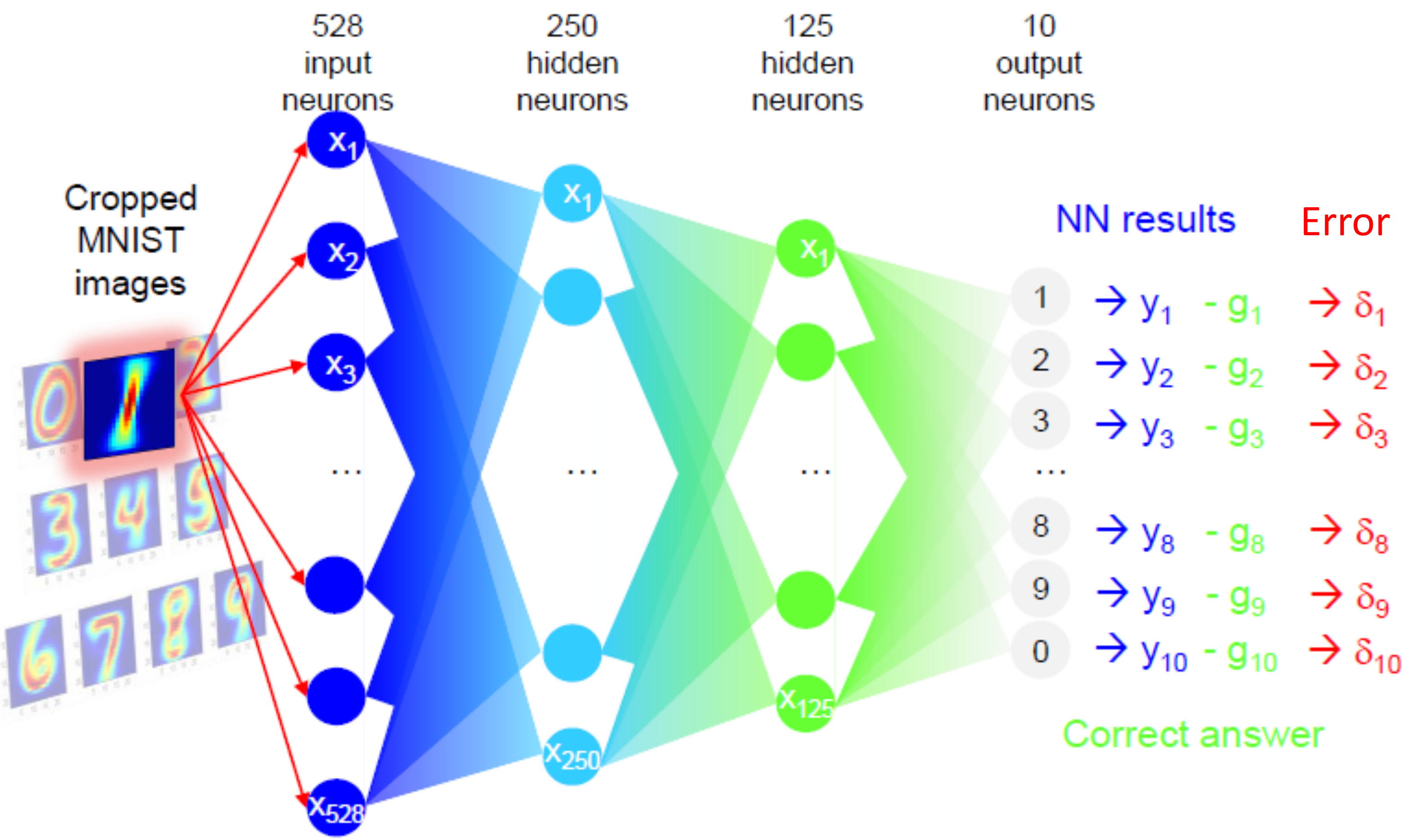
In-memory MVM



$$\sum_i x_i w_{ij}$$

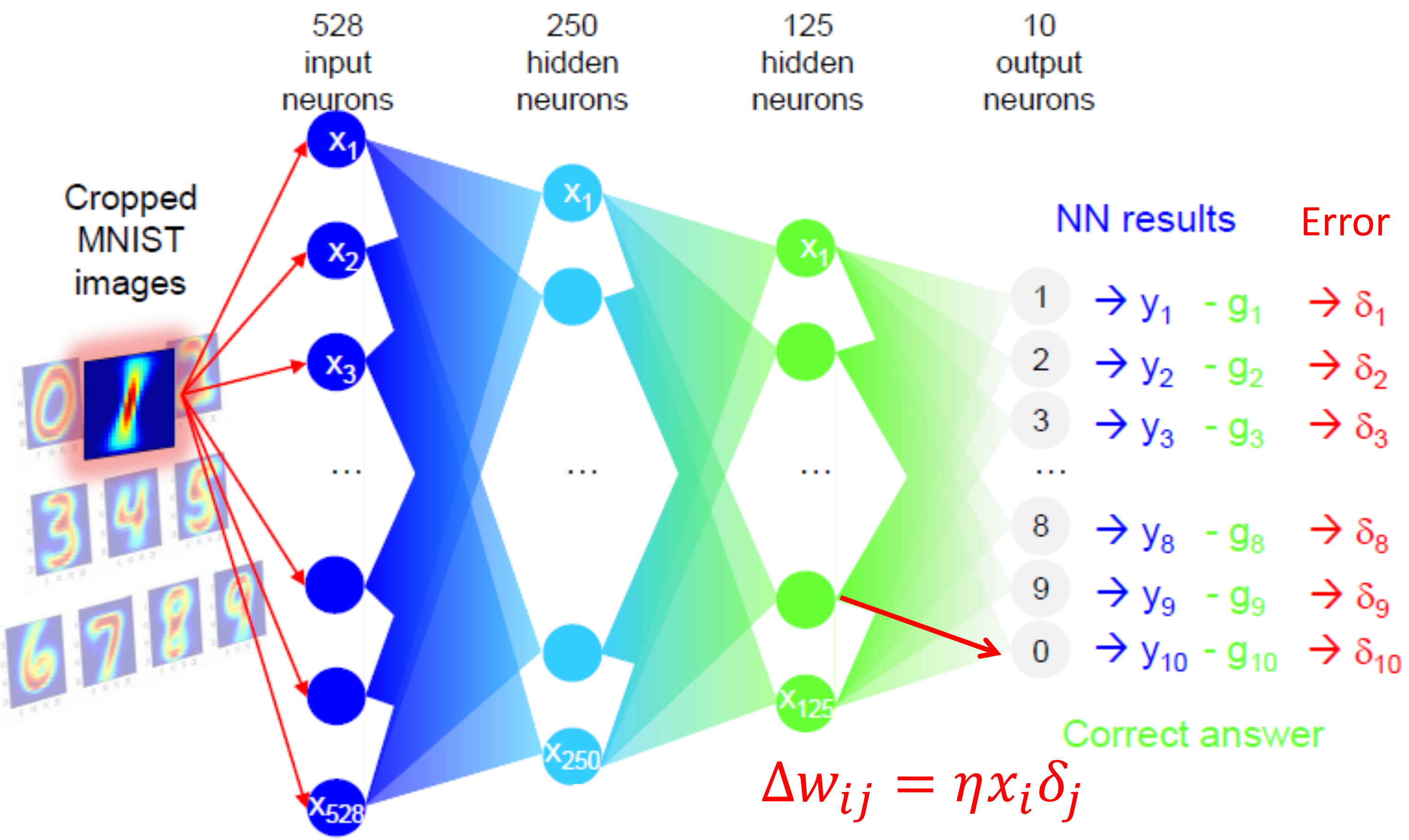
S. Agarwal, et al., IJCNN 929 (2016)

2 – Error evaluation



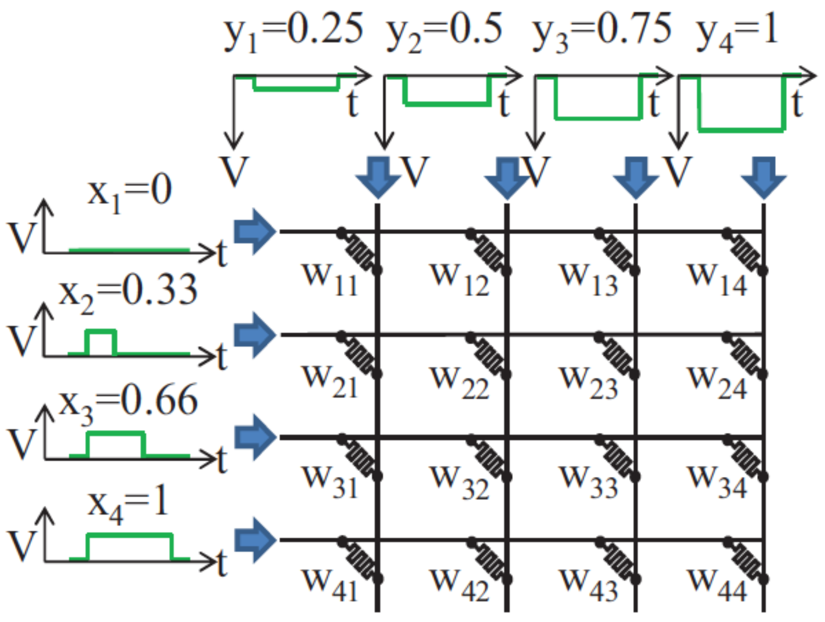
Supervised training = pattern is submitted with the corresponding label → we know the correct answer

3 – Weight update



$$\Delta w_{ij} = \eta x_i \delta_j$$

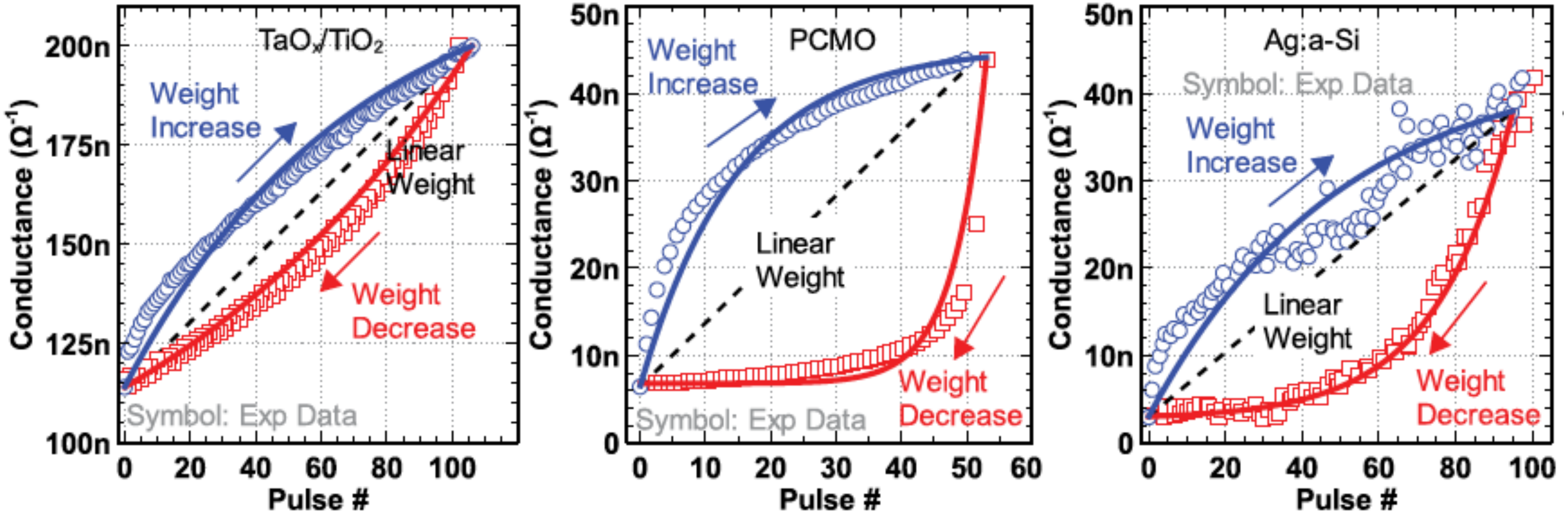
In-memory outer-product

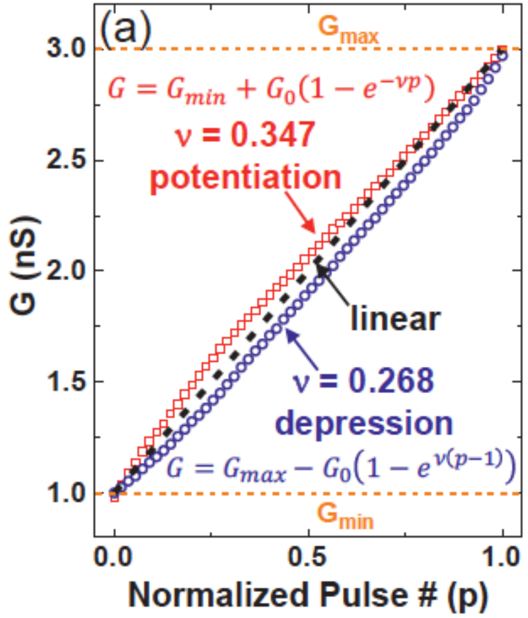


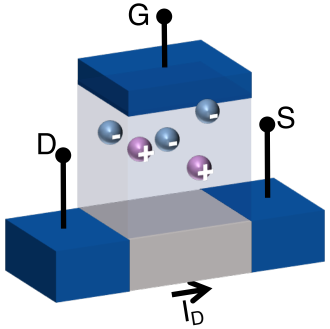
$$\Delta w_{ij} = x_i y_j$$

S. Agarwal, et al., IJCNN 929 (2016)

Device non-linearity

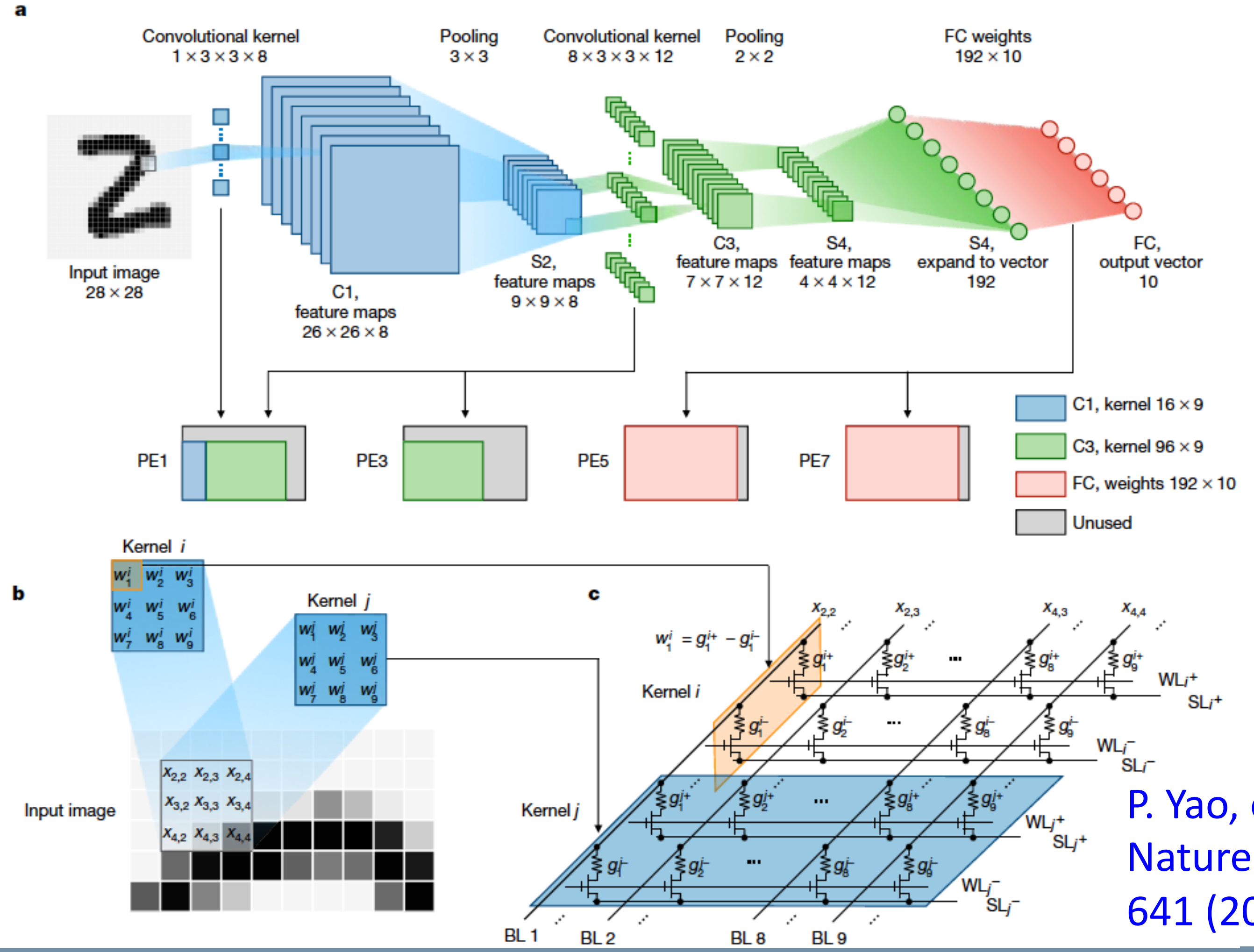




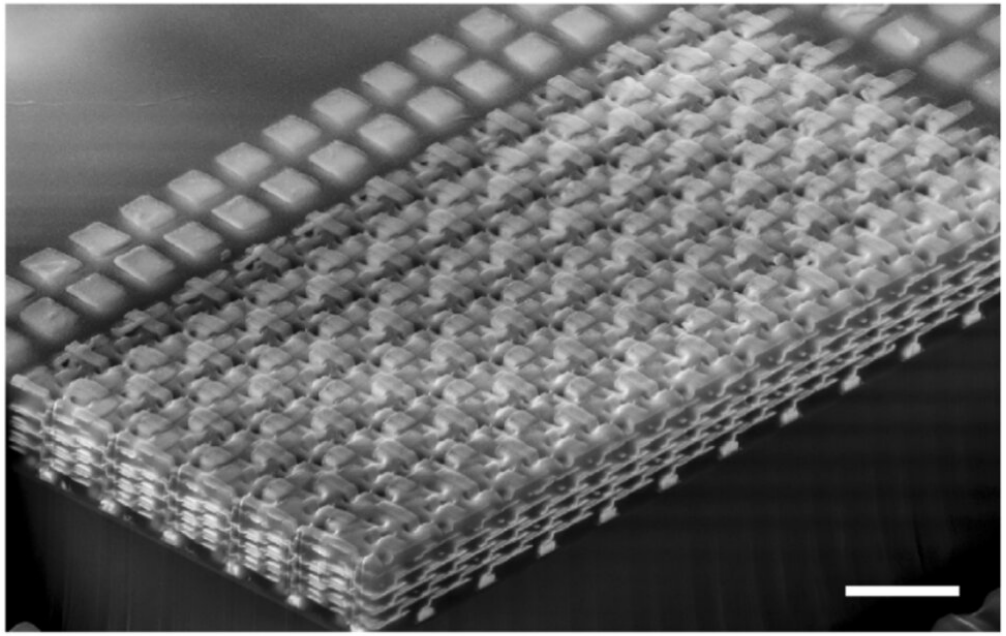


- In general, physical devices are not linear in time and voltage
- Record linearity for Li-based ECRAM (IBM, IEDM 2018)

In-memory convolutional neural networks (CNNs)



P. Yao, et al., Nature 577, 641 (2020)

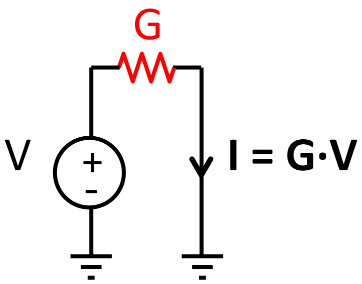


3D RRAM array

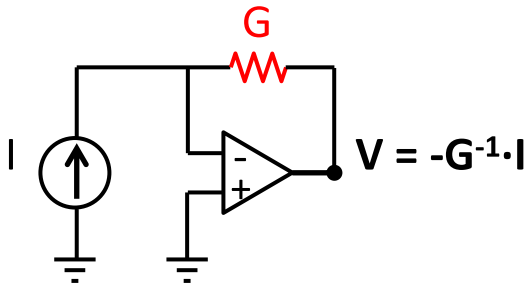
P. Lin, et al., Nat Electron 3, 225 (2020)

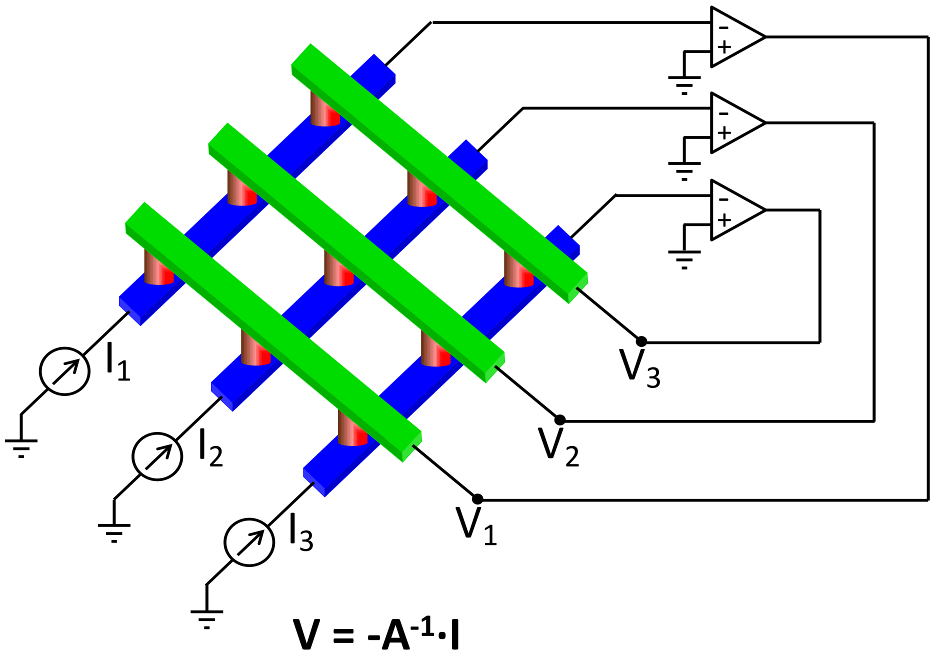
Inverting MVM


Georg Ohm




Harold Black





Matrix-vector division is equivalent to solving $Ax = b$, with $x = A^{-1}b$

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

O(1) complexity of inverse MVM

PRL **103,** 150502 (2009) PHYSICAL REVIEW LETTERS week ending 9 OCTOBER 2009

Quantum Algorithm for Linear Systems of Equations

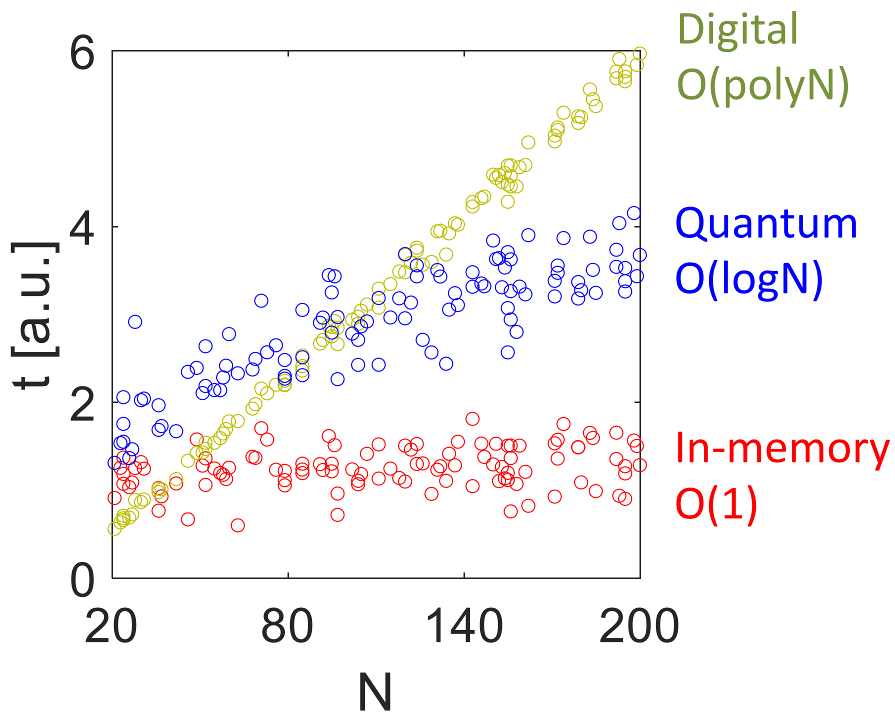
Aram W. Harrow,[1] Avinatan Hassidim,[2] and Seth Lloyd[3]

[1]*Department of Mathematics, University of Bristol, Bristol, BS8 1TW, United Kingdom*
[2]*Research Laboratory for Electronics, MIT, Cambridge, Massachusetts 02139, USA*
[3]*Research Laboratory for Electronics and Department of Mechanical Engineering, MIT, Cambridge, Massachusetts 02139, USA*

(Received 5 July 2009; published 7 October 2009)

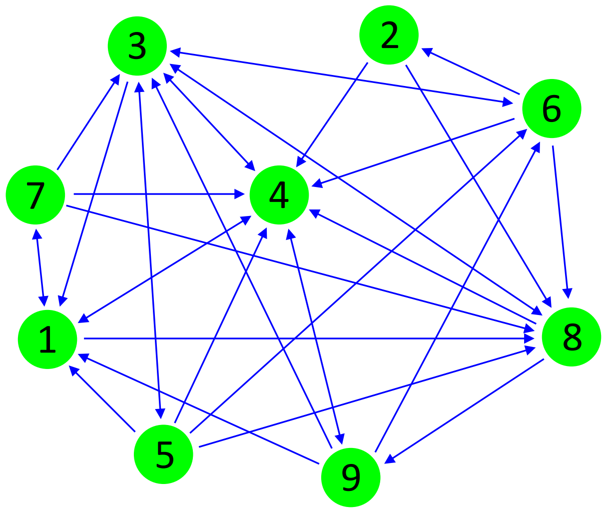
Solving linear systems of equations is a common problem that arises both on its own and as a subroutine in more complex problems: given a matrix A and a vector $\vec{b}$, find a vector $\vec{x}$ such that $A\vec{x} = \vec{b}$. We consider the case where one does not need to know the solution $\vec{x}$ itself, but rather an approximation of the expectation value of some operator associated with $\vec{x}$, e.g., $\vec{x}^\dagger M \vec{x}$ for some matrix M. In this case, when A is sparse, $N \times N$ and has condition number κ, the fastest known classical algorithms can find $\vec{x}$ and estimate $\vec{x}^\dagger M \vec{x}$ in time scaling roughly as $N\sqrt{\kappa}$. Here, we exhibit a quantum algorithm for estimating $\vec{x}^\dagger M \vec{x}$ whose runtime is a polynomial of $\log(N)$ and κ. Indeed, for small values of κ [i.e., poly $\log(N)$], we prove (using some common complexity-theoretic assumptions) that any classical algorithm for this problem generically requires exponentially more time than our quantum algorithm.

DOI: 10.1103/PhysRevLett.103.150502 PACS numbers: 03.67.Ac, 02.10.Ud, 89.70.Eg

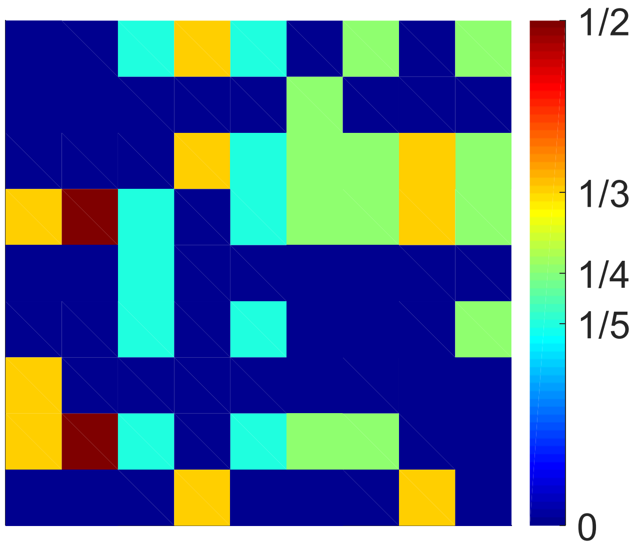


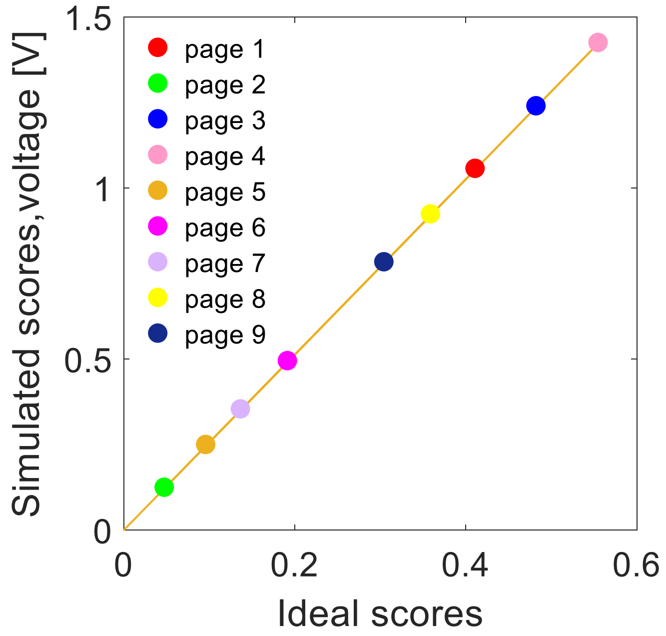
Z. Sun, *et al.*, IEEE Trans. Electron Devices 67, 2945 (2020)

In-memory PageRank



Page No.	Company
1	Amazon
2	Baidu
3	Facebook
4	Google
5	LinkedIn
6	PayPal
7	Quora
8	Twitter
9	YouTube



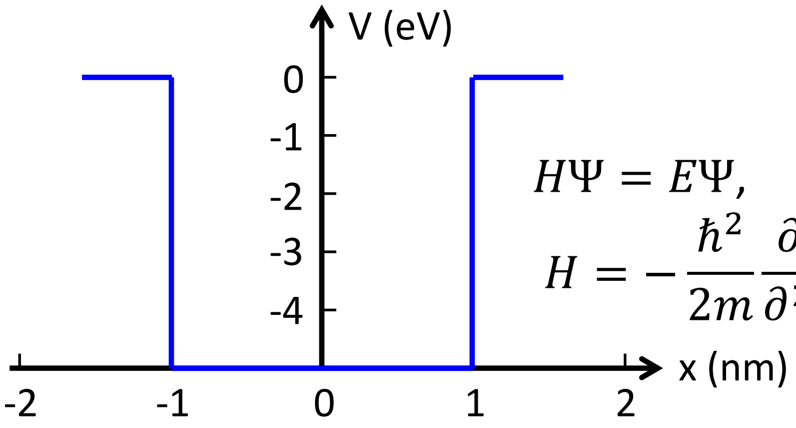


Z. Sun, *et al.*, PNAS 116, 4123 (2019)

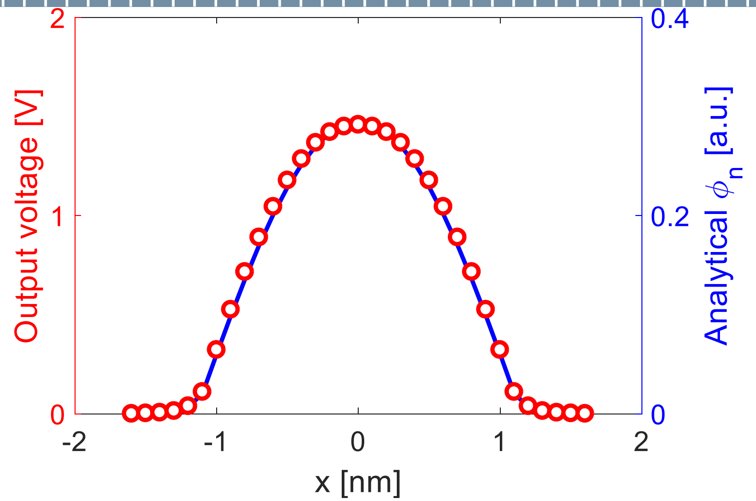
	Throughput [TOPS]	Energy efficiency [TOPS/W]
In-memory	0.183	362
TPU	92	2.3

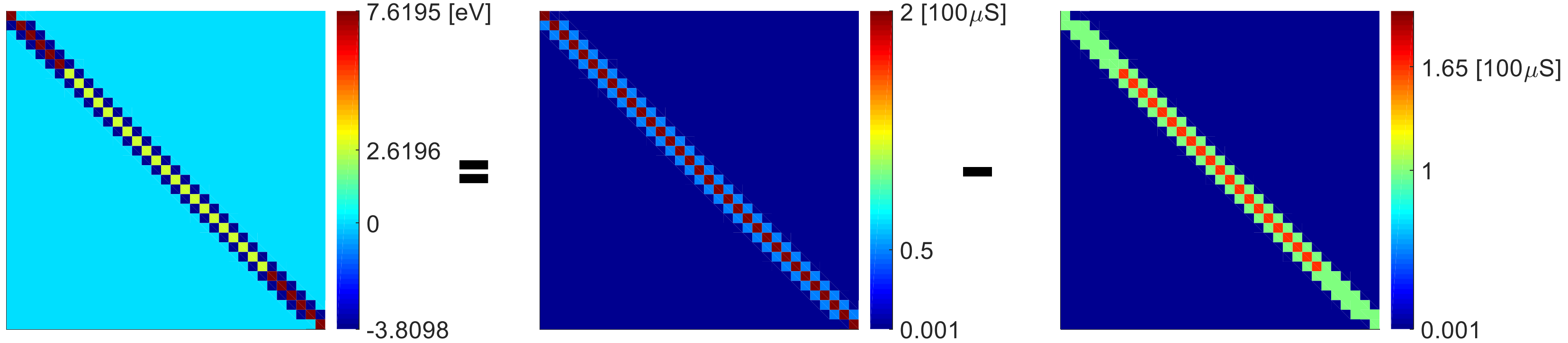
150X

Solving a Schrödinger equation



$$H\Psi = E\Psi,$$
$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x} + V$$

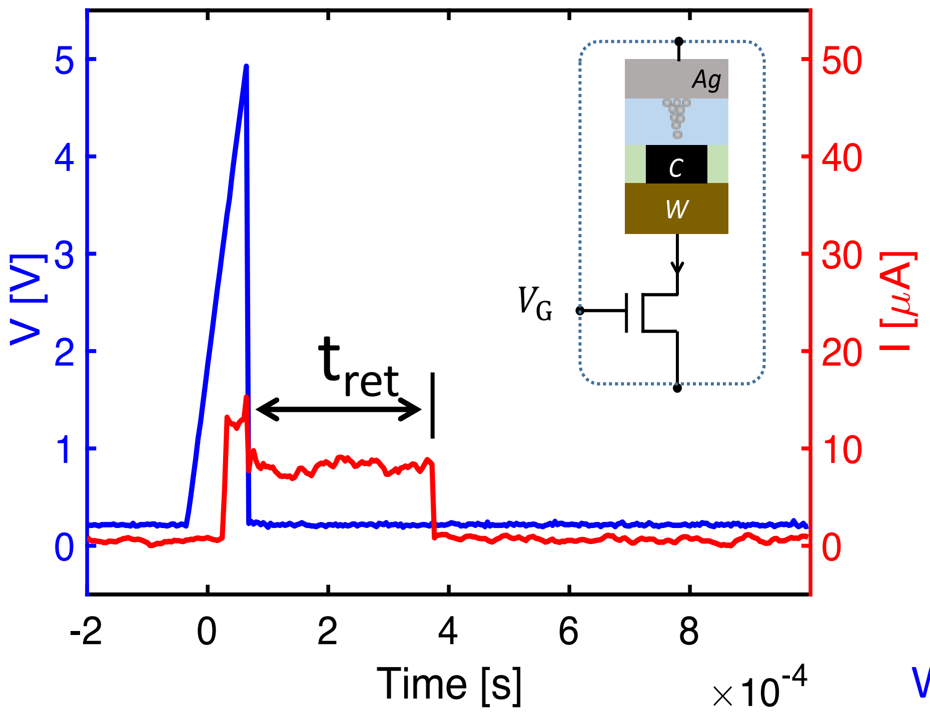


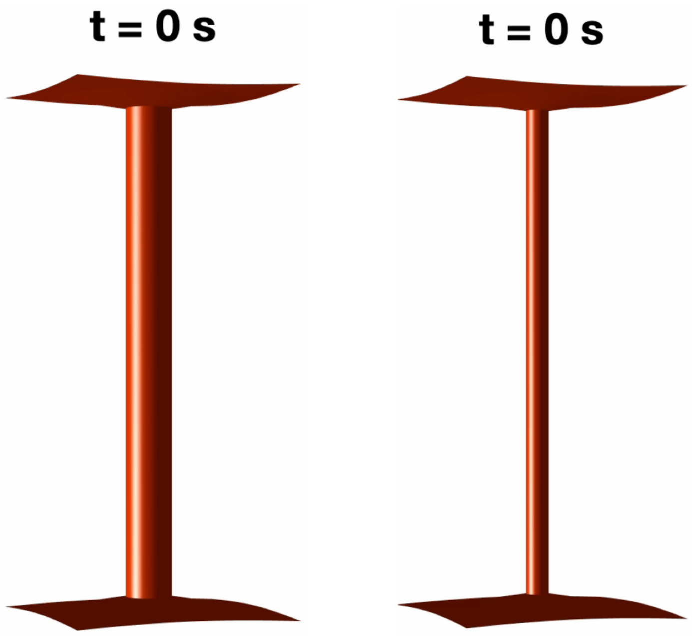


matrix **A** = matrix **B** - matrix **C**

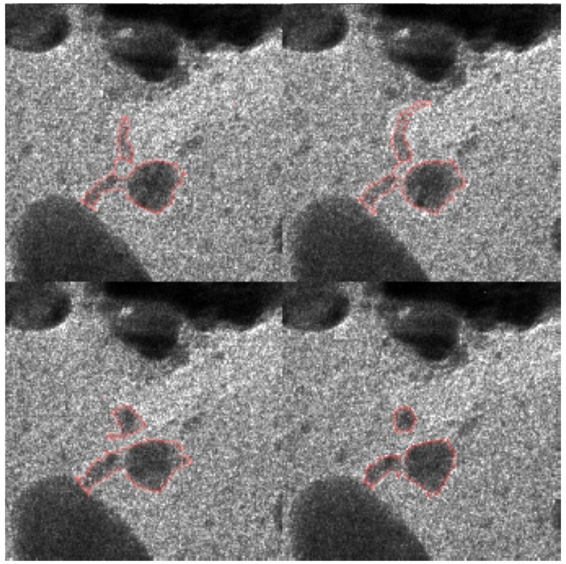
Z. Sun, *et al.*, PNAS 116, 4123 (2019)

Volatile RRAM





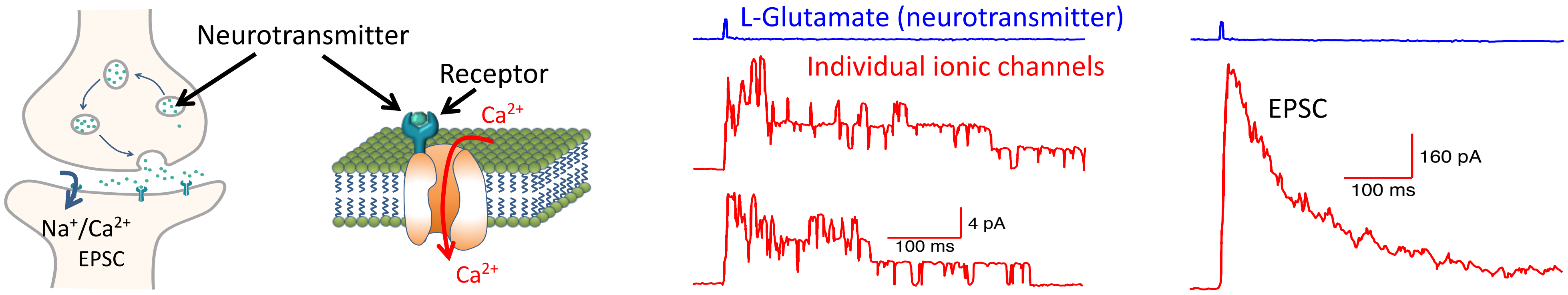
W. Wang, et al., Nat. Commun., 10, 81 (2019)



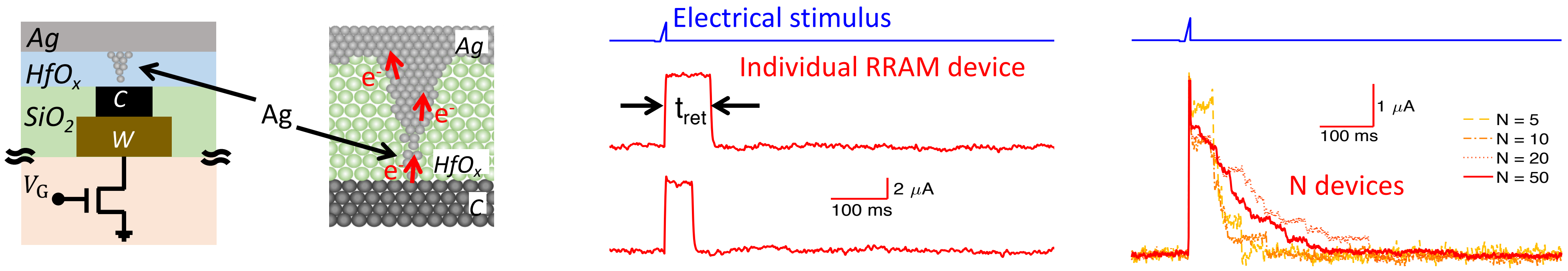
Z. Wang et al.,
Nat. Mater. **16** 101 (2017)

- Volatile behavior due to Ag diffusion and filament disconnection in the µs-ms timescale

Analogy with biological synapses

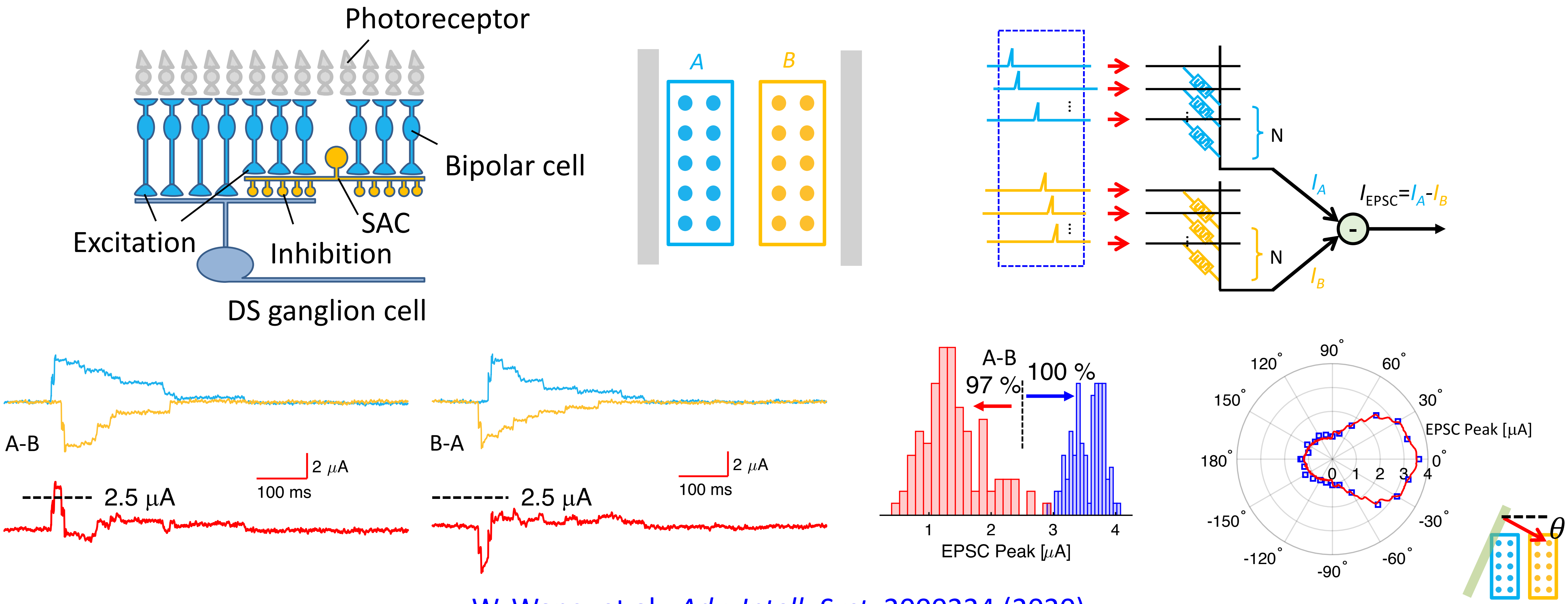


R. A. J. Lester and C. E. Jahr, *J. Neurosci.*, 12, 635 (1992)



W. Wang, et al., *Adv. Intell. Syst.* 2000224 (2020)

Direction selectivity in the human retina



W. Wang, et al., *Adv. Intell. Syst.* 2000224 (2020)

Conclusions

- In-memory computing by crosspoint operations:
  - MVM (dot product) → DNN inference
  - Outer product → DNN training
  - MVM + feedback (inverse MVM) → linear algebra
- Device physics for brain-inspired neuromorphic computing (short-term RRAM)

grazie!