Phase-Change Memory Devices: Fundamentals and Applications (Part II)

Abu Sebastian
Principal Research Staff Member
IBM Research - Zurich

Oxford, March 18, 2019

Acknowledgements

- **Neuromorphic and in-memory computing**
  - ✓ Thomas Bohnstingl
  - ✓ Irem Boybat
  - ✓ Iason Giannopoulos
  - ✓ Riduan Khaddam-Aljameh
  - ✓ Benedikt Kersting
  - ✓ Christophe Piveteou
  - ✓ Vinay Joshi
  - ✓ S. R. Nandakumar
  - ✓ Timoleon Moraitis
  - ✓ Stanislaw Wozniak
  - ✓ Varaprasad Jonnalagadda
  - ✓ Manuel Le Gallo
  - ✓ Angeliki Pantazi
  - ✓ Giovanni Cherubini
  - ✓ Evangelos Eleftheriou
- **Foundations of cognitive solutions**
- **Cloud storage and analytics**
- **IBM TJ Watson Research Center**
- **IBM Research-Almaden**
- **NJIT, Univ. of Patras, RWTH Aachen, ETH, EPFL, Exeter, Oxford**

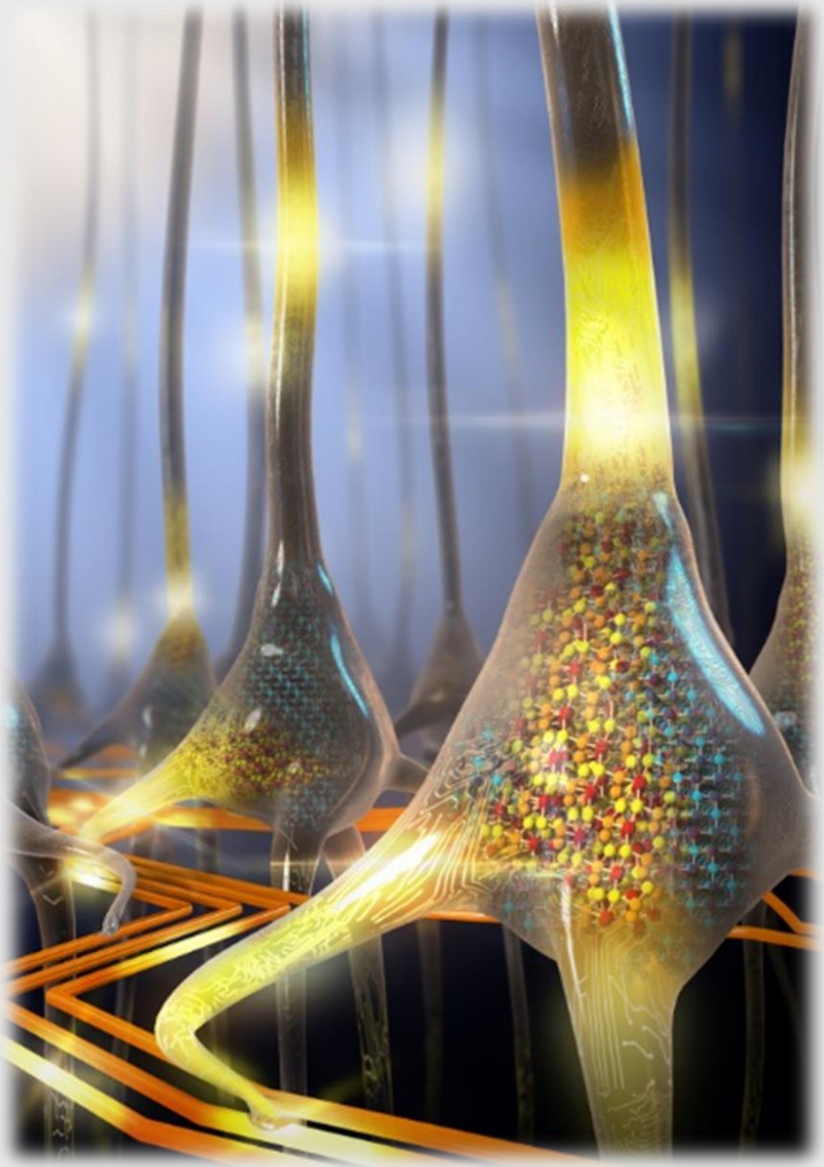
erc
European Research Council
Established by the European Commission

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

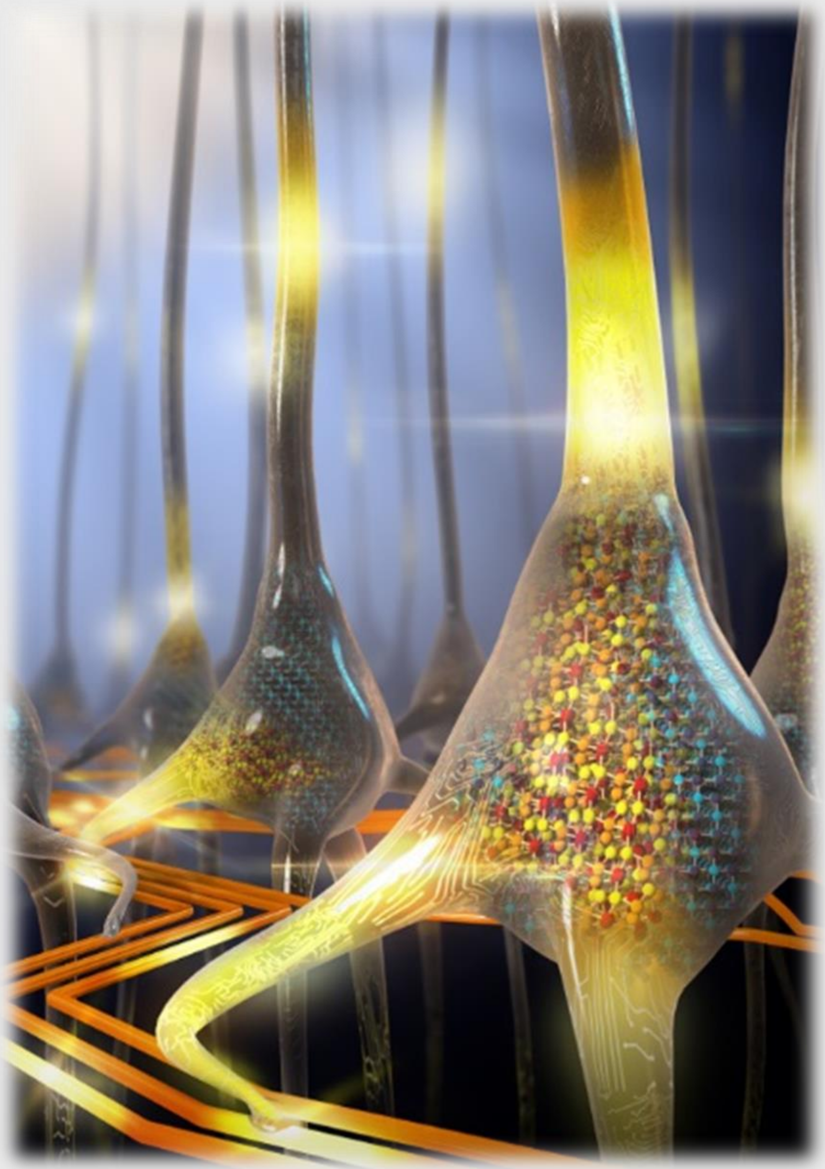
Outline

- **Introduction**
  - ✓ The computing efficiency problem of AI
  - ✓ Brain-inspired computing and the role of memory
  - ✓ Key enablers for brain-inspired computing
- **First level of inspiration: In-memory computing**
  - ✓ Matrix-vector multiplication and applications
  - ✓ Computing with device dynamics
- **Second level of inspiration: Co-processors for deep learning**
  - ✓ Mixed-precision deep learning
- **Third level of inspiration: Spiking neural networks**
  - ✓ Neuronal and synaptic emulations
  - ✓ Unsupervised learning
- **Summary & Outlook**

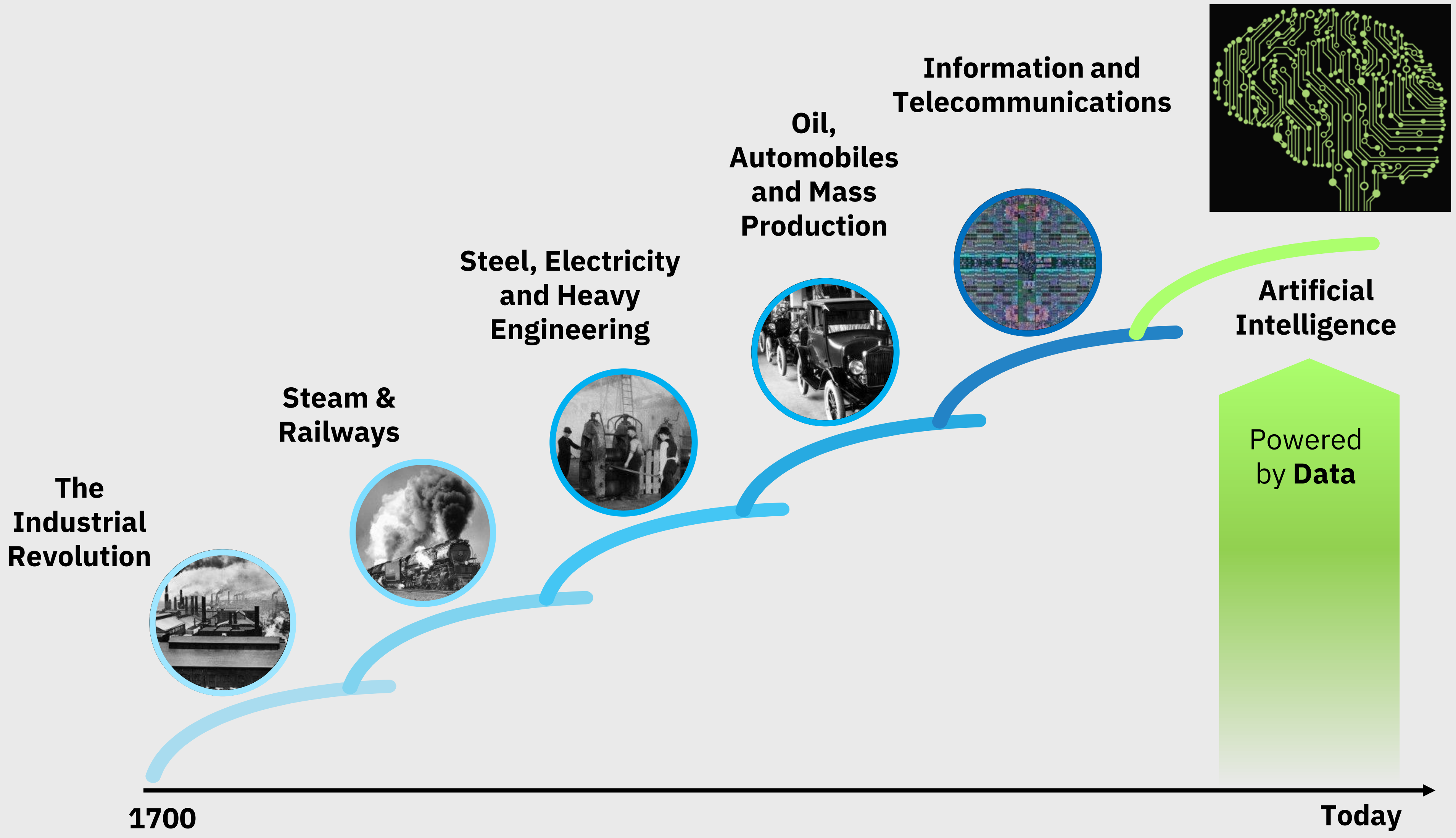


Outline

- **Introduction**
  - ✓ The computing efficiency problem of AI
  - ✓ Brain-inspired computing and the role of memory
  - ✓ Key enablers for brain-inspired computing
- **First level of inspiration: In-memory computing**
  - ✓ Matrix-vector multiplication and applications
  - ✓ Computing with device dynamics
- **Second level of inspiration: Co-processors for deep learning**
  - ✓ Mixed-precision deep learning
- **Third level of inspiration: Spiking neural networks**
  - ✓ Neuronal and synaptic emulations
  - ✓ Unsupervised learning
- **Summary & Outlook**



The AI Revolution



The Industrial Revolution

Steam & Railways

Steel, Electricity and Heavy Engineering

Oil, Automobiles and Mass Production

Information and Telecommunications

Artificial Intelligence

Powered by **Data**

1700

Today

Jeopardy! (2011)

~80,000 W

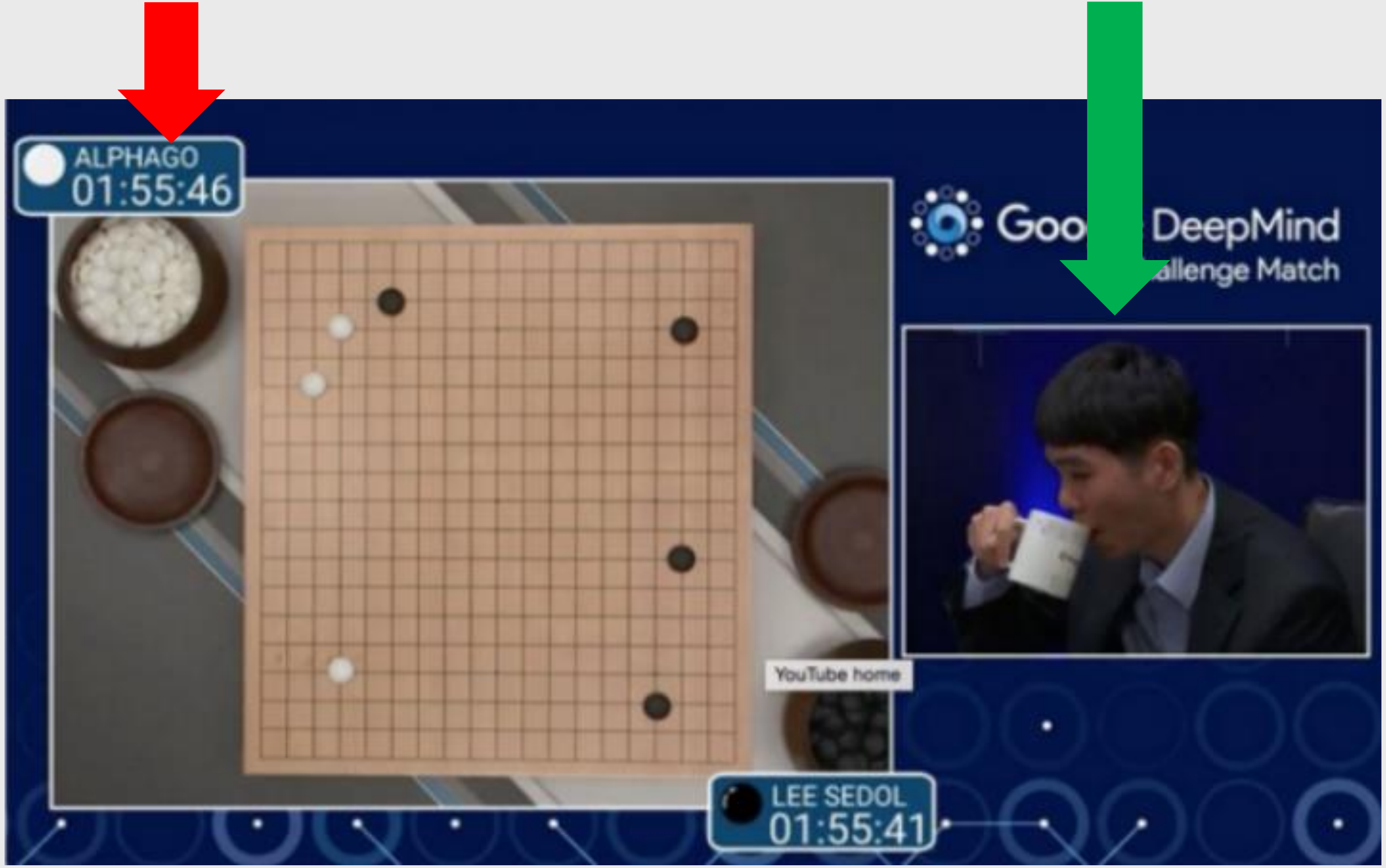
~20W

- 2880 processor threads
- 16 terabytes of RAM
- 20 tons of air-conditioned cooling capacity

AlphaGo (2016)

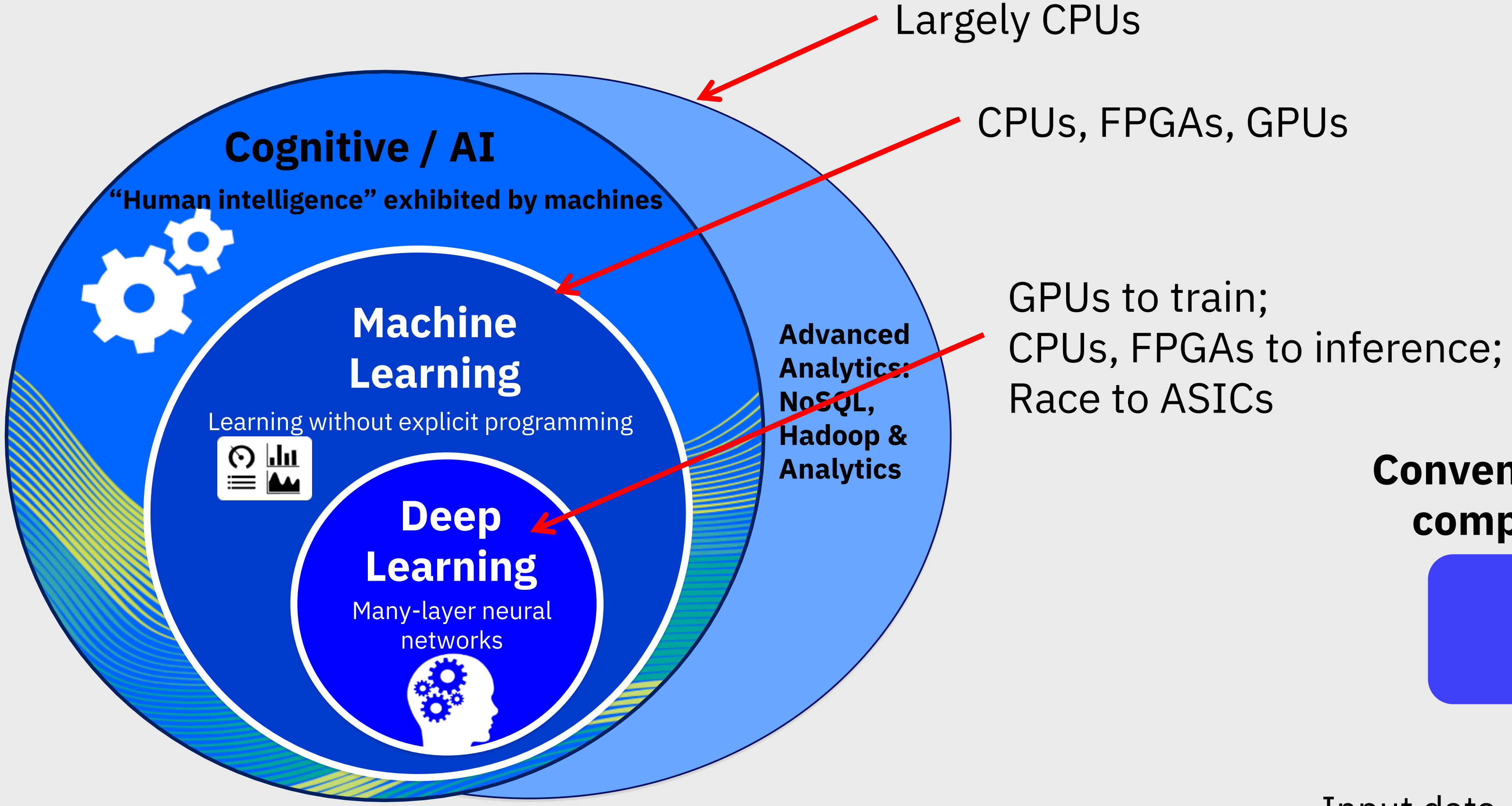
~1,000,000 W

~20W



- 1202 CPUs
- 176 GPUs

AI’s computing efficiency problem



Cognitive / AI

“Human intelligence” exhibited by machines

Machine Learning

Learning without explicit programming

Deep Learning

Many-layer neural networks

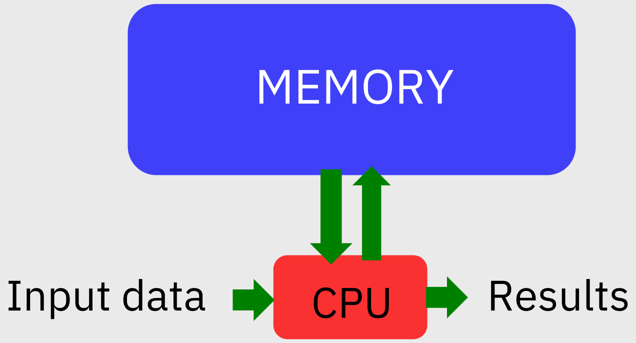
Advanced Analytics: NoSQL, Hadoop & Analytics

Largely CPUs

CPUs, FPGAs, GPUs

GPUs to train;
CPUs, FPGAs to inference;
Race to ASICs

Conventional von Neumann computing architecture



MEMORY

Input data

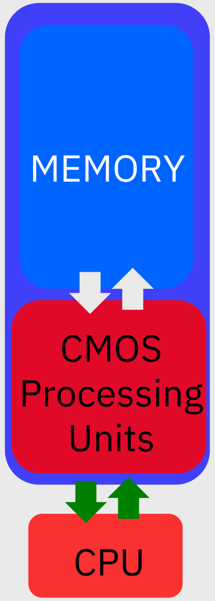
CPU

Results

Pedram et al., IEEE MICRO, 2017

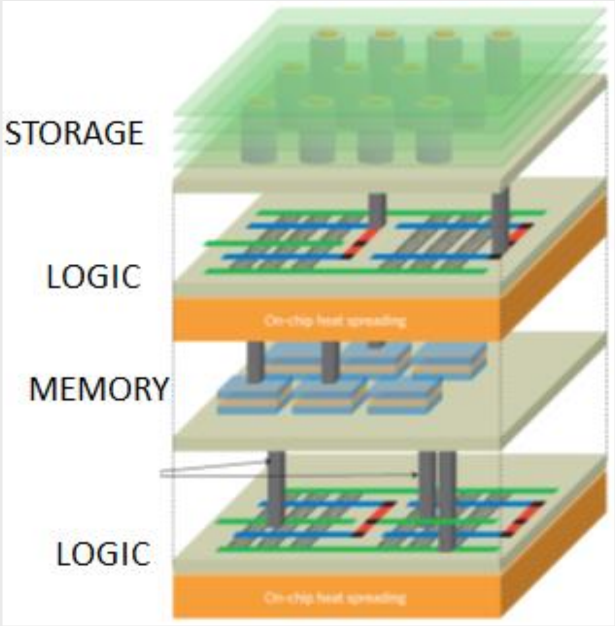
Advances in von Neumann computing

Processor-in-memory (near memory computing)



Vermij et al., Proc. ACM CF, 2016

Monolithic 3D integration

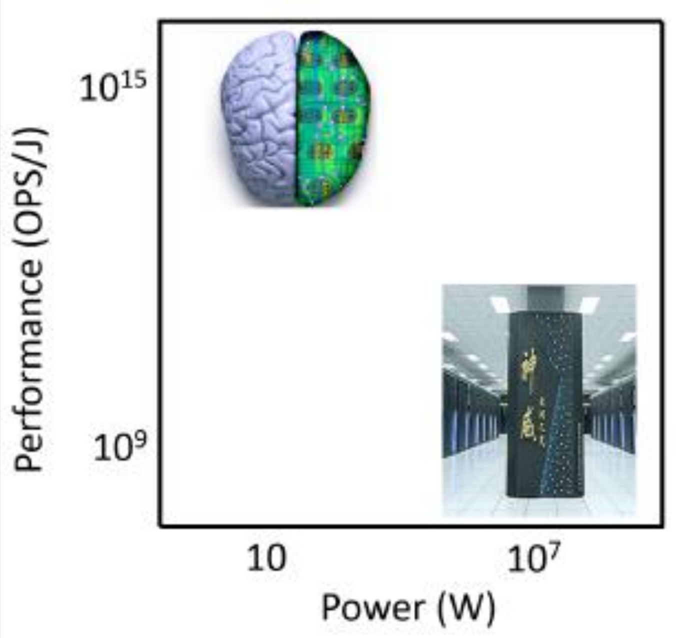


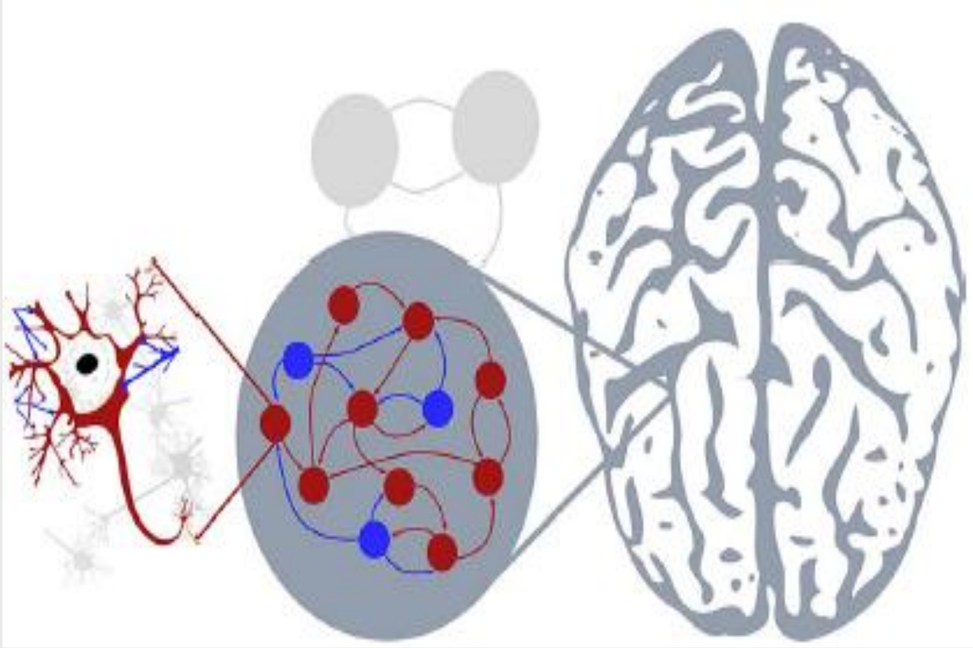
Shulaker et al., Nature, 2017

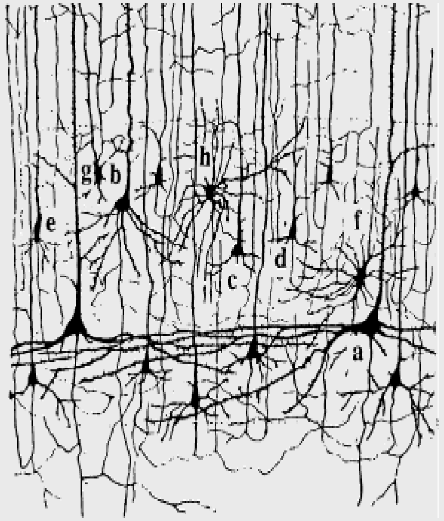
- Minimize the time and distance to memory access

Going beyond von Neumann computing: Brain-inspired computing

An “existence proof” for an ultra-low power AI computer





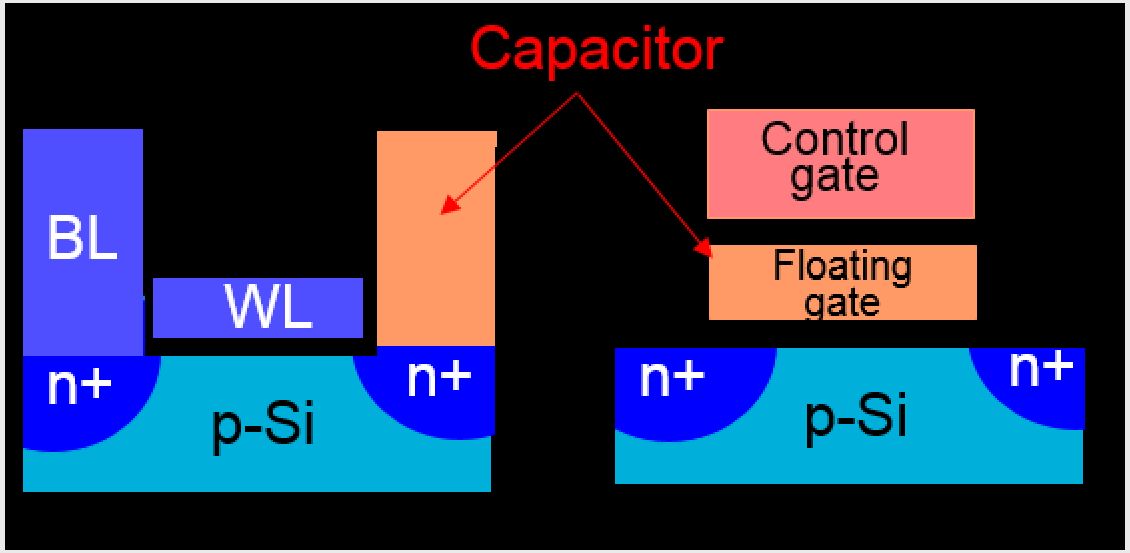


Ramón y Cajal

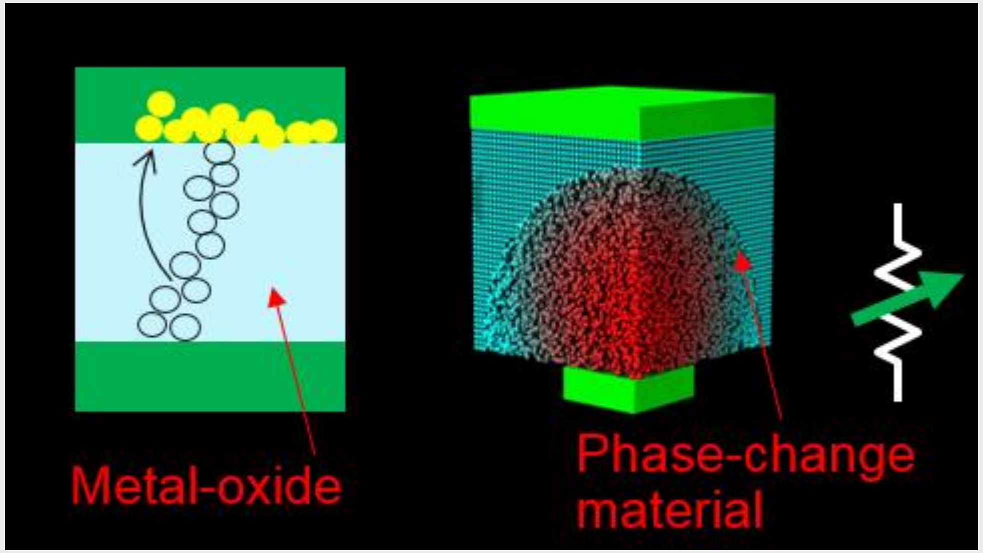
- Trades **accuracy for efficiency**
- Highly entwined, **collocated memory and processing**
- Computing fabric comprising large-scale networks of **neurons and synapses**
- **Spike-based** communication and processing of information

The role of memory

“Charge on a capacitor”

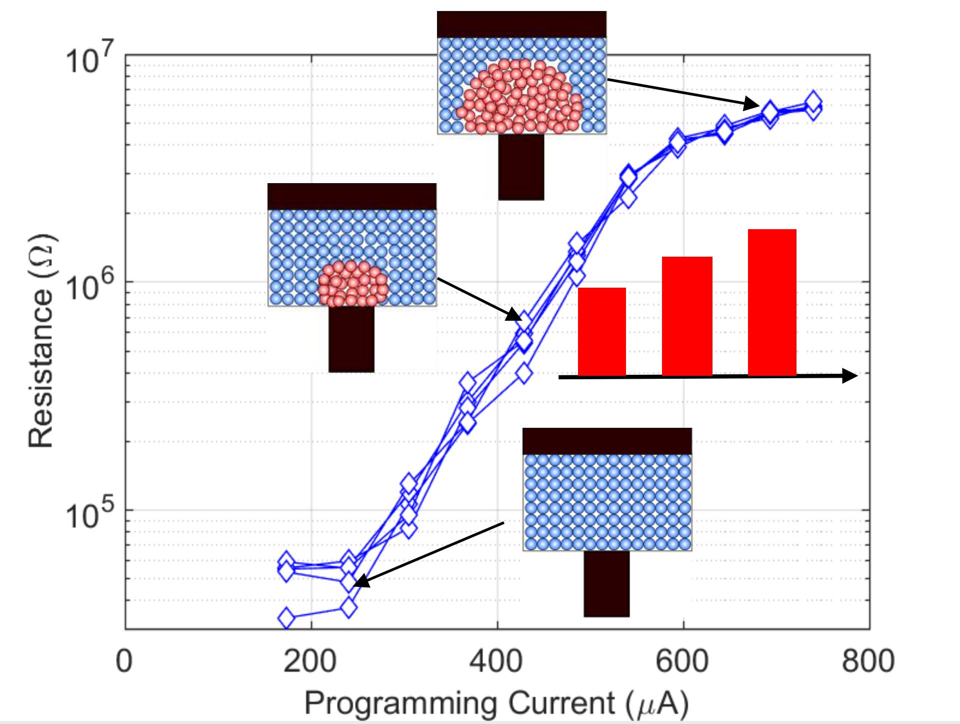


“Alternate atomic arrangements”

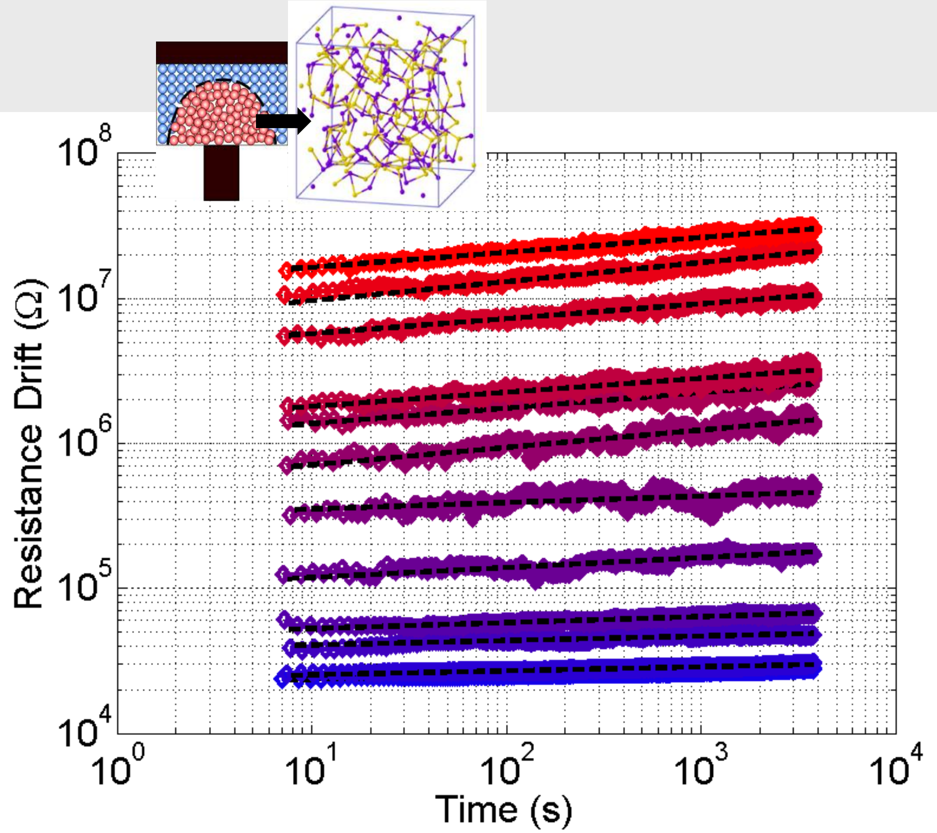


- Difference in atomic arrangements induced by the application of electrical pulses and measured as a difference in electrical resistance
- **Resistive memory devices** or “**memristive**” devices
- Based on physical mechanisms such as **ionic drift** and **phase transition**
- **Particularly well-suited for brain-inspired computing**

1st key enabler: Multi-level storage capability



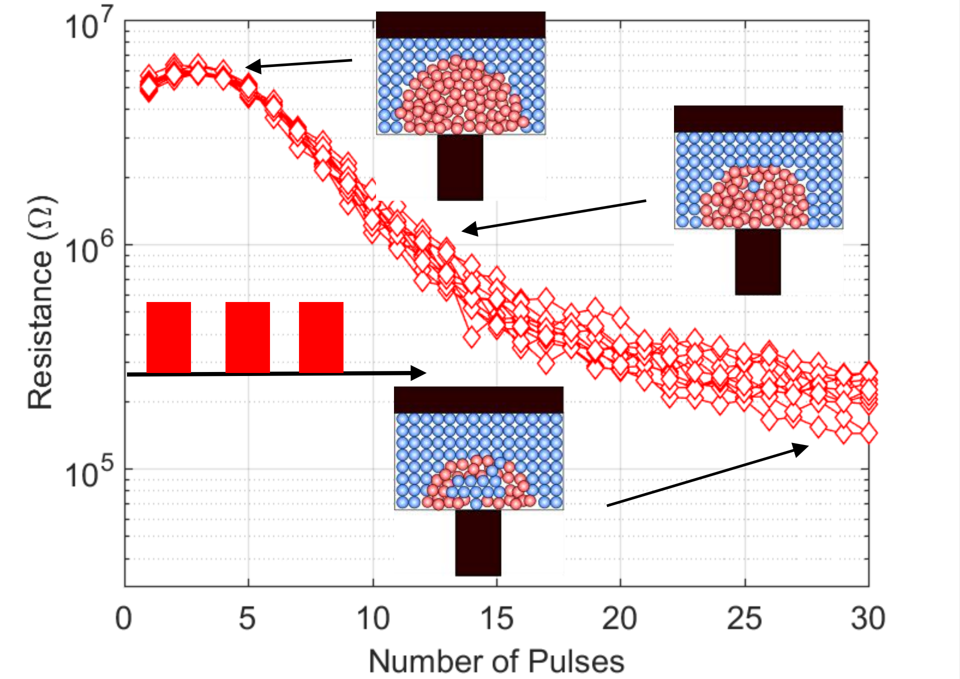
Sebastian et al., J. Appl. Phys., 2018



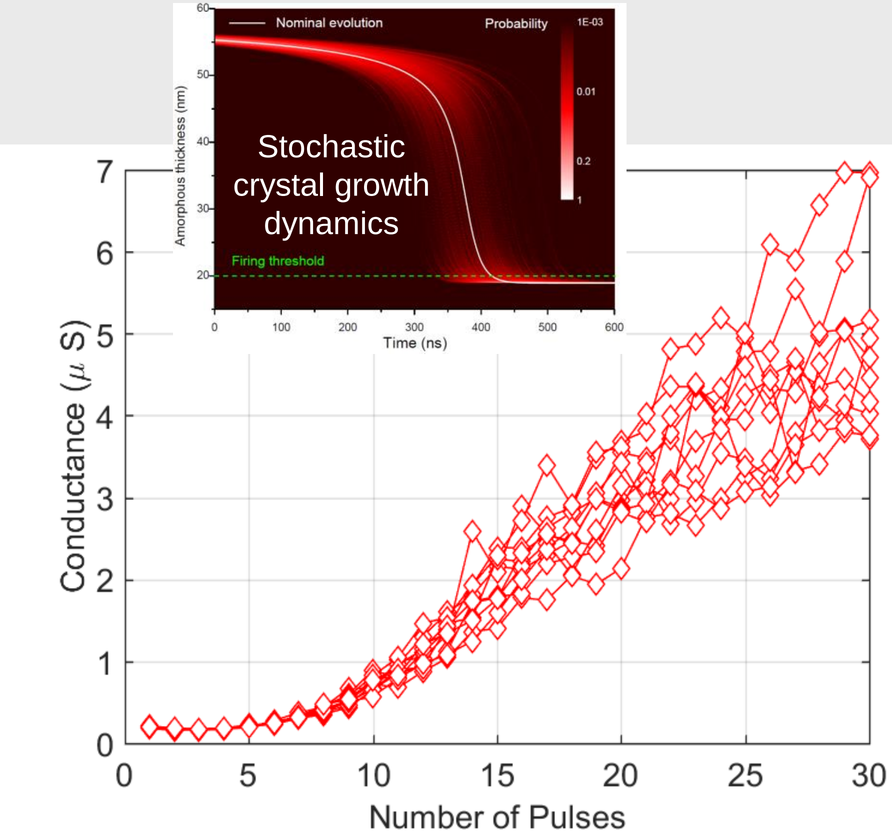
Le Gallo et al., Adv. Electr. Mat., 2018

- Essentially an analog storage device, but with **drift and noise**

2nd key enabler: Accumulative dynamics



Sebastian et al., Nature Comm., 2014

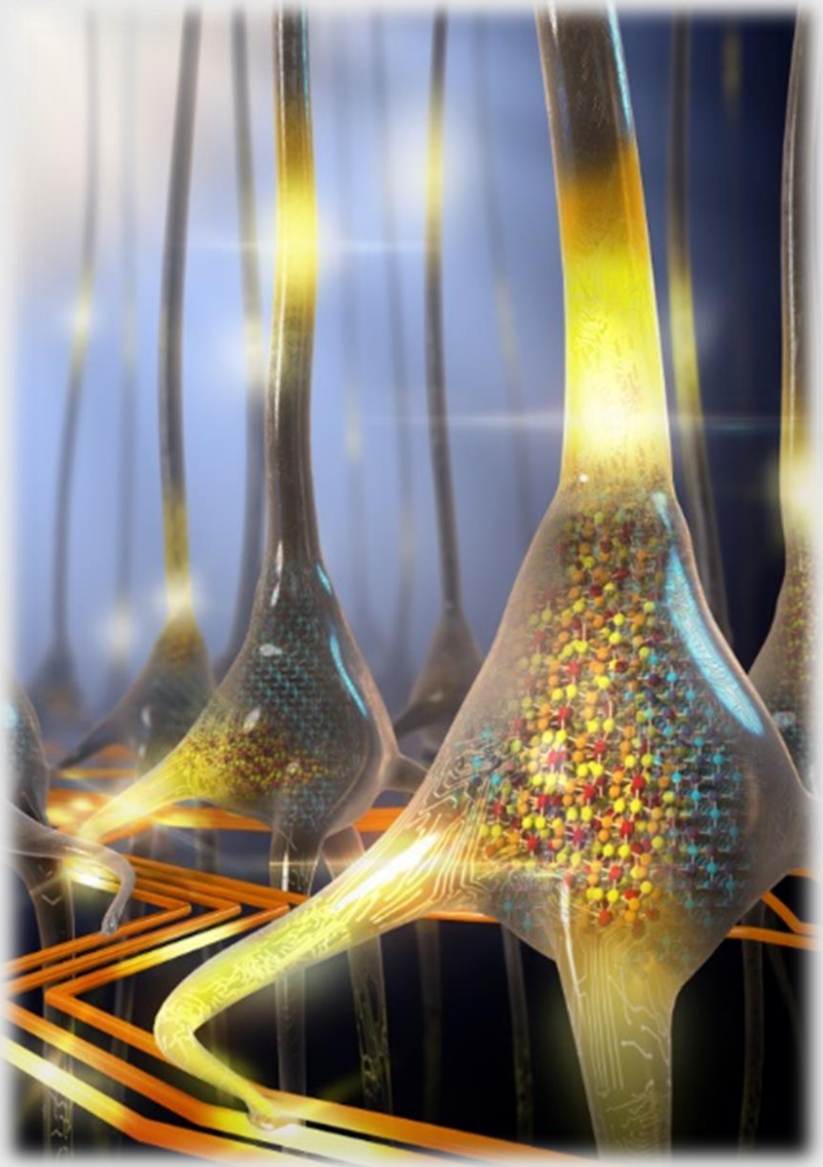


Le Gallo et al., ESSDERC, 2016

- Nonvolatile nanoscale integrator but **stochastic and nonlinear**

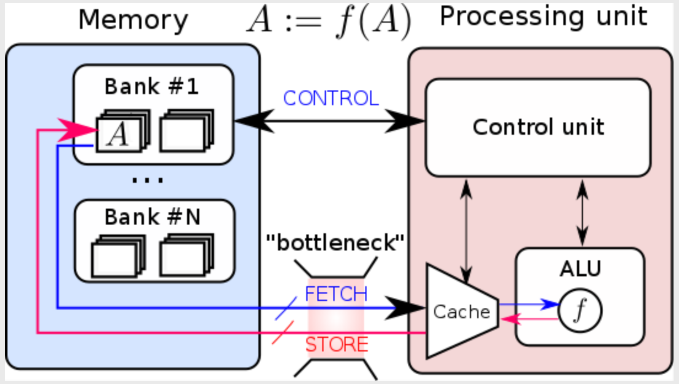
Outline

- **Introduction**
  - ✓ The computing efficiency problem of AI
  - ✓ Brain-inspired computing and the role of memory
  - ✓ Key enablers for brain-inspired computing
- **First level of inspiration: In-memory computing**
  - ✓ Matrix-vector multiplication and applications
  - ✓ Computing with device dynamics
- **Second level of inspiration: Co-processors for deep learning**
  - ✓ Mixed-precision deep learning
- **Third level of inspiration: Spiking neural networks**
  - ✓ Neuronal and synaptic emulations
  - ✓ Unsupervised learning
- **Summary & Outlook**

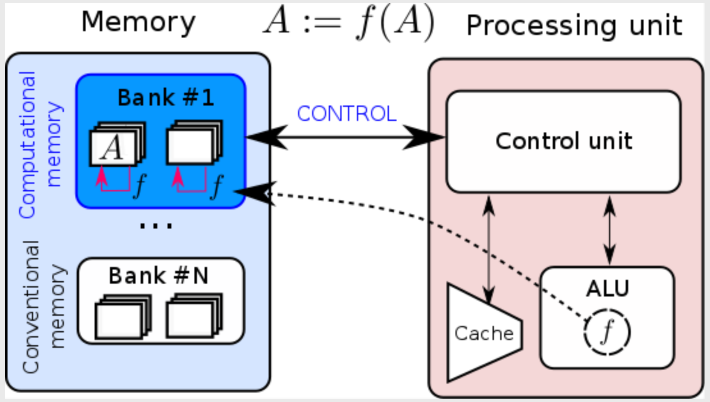


In-memory computing

Processing unit & Conventional memory



Processing unit & Computational memory



- Perform “certain” computational tasks using “certain” memory cores/units without the need to shuttle data back and forth in the process
  - ✓ Logical operations
  - ✓ Arithmetic operations
  - ✓ Machine learning algorithms
- Exploits the physical attributes and state dynamics of the memory devices

Hosseini et al., Electr. Dev. Lett., 2015
Sebastian et al., Nature Comm., 2017
Le Gallo et al., Nature Electronics, 2018

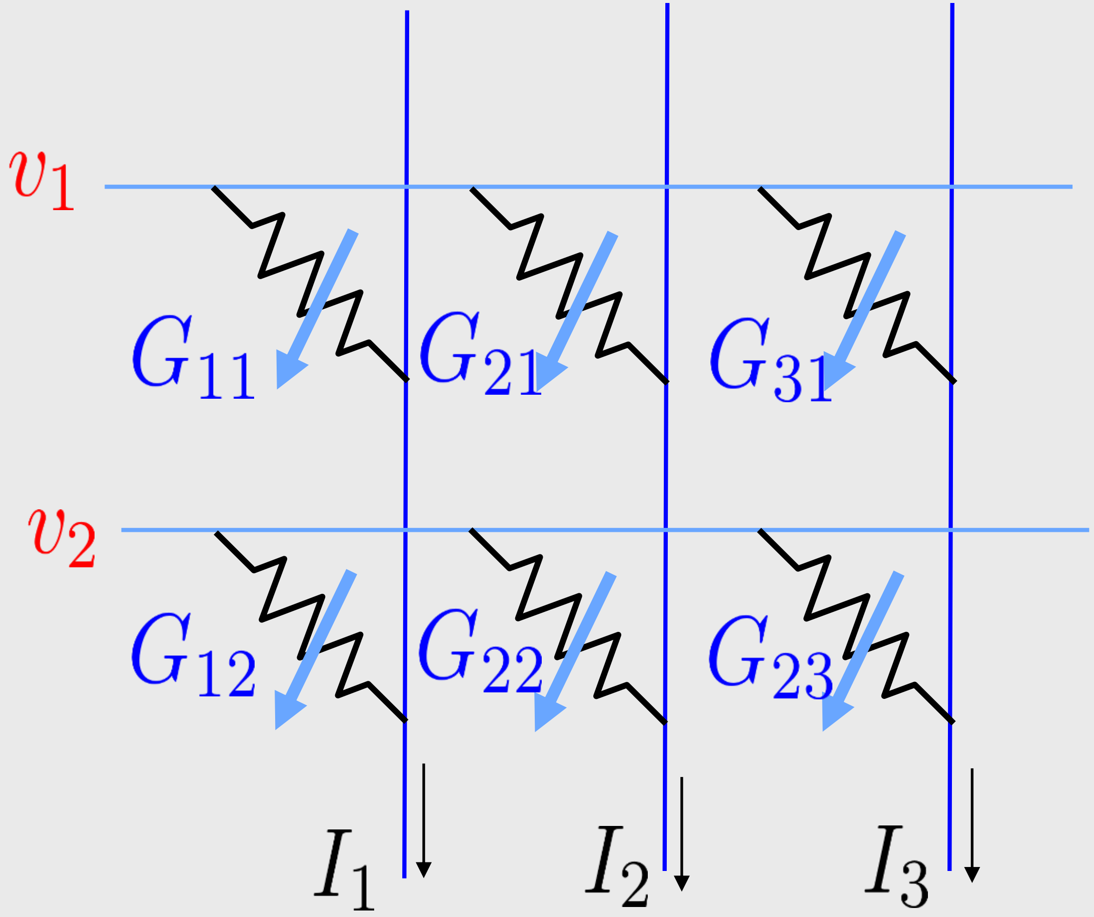
Matrix-vector multiplication

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

MAP to conductance values

MAP to read voltage

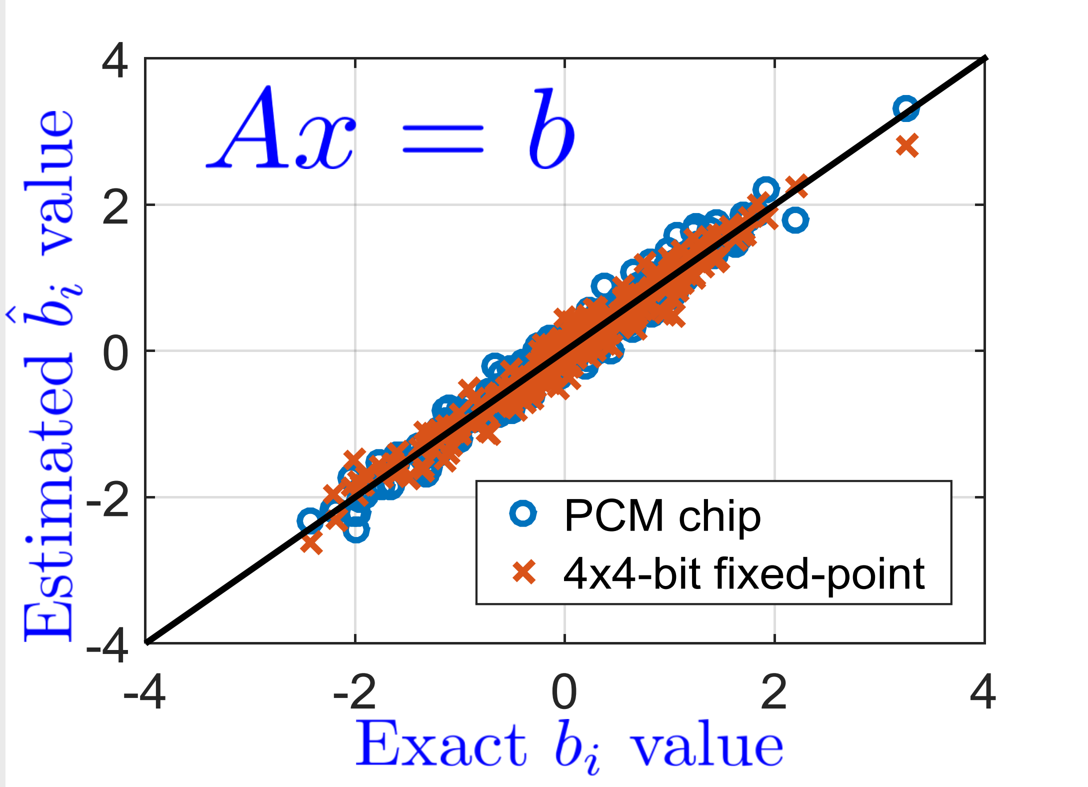
DECIPHER from the current



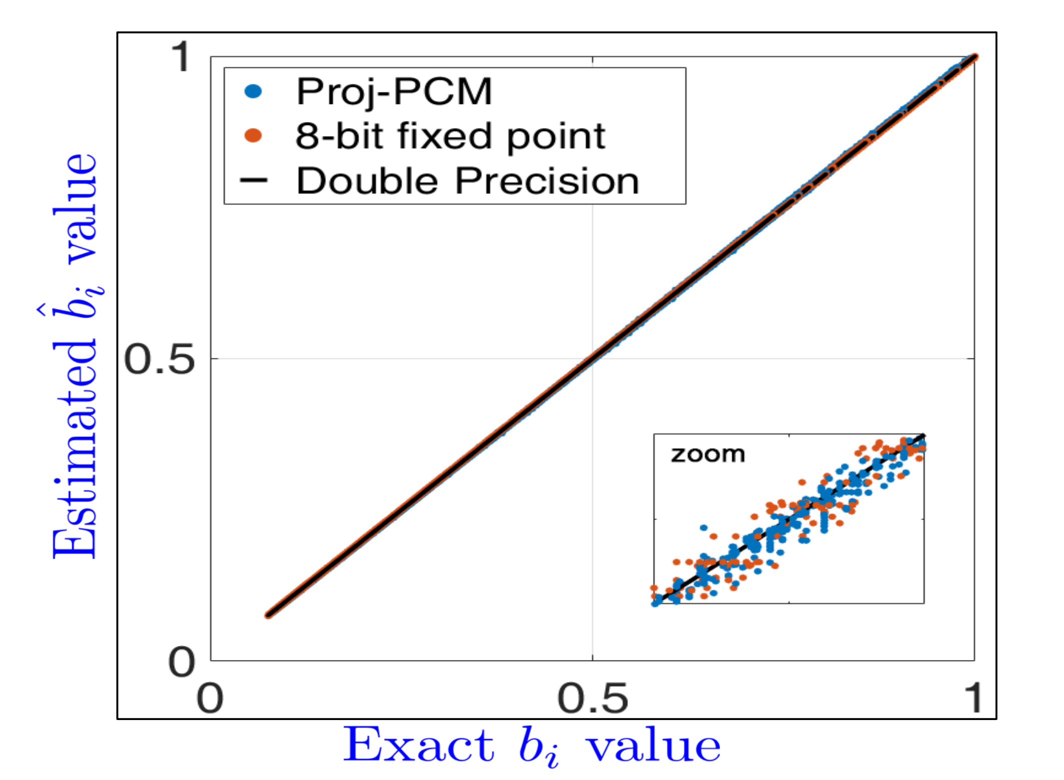
Burr et al., Adv. Phys. X, 2017
Le Gallo et al., Nature Electronics, 2018

- By arranging the resistive memory devices in a cross-bar configuration, one can perform matrix-vector operation with **O(1) time complexity**
- Exploits multi-level storage capability and Kirchhoff’s circuits laws
- Can also implement multiplication with the **matrix transpose**

Matrix-vector multiplication



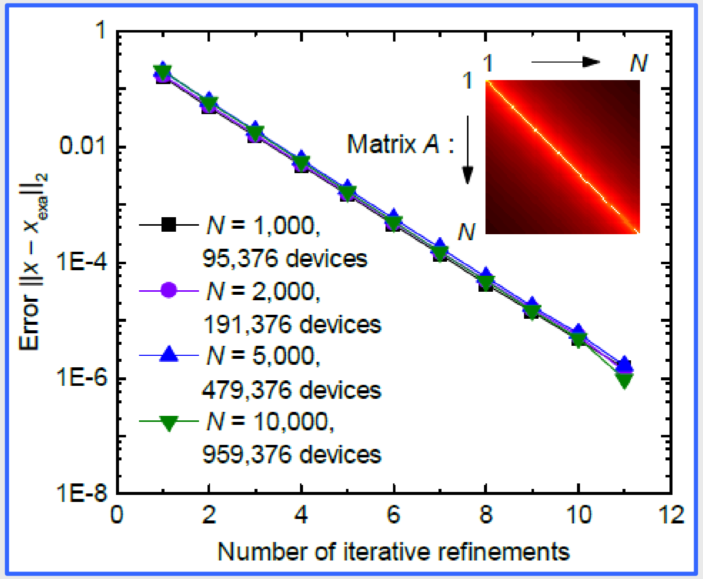
Le Gallo et al., Proc. IEDM, 2017

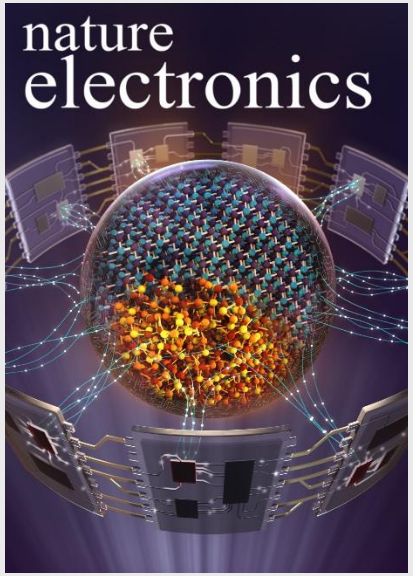


Giannopoulos et al., Proc. IEDM, 2018

Applications

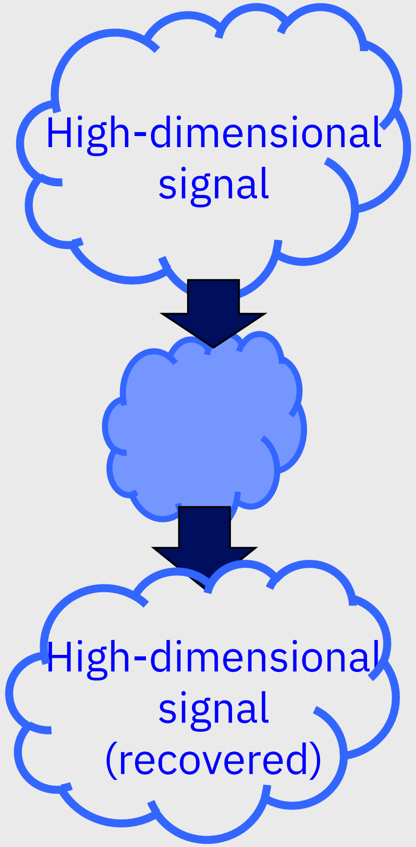
Solving systems of linear equations





Le Gallo et al., Nature Electronics, 2018

Compressed sensing and recovery

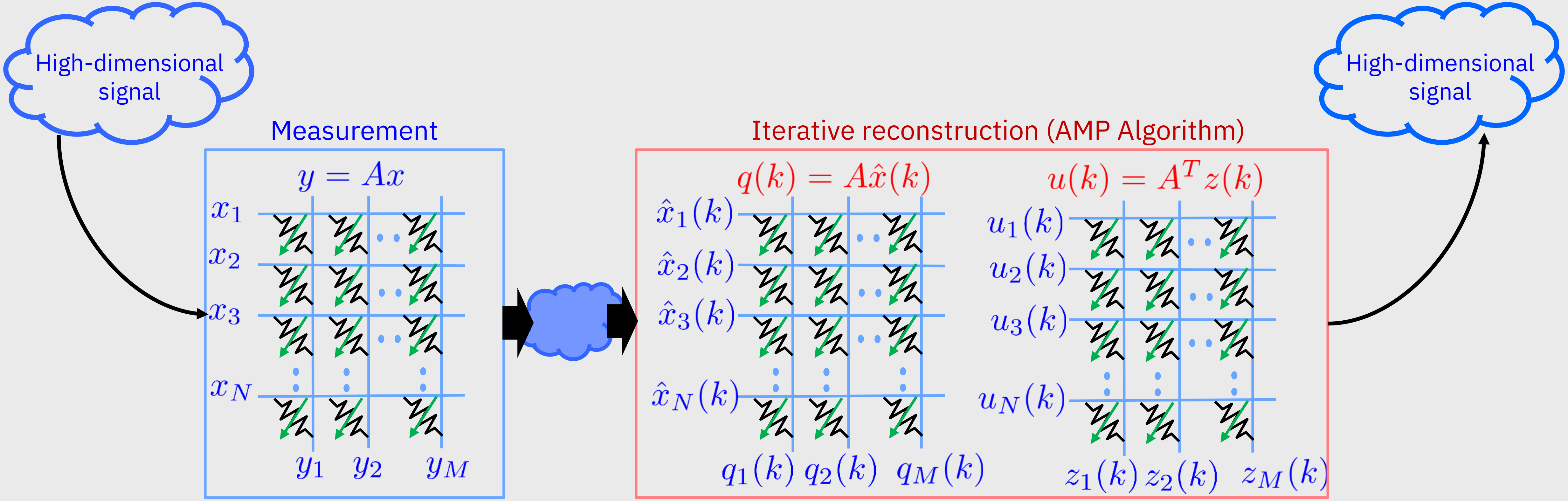


High-dimensional signal

High-dimensional signal (recovered)

Le Gallo et al., Proc. IEDM, 2017
Le Gallo et al., IEEE Trans. Electr. Dev., 2018

Compressed sensing and recovery



High-dimensional signal

Measurement

$y = Ax$

Iterative reconstruction (AMP Algorithm)

$q(k) = A\hat{x}(k)$

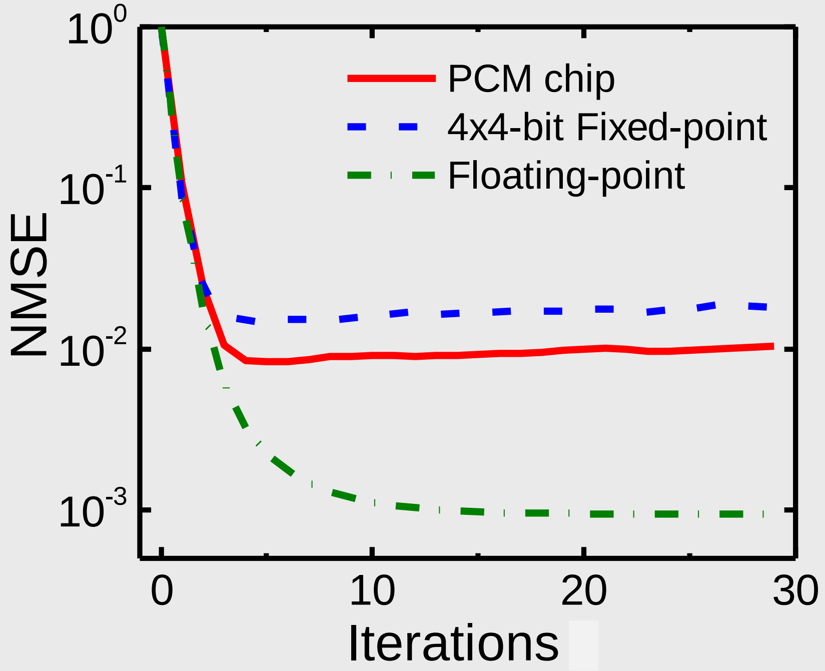
$u(k) = A^T z(k)$

High-dimensional signal

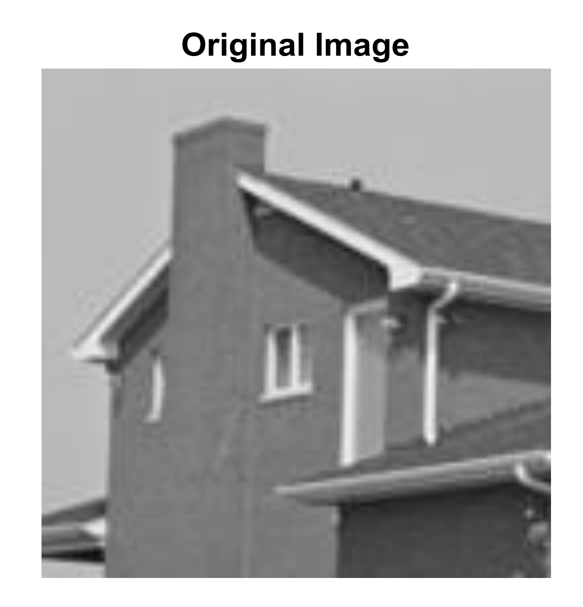
- Store the measurement matrix in a cross-bar array of resistive memory devices
- The same array used for both compression and reconstruction
- Reconstruction complexity reduction: O(NM) → O(N)

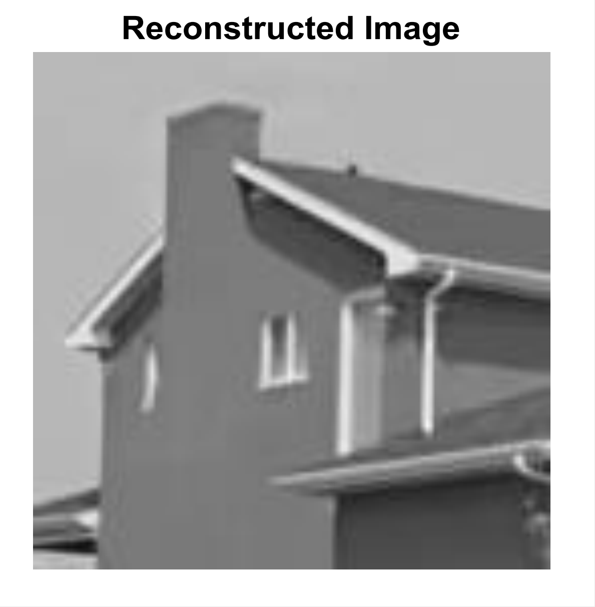
Le Gallo et al., Proc. IEDM, 2017
Le Gallo et al., IEEE Trans. Electr. Dev., 2018

Compressed sensing and recovery



Experimental result: 128X128 image, 50% sampling rate, Computation memory unit with 131,072 PCM devices

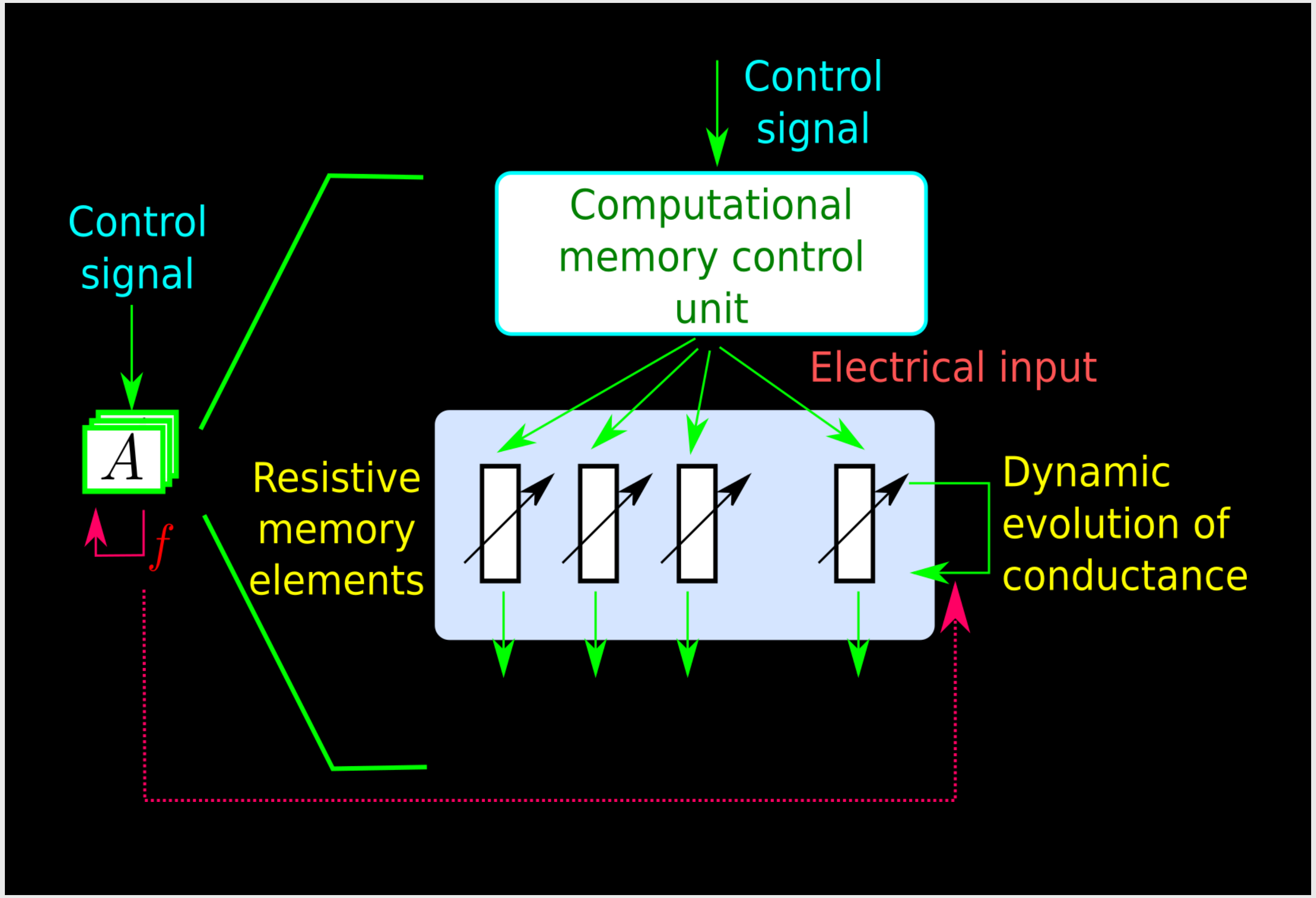


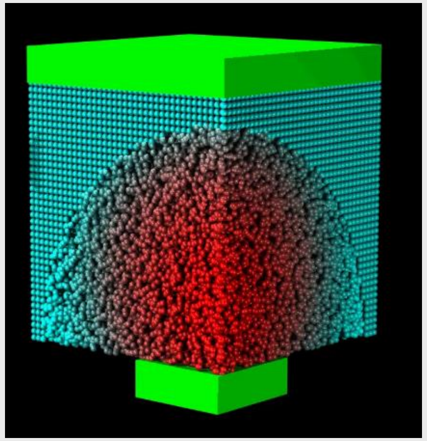


- Estimated power reduction of 50x compared to using an optimized 4-bit FPGA matrix-vector multiplier that delivers same reconstruction accuracy at same speed

Le Gallo et al., Proc. IEDM, 2017
Le Gallo et al., IEEE Trans. Electr. Dev., 2018

Can we compute with device dynamics?

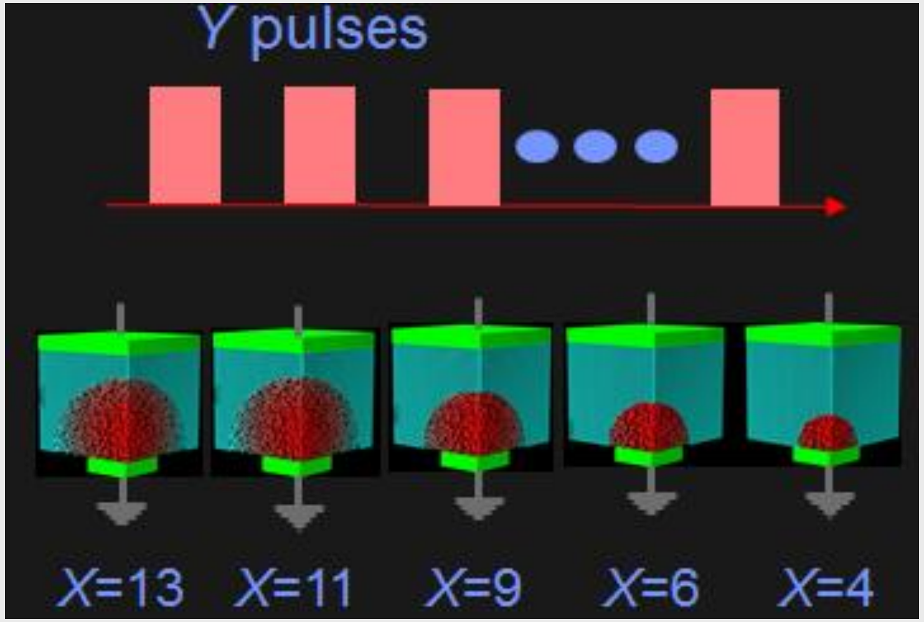

Control signal
Computational memory control unit
Control signal
Electrical input
A
f
Resistive memory elements
Dynamic evolution of conductance

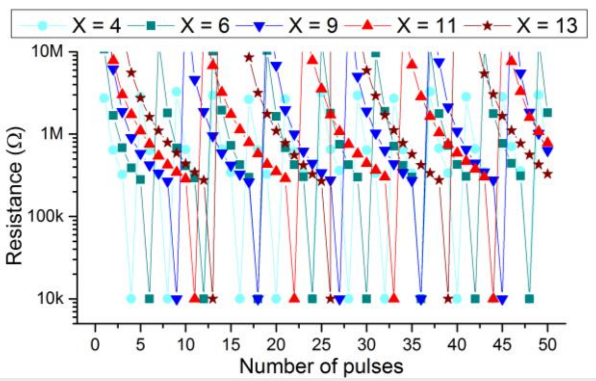


Sebastian et al., Nature Communications, 2014
Sebastian et al., Nature Communications, 2017

Applications

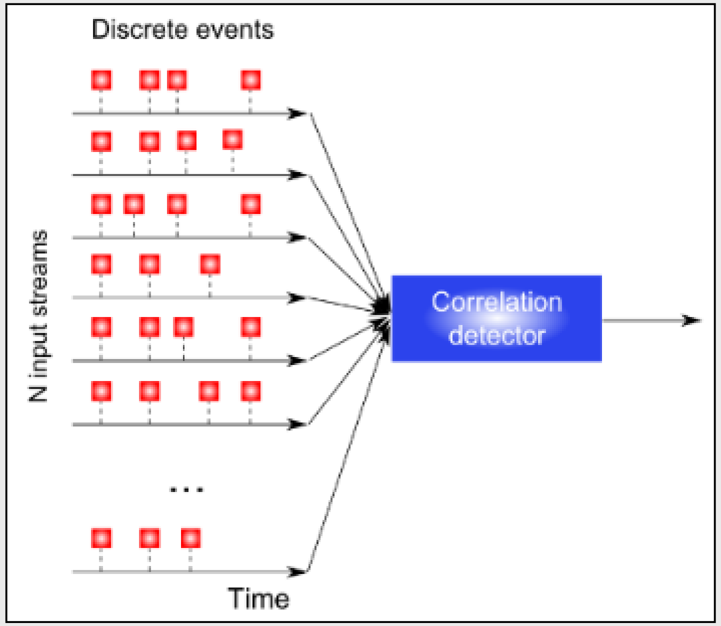
Finding factors in parallel





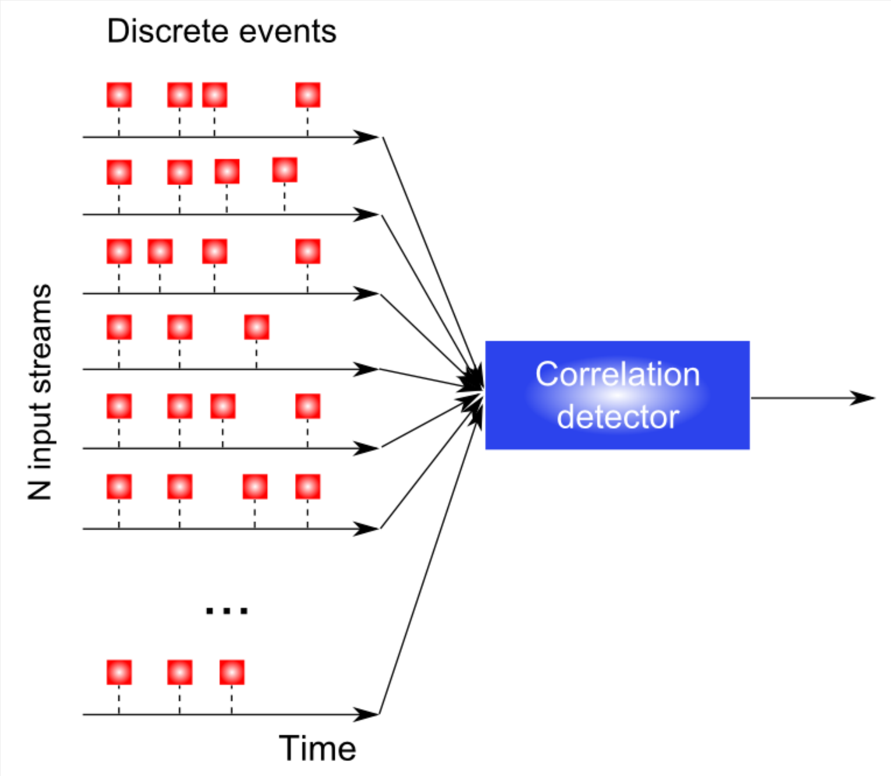
Hosseini et al., Electr. Dev. Lett., 2015

Detecting temporal correlations



Sebastian et al., Nature Comm., 2017

Detecting temporal correlations



- Find temporal correlations between event-based data streams in an **unsupervised manner**
- Gain selectivity specifically to the correlated inputs
- Observe **variations in the activity** of the correlated input
- Quickly react to occurrence of coincident inputs in the correlated inputs
- **Continuously and dynamically re-evaluate** the learned statistics

FINANCE

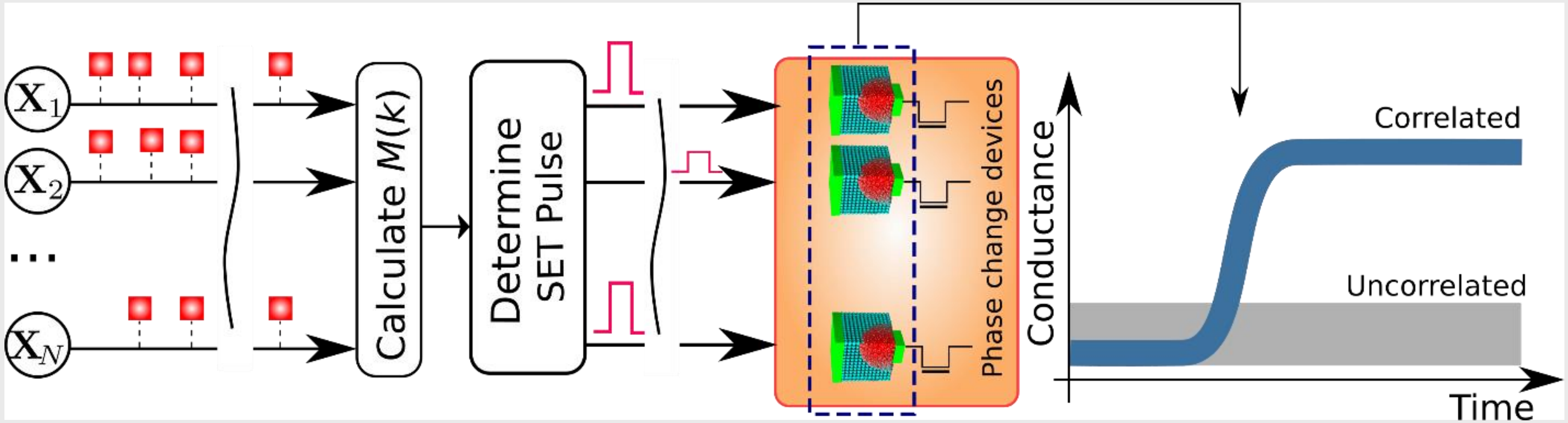
SCIENCE

MEDICINE

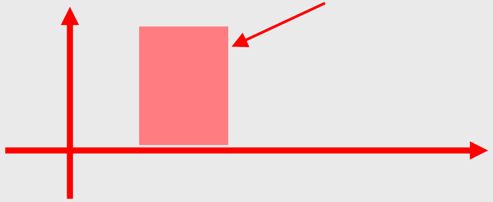
BIG DATA

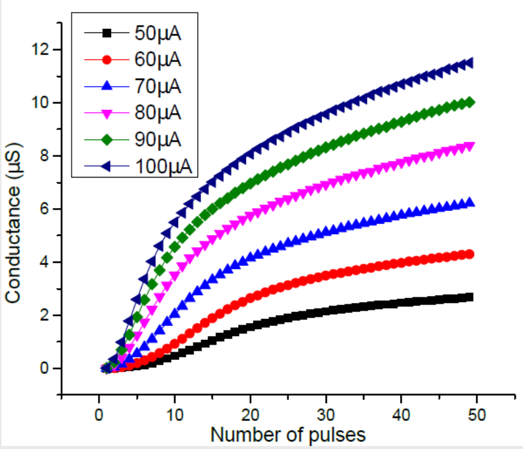
...AND MORE

Detecting temporal correlations



Modulate the amplitude based on $M(k) = \sum_{j=1}^{N} X_j(k)$

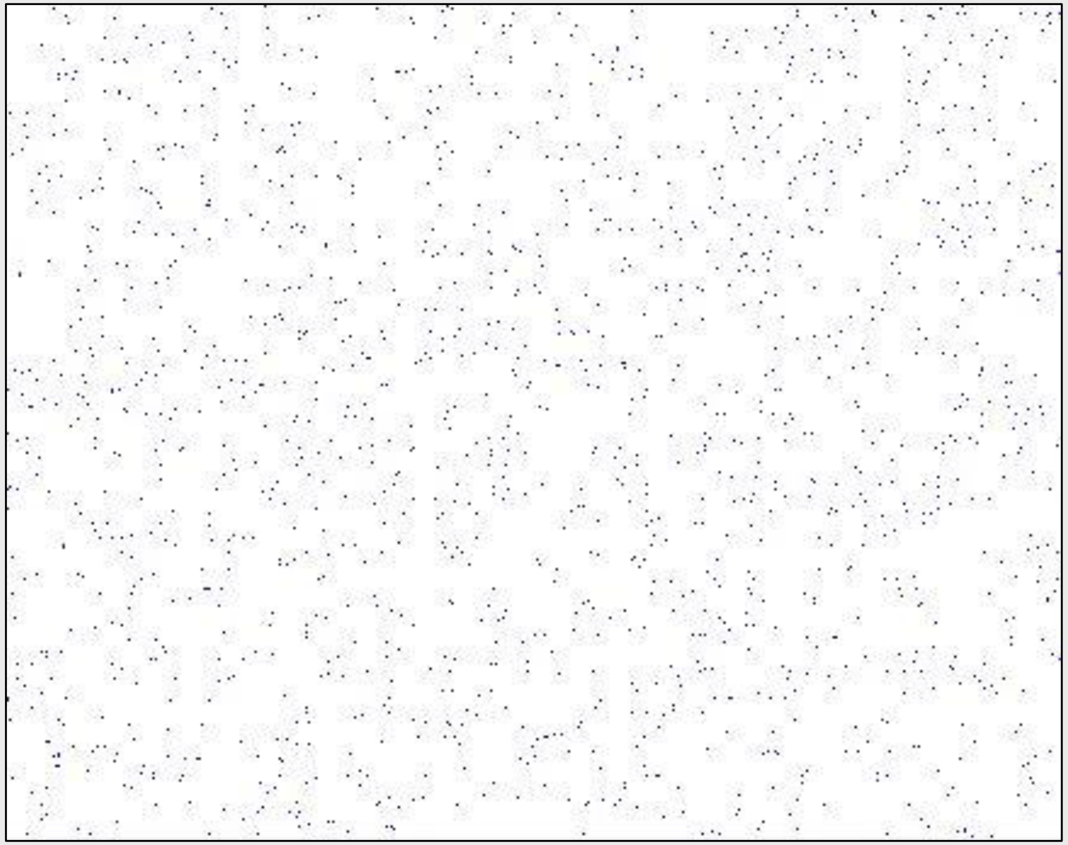


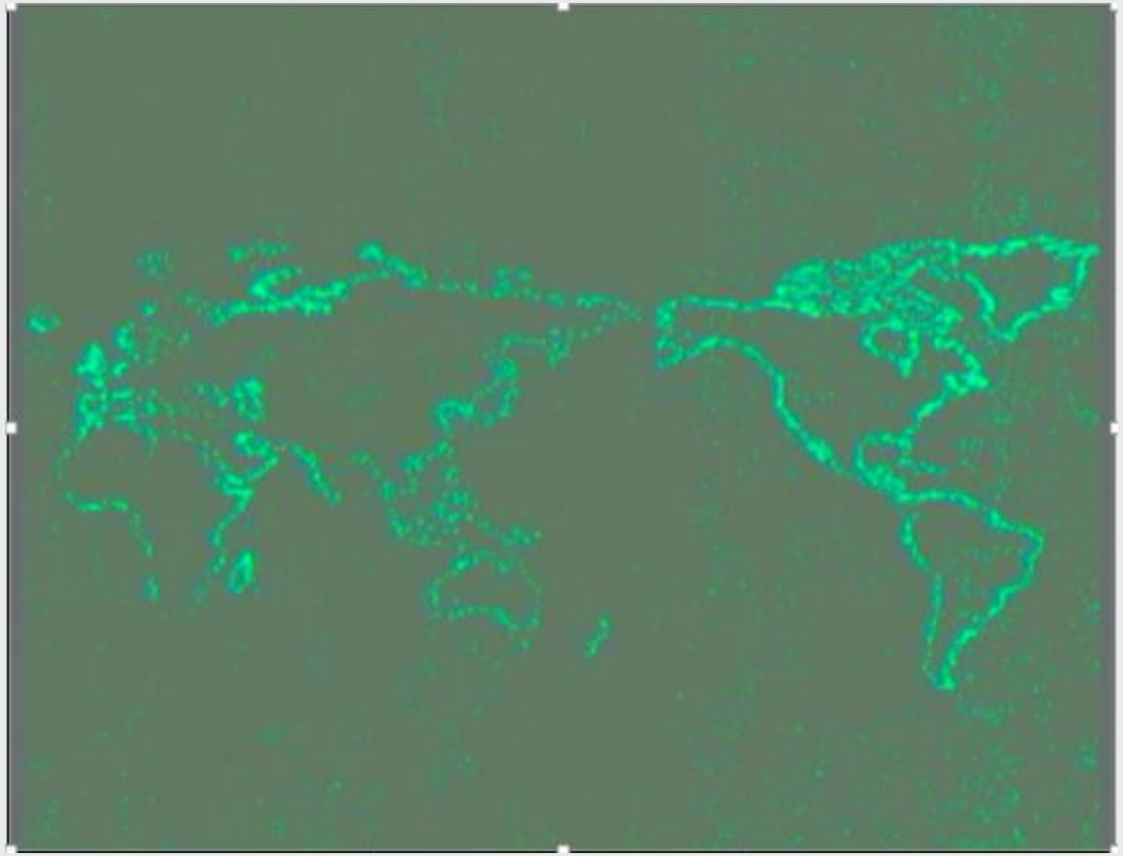


$$
\begin{aligned}
\Delta u_{a_i}(K) &= \sum_{k=1}^{K} \delta u_{a_i}(k) X_i(k) \\
&= C\mathscr{G} \sum_{k=1}^{K} \sum_{j=1}^{N} X_i(k) X_j(k) \\
&= C\mathscr{G} \sum_{j=1}^{N} \sum_{k=1}^{K} X_i(k) X_j(k) \\
&= KC\mathscr{G} \sum_{j=1}^{N} \hat{R}_{ij} \\
&= KC\mathscr{G} \hat{W}_i .
\end{aligned}
$$

Sebastian et al., Nature Comm., 2017

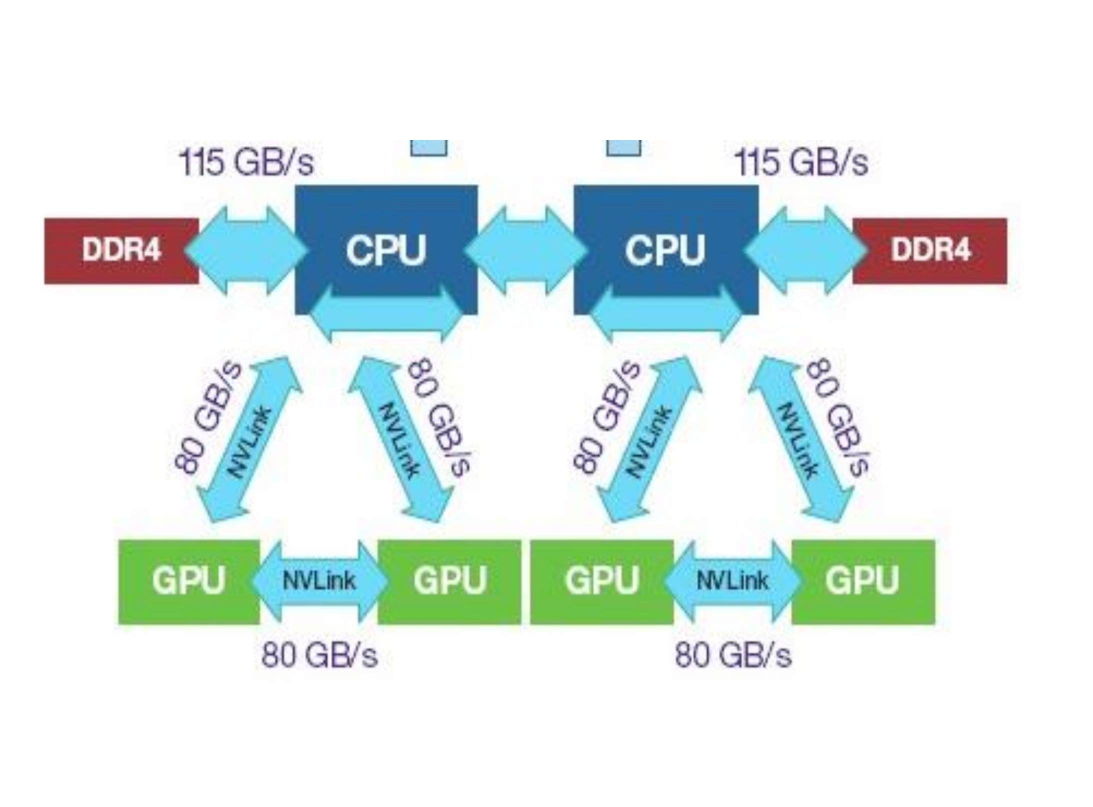
Detecting temporal correlations: Experiments (1 Million PCM devices)

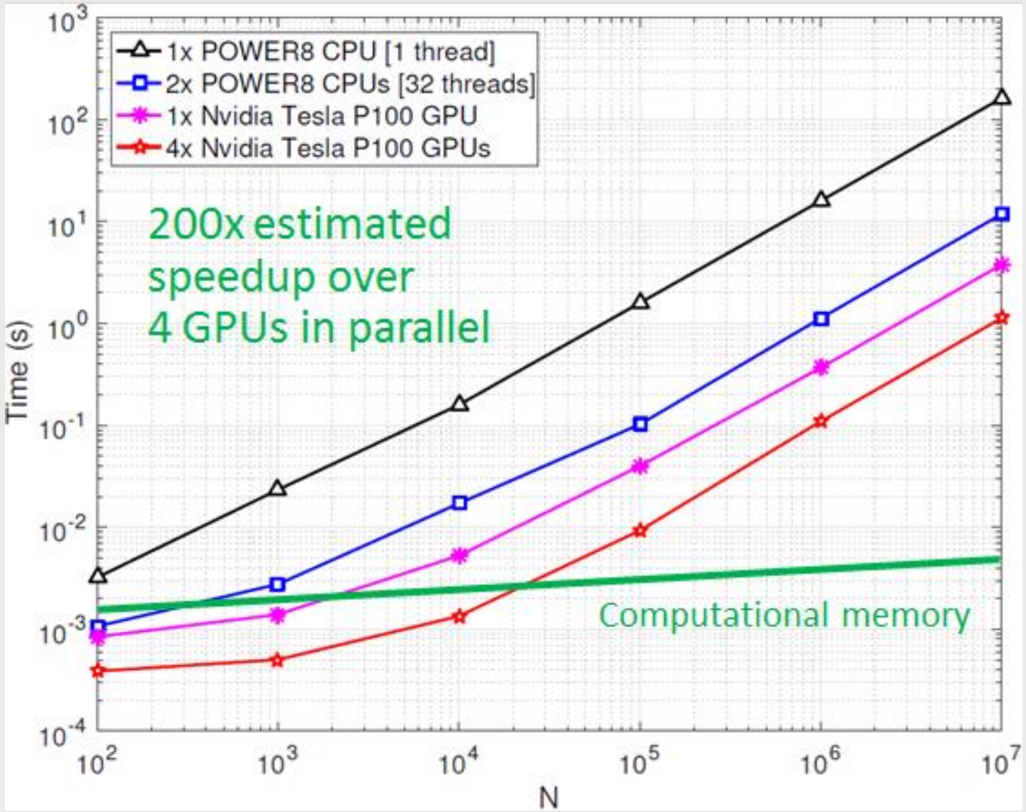




Sebastian et al., Nature Comm., 2017

Detecting temporal correlations: Comparative study

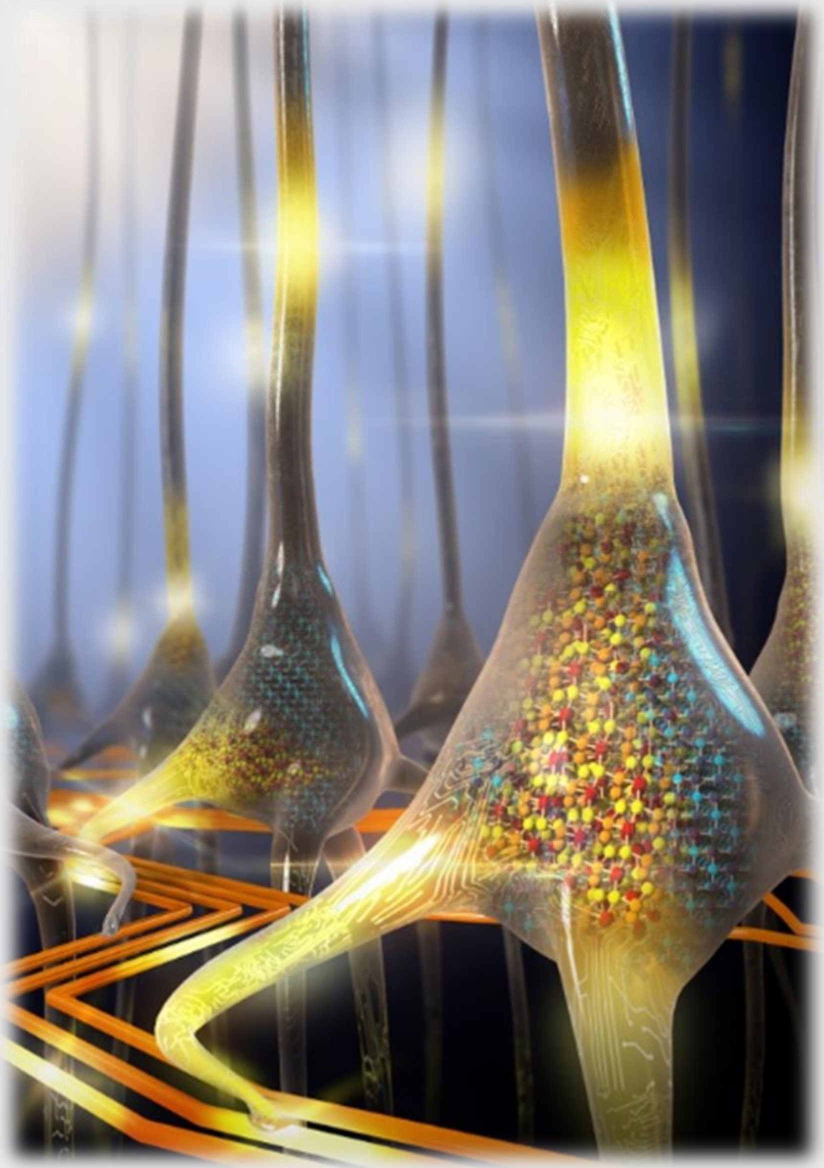




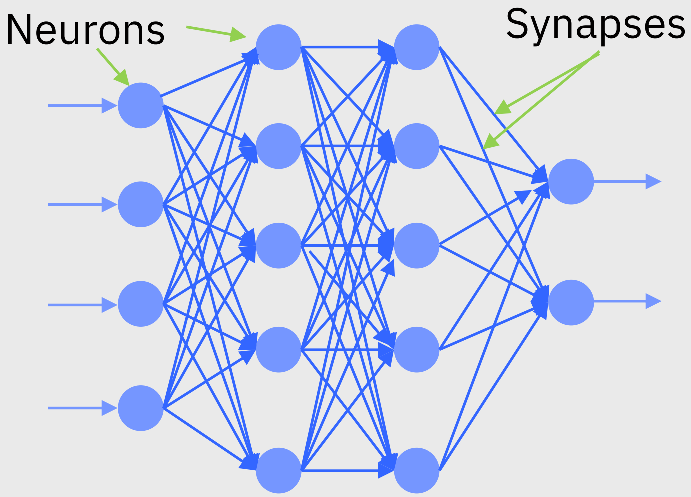
Sebastian et al., Nature Comm., 2017

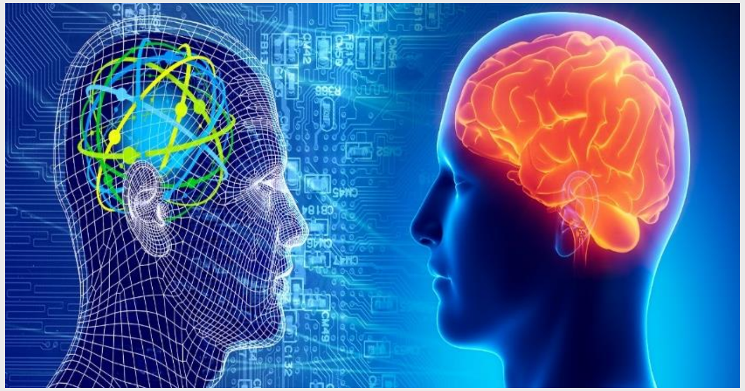
Outline

- **Introduction**
  - ✓ The computing efficiency problem of AI
  - ✓ Brain-inspired computing and the role of memory
  - ✓ Key enablers for brain-inspired computing
- **First level of inspiration: In-memory computing**
  - ✓ Matrix-vector multiplication and applications
  - ✓ Computing with device dynamics
- **Second level of inspiration: Co-processors for deep learning**
  - ✓ Mixed-precision deep learning
- **Third level of inspiration: Spiking neural networks**
  - ✓ Neuronal and synaptic emulations
  - ✓ Unsupervised learning
- **Summary & Outlook**



Co-processors for deep neural networks

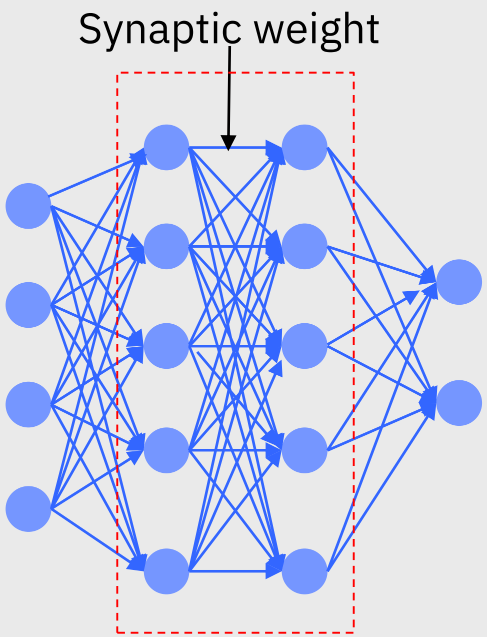


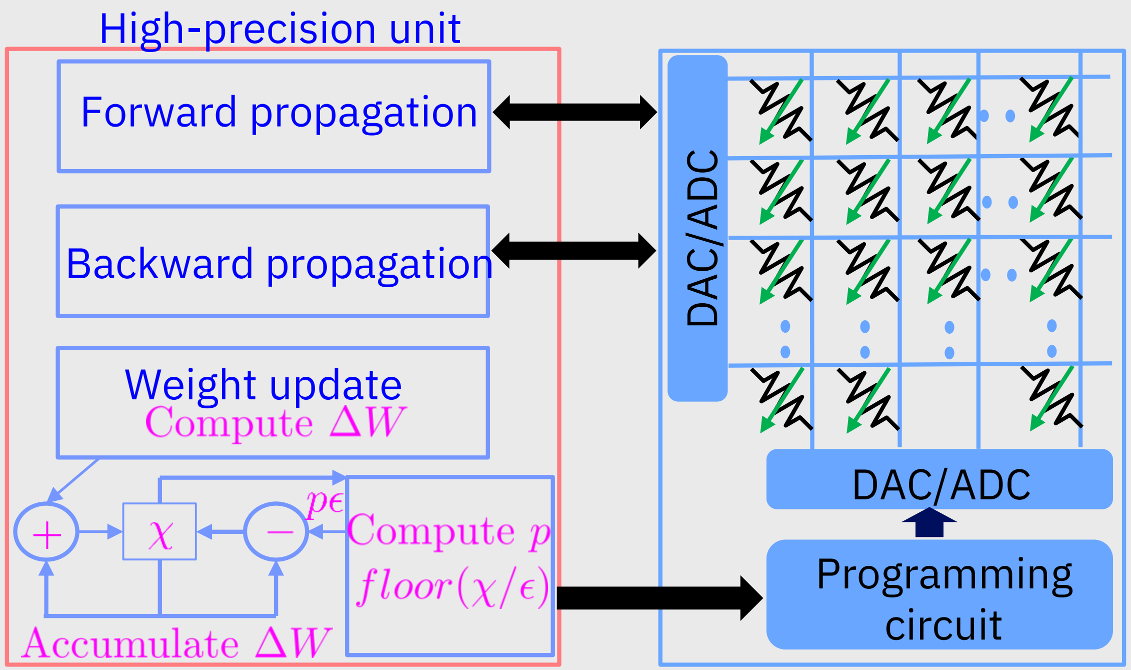


- Multiple layers of parallel processing units (neurons) interconnected by plastic synapses
- By tuning the synaptic weights (training), able to solve certain classification tasks remarkably well
- Training based on a global supervised learning algorithm → **gradient descent with backpropagation**
- **Brute force optimization**: Multiple days or weeks to train state-of-the-art networks on von Neumann machines (CPU,GPU clusters)
- Can we design **non-von Neumann co-processors for training deep neural networks**?

Burr et al., IEEE TED, 2015
Nandakumar et al., ISCAS, 2018
Ambrogio et al., Nature, 2018

Mixed-precision deep learning

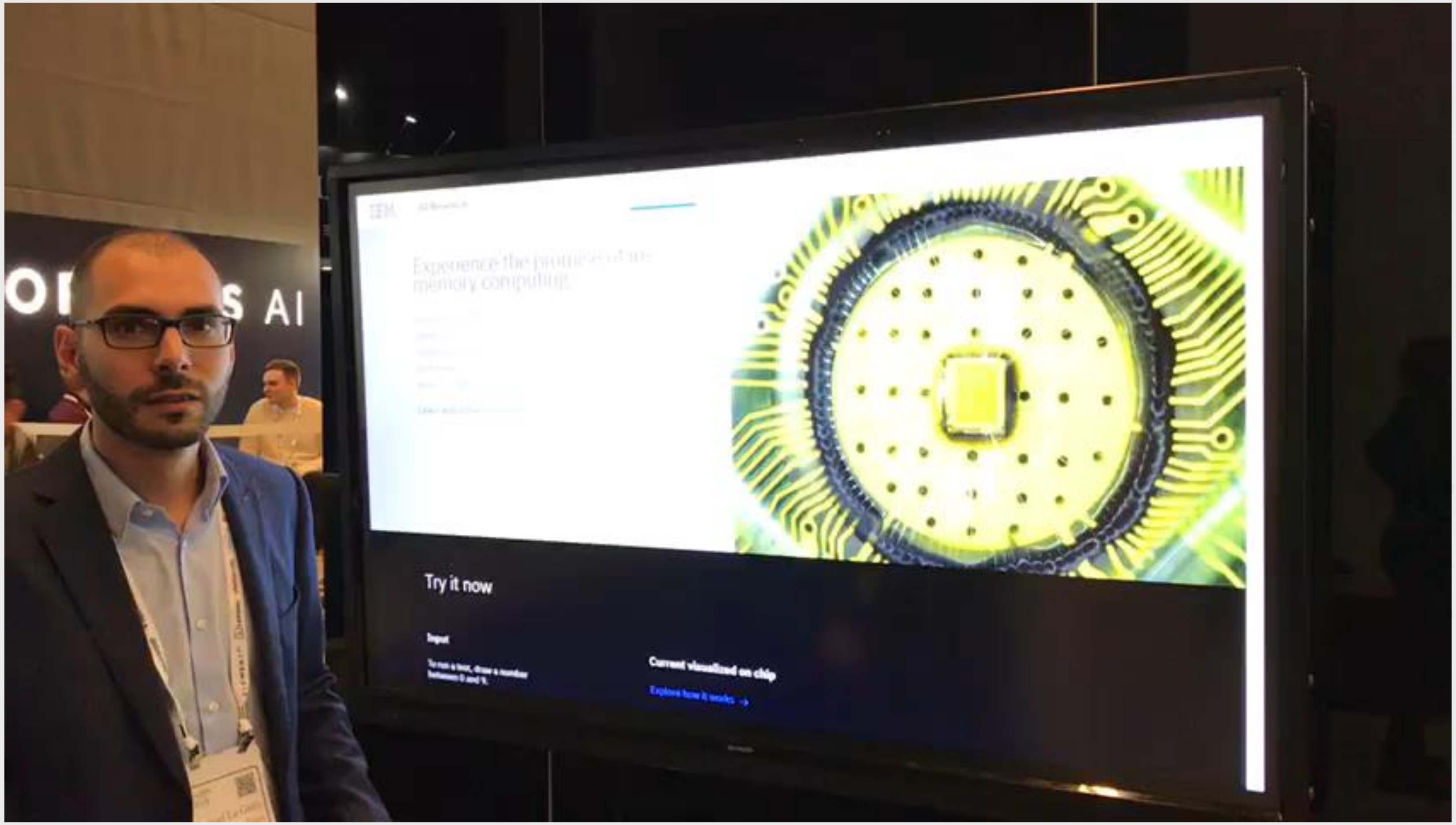




- **Synaptic weights always reside in the computational memory**
- **Forward/backward propagation** performed in place (with low precision)
- The desired weight updates **accumulated in high precision**
- Programming pulses issued to the memory devices to alter the synaptic weights
- **Exploits both multi-level storage capability and accumulative behavior!**

Nandakumar et al., ISCAS, 2018

Demo @ NeurIPS, Montreal, 2018

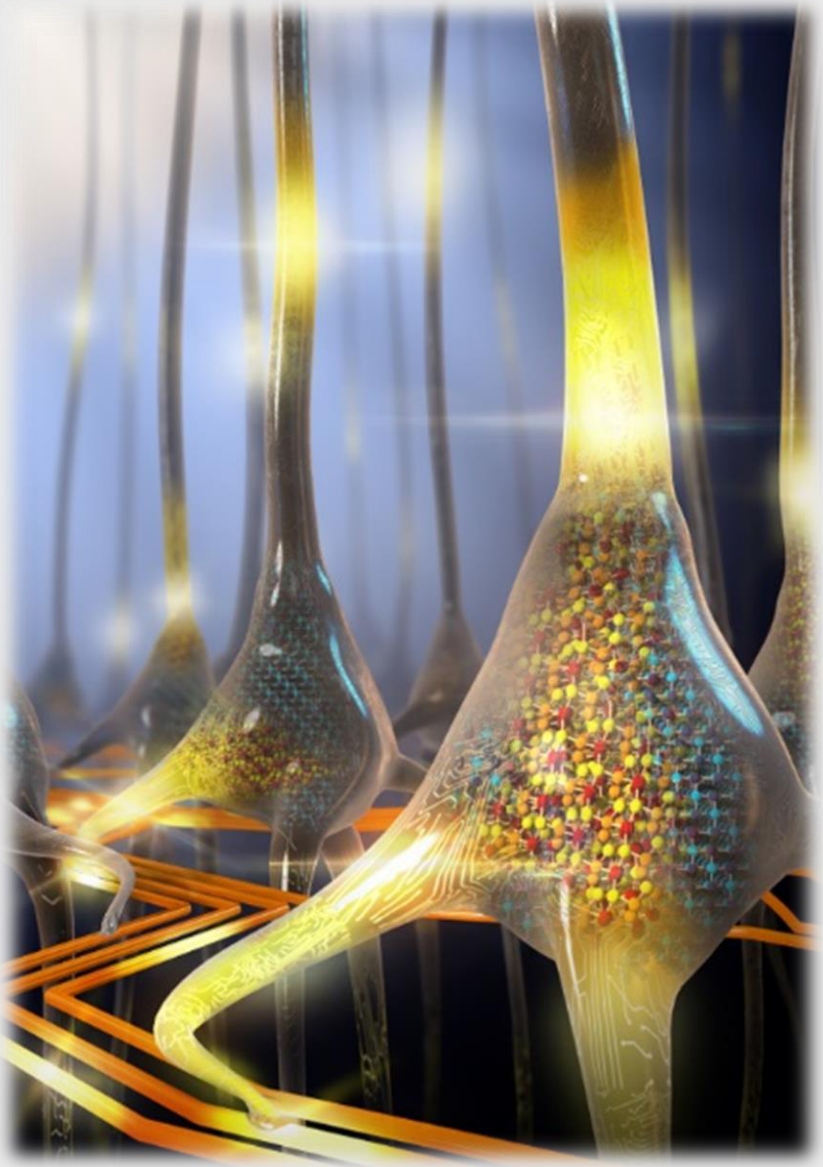

Experience the promise of in-memory computing
Try it now
Input
To run a test, draw a number between 0 and 9.
Current visualized on chip
Explore how it works →

Experience the promise of in-memory computing

https://analog-ai-demo.mybluemix.net/?cm_mc_uid=62608486854615522234476&cm_mc_sid_50200000=20414201553078943175

Outline

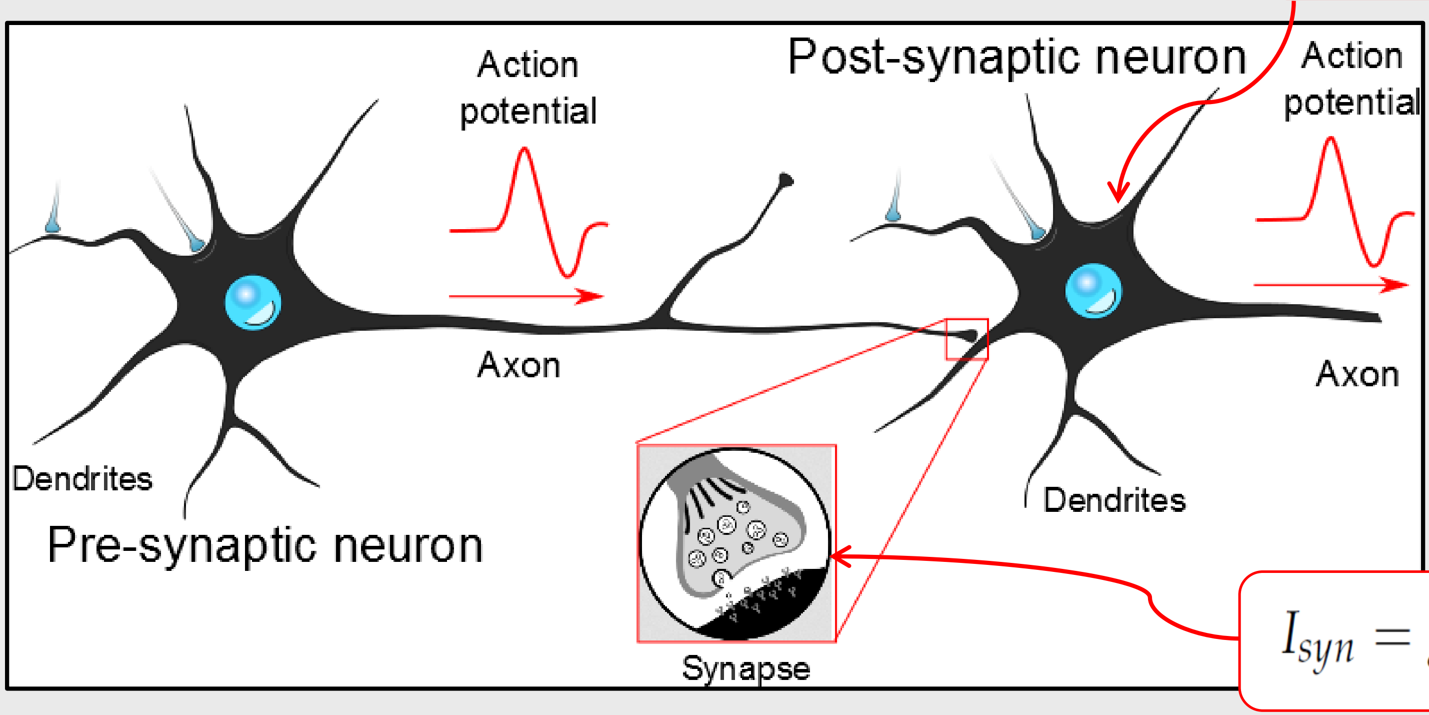
- **Introduction**
  - ✓ The computing efficiency problem of AI
  - ✓ Brain-inspired computing and the role of memory
  - ✓ Key enablers for brain-inspired computing
- **First level of inspiration: In-memory computing**
  - ✓ Matrix-vector multiplication and applications
  - ✓ Computing with device dynamics
- **Second level of inspiration: Co-processors for deep learning**
  - ✓ Mixed-precision deep learning
- **Third level of inspiration: Spiking neural networks**
  - ✓ Neuronal and synaptic emulations
  - ✓ Unsupervised learning
- **Summary & Outlook**



Spiking neural networks

Neuronal dynamics

$$\mathrm{d}u/\mathrm{d}t = F(u) + G(u)I$$



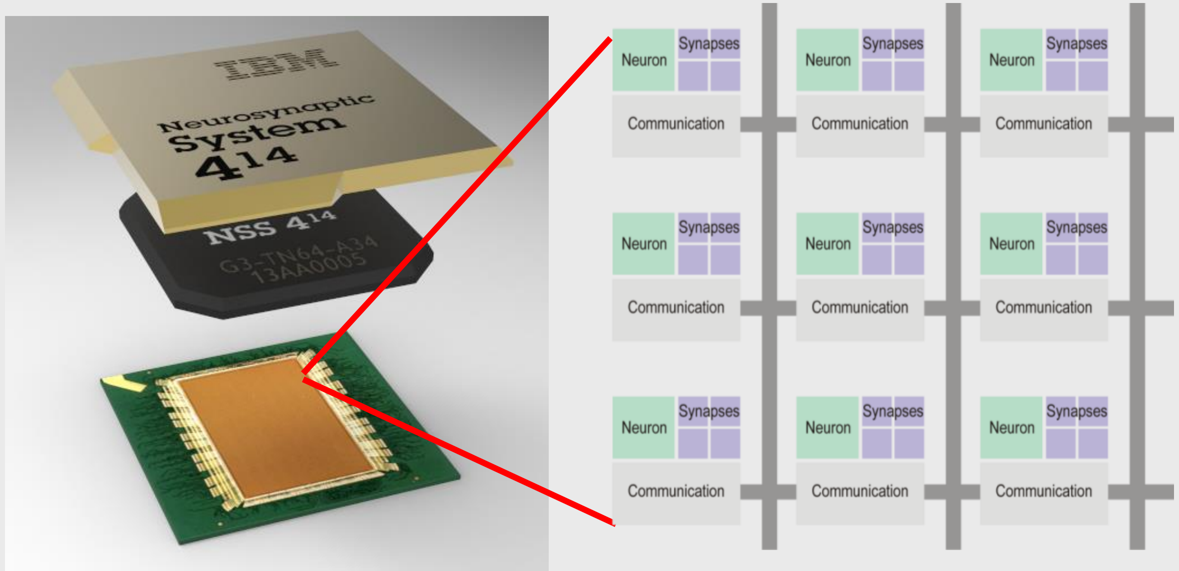
- Employed by the brain
- **Asynchronous, low-latency,** massively-distributed computation
- **Local, event-based learning**
- Continuously learning systems
- **Computationally superior?**

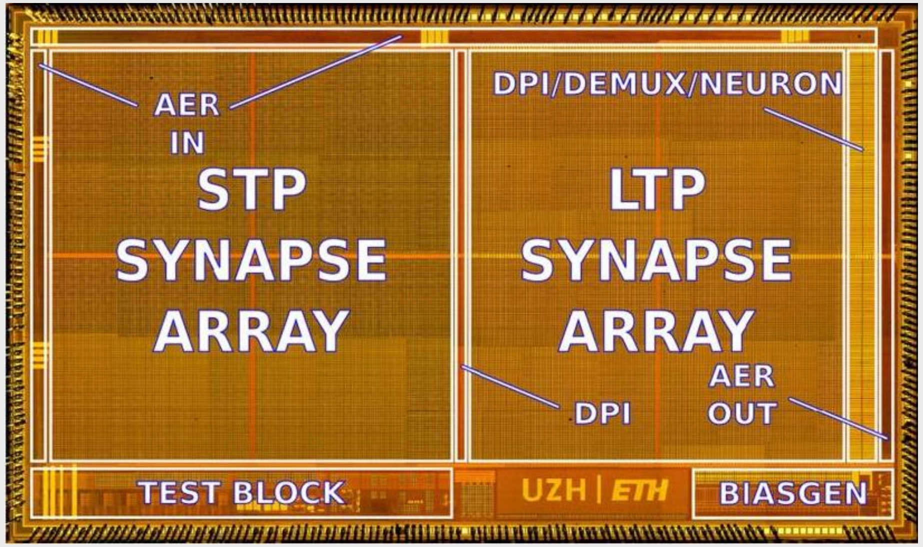
Synaptic dynamics

$$I_{syn} = g_{syn} S(V - E_{syn})$$

- Challenge 1: Learning rules and killer applications
- Challenge 2: Substrates for efficient realization: Emulate neuronal and synaptic dynamics

SNN co-processors (Digital and Analog CMOS-based)





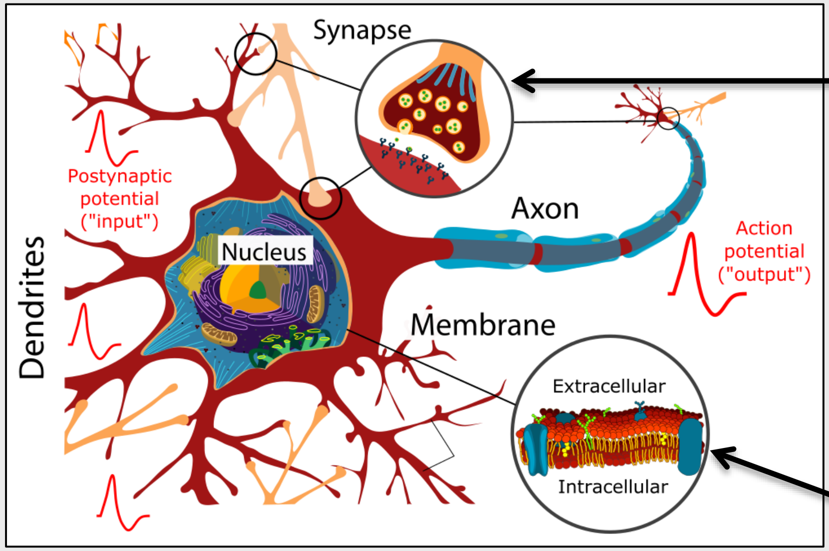
- Emulation of neuronal and synaptic dynamics in digital CMOS circuitry
- No in-situ learning

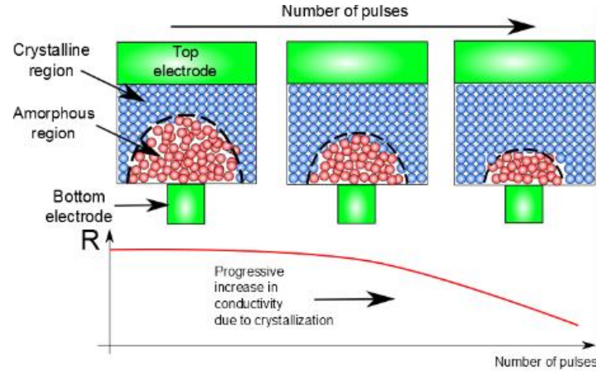
Merolla et al., Science, 2014

- Exploit subthreshold MOSFET characteristics to directly emulate neuronal and synaptic dynamics
- Highly susceptible to process induced variations

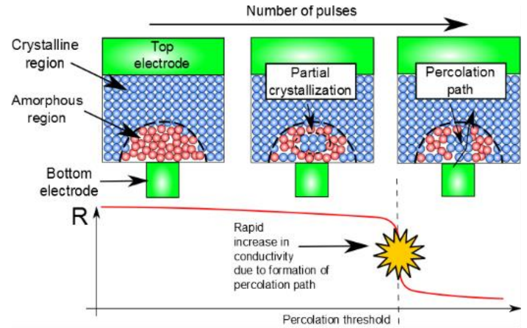
Qiao et al., Front. Neuroscience, 2015

Phase change devices in spiking neural networks





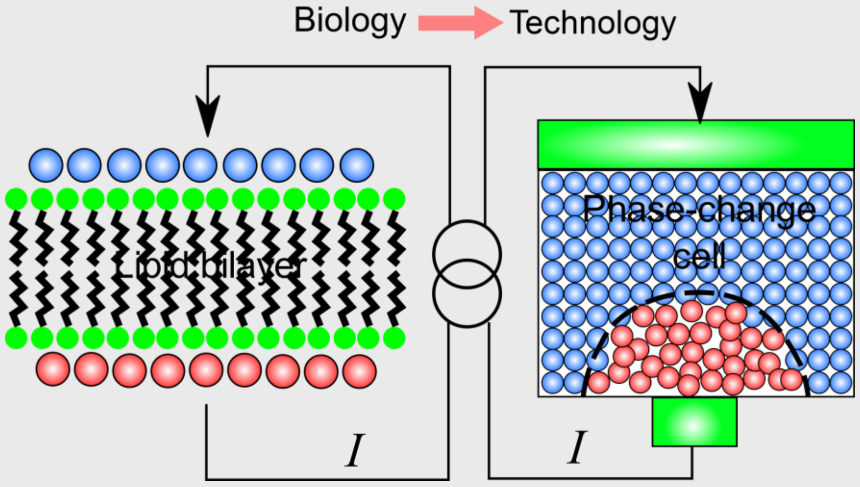
Kuzum et al., NanoLetters, 2012
Jackson et al., ACM JETCS, 2013

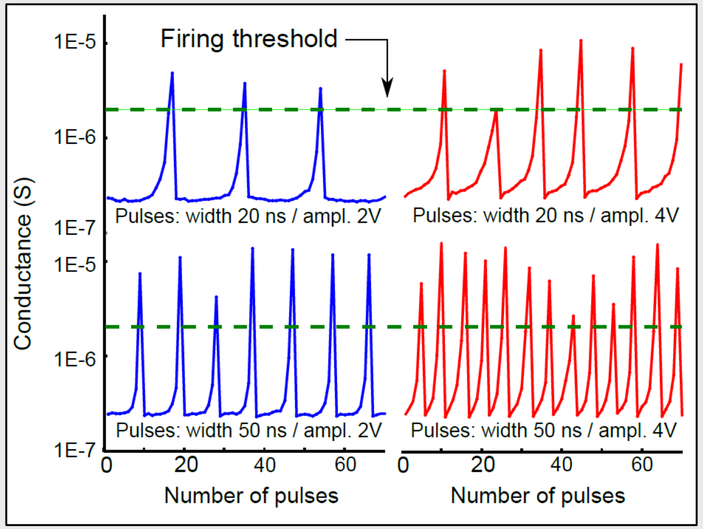


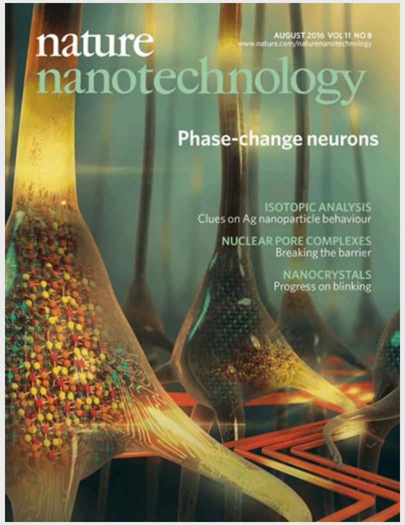
Ovshinsky, E/PCOS, 2004
Wright, Advanced Materials, 2011
Tuma et al., Nature Nanotech., 2016

- Areal/energy efficiency
- **Can we exploit some unique physical attributes?**

Stochastic phase-change neurons





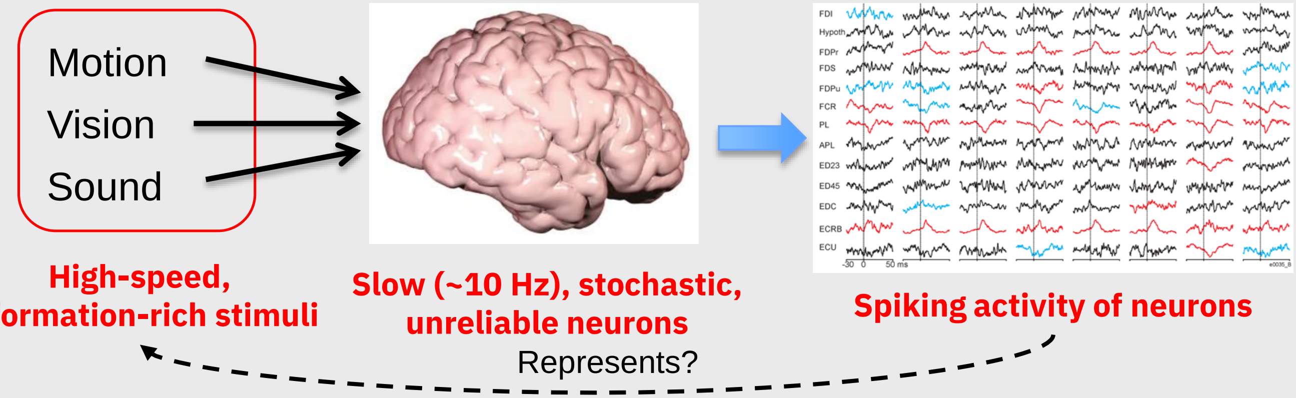


- The **internal state of the neuron** is stored in the phase configuration of a PCM device
- Neuronal dynamics emulated using the **physics of crystallization**
- **Exhibit inherent stochasticity**, which is key for **neuronal population coding**

Tuma et al., Nature Nano., 2016

Neuronal population coding

How does the brain store and represent complex stimuli given the slowness, unreliability and uncertainty of individual neurons?



Motion

Vision

Sound

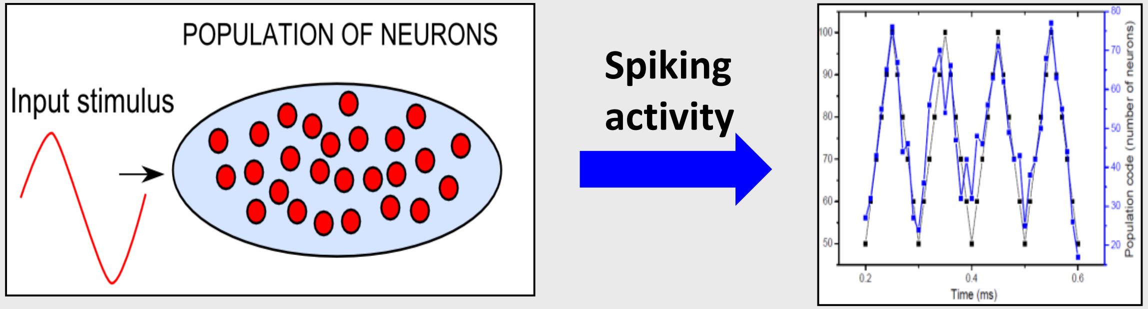
High-speed, information-rich stimuli

Slow (~10 Hz), stochastic, unreliable neurons

Spiking activity of neurons

Represents?

> "As in any good democracy, **individual neurons count for little**; it is **population activity** that matters. For example, as with control of eye and arm movements, visual discrimination is much more accurate than would be predicted from the responses of single neurons." (***Averbeck et al., Nature Reviews, 2006***)



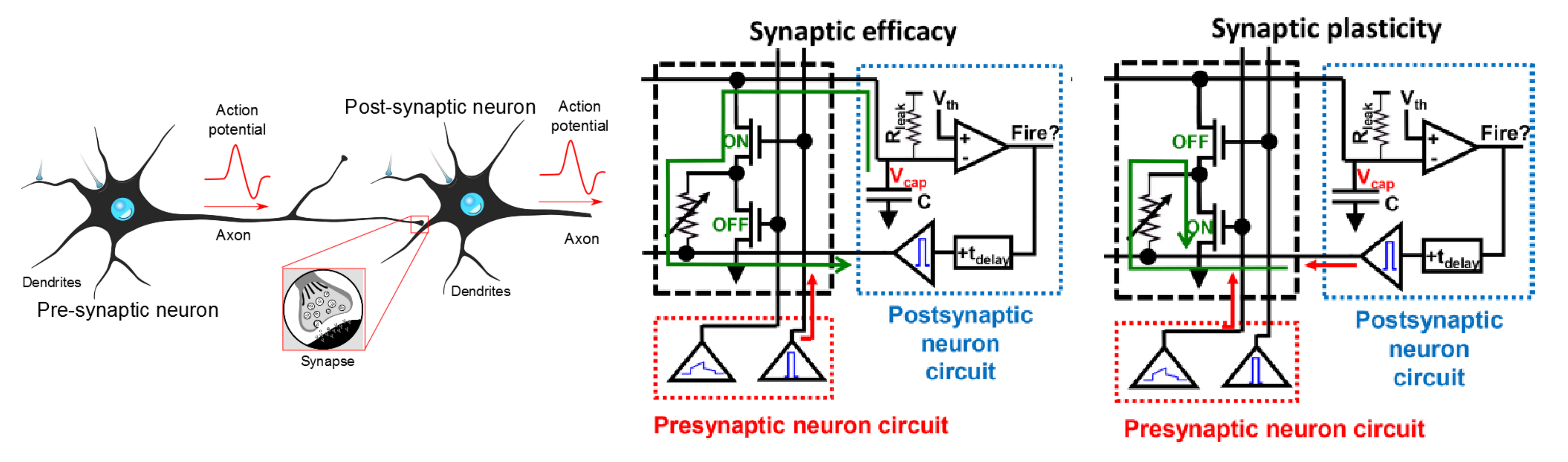
POPULATION OF NEURONS

Input stimulus

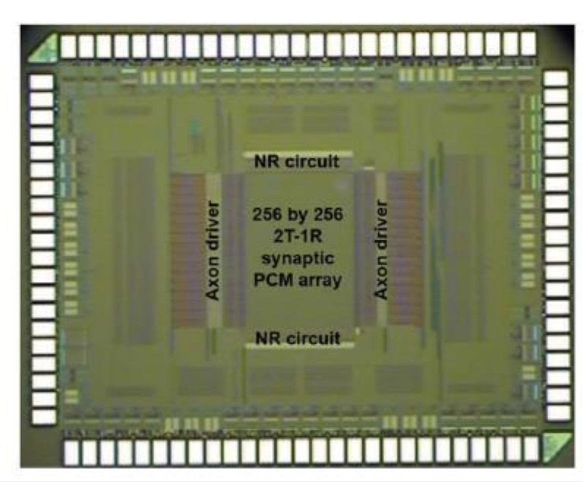
Spiking activity

Tuma et al., Nature Nano., 2016

2T-1R PCM Synapses



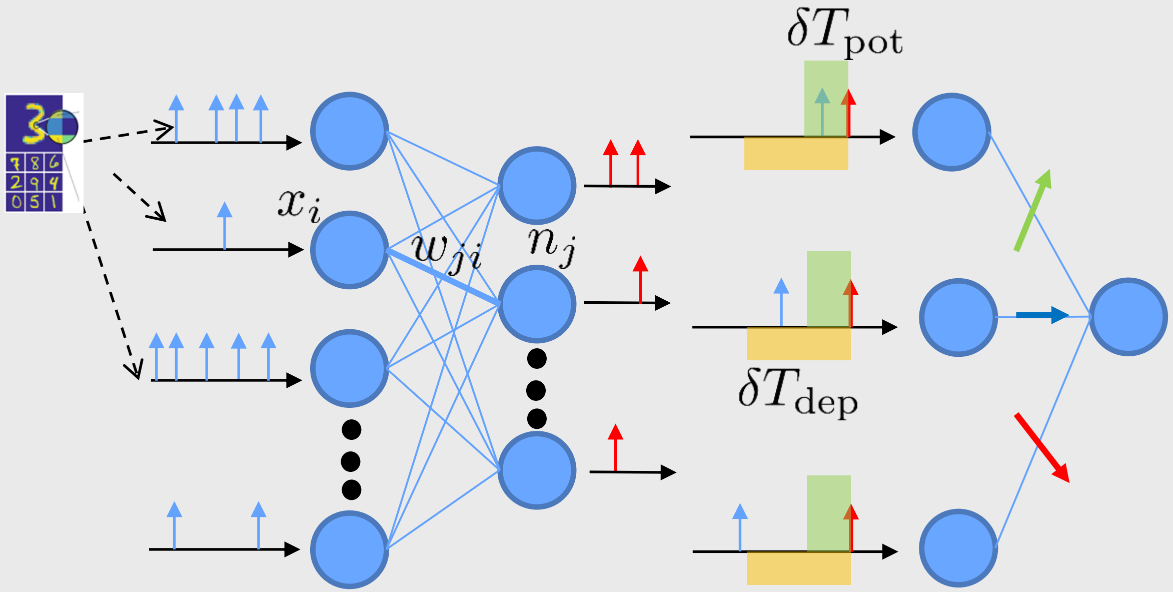
- A 2T-1R PCM unit can implement both synaptic efficacy and plasticity in a very efficient manner
- Neuromorphic core with 64k-cell PCM synaptic array and in-situ learning capability was demonstrated

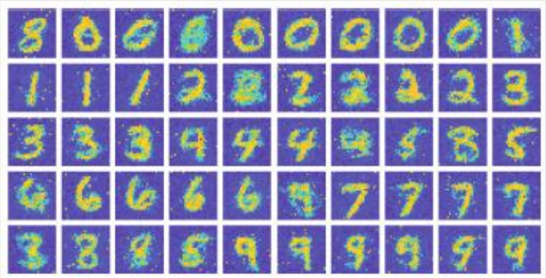


Kim et al., IEDM., 2015

Applications of SNNs

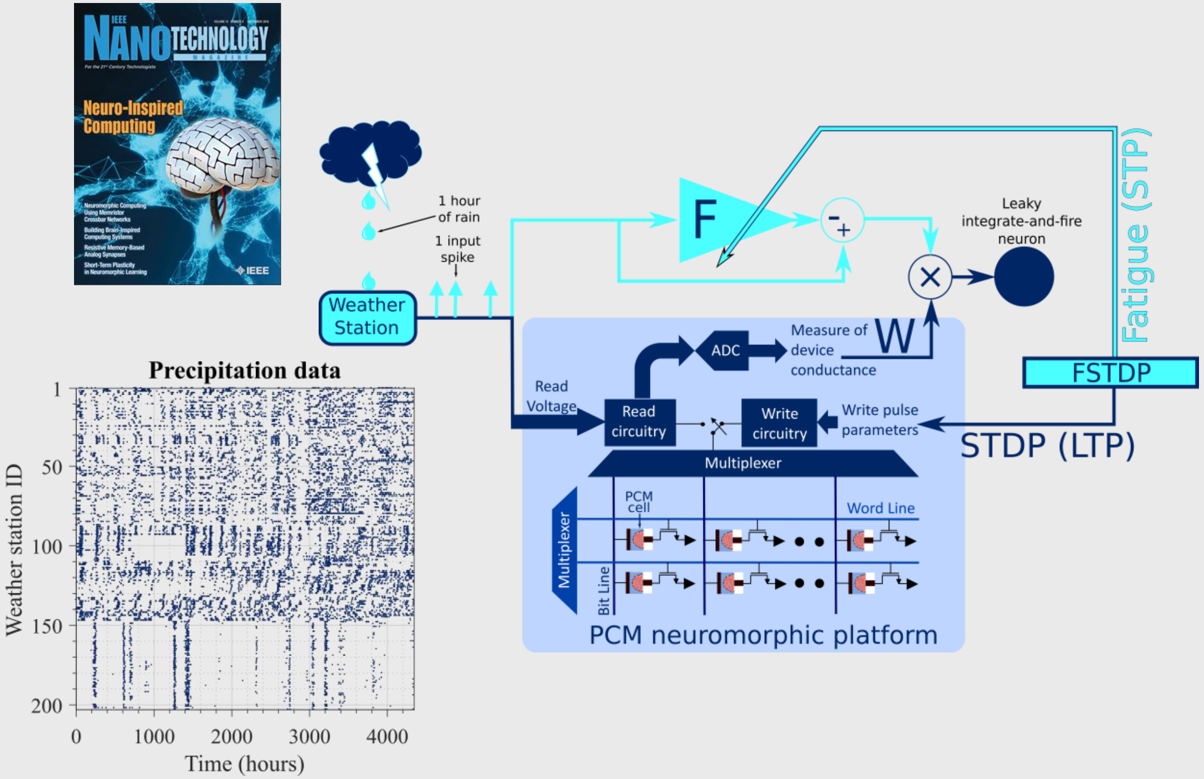
Efficient unsupervised learning via local learning rules





Sidler et al., ICANN, 2017
Wozniak et al., IJCNN, 2017, 2018

Multi-time scale learning using short-term plasticity



Moraitis et al., IJCNN, 2017, 2018
Moraitis et al., IEEE Nanotech. Magazine, 2018

Summary

- **The AI revolution** is a significant driver for brain-inspired computing
- Brain-inspired computing can be realized at **multiple levels of inspiration** and resistive memory devices such as PCM could play a key role
- **First level of inspiration: In-memory computing**
  - ✓ Matrix-vector multiplication is a computational primitive that can be applied to a range of applications such as compressed sensing and solving systems of linear equations
  - ✓ Detecting temporal correlations is a fascinating application of computing with device dynamics
- **Second level of inspiration: Co-processors for deep learning**
  - ✓ Mixed-precision in-memory co-processors for inference and training
- **Third level of inspiration: Computational substrates for Spiking Neural Networks**
  - ✓ Emulation of neuronal and synaptic dynamics
  - ✓ Unsupervised multi-time-scale learning is a very promising application domain

Outlook

BRAIN-INSPIRED COMPUTING

STORAGE
(e.g. Flash, HDD)
(nonvolatile, slow)

STORAGE-CLASS MEMORY

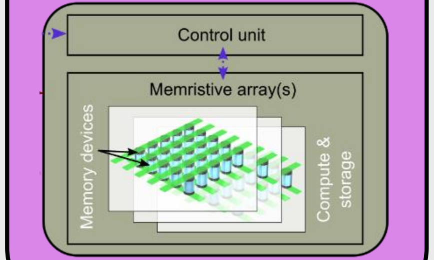
MEMORY (e.g. DRAM)
(volatile, fast)

CMOS processing units

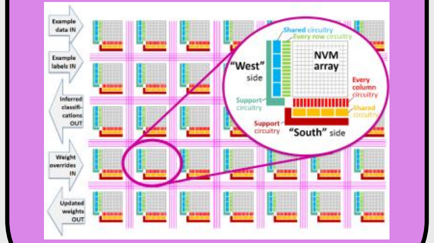
Von Neumann ACCELERATORS (e.g. GPUs, ASICs)

High-speed memory

In-memory computing for ML/DL inference

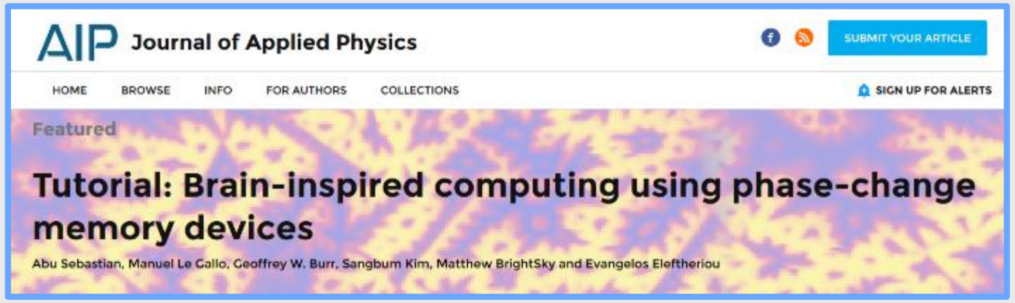


Co-processors for DL training



Neuromorphic co-processors (SNNs?)

CENTRAL PROCESSING UNIT (CPU)



Journal of Applied Physics

Featured

Tutorial: Brain-inspired computing using phase-change memory devices

Abu Sebastian, Manuel Le Gallo, Geoffrey W. Burr, Sangbum Kim, Matthew BrightSky and Evangelos Eleftheriou