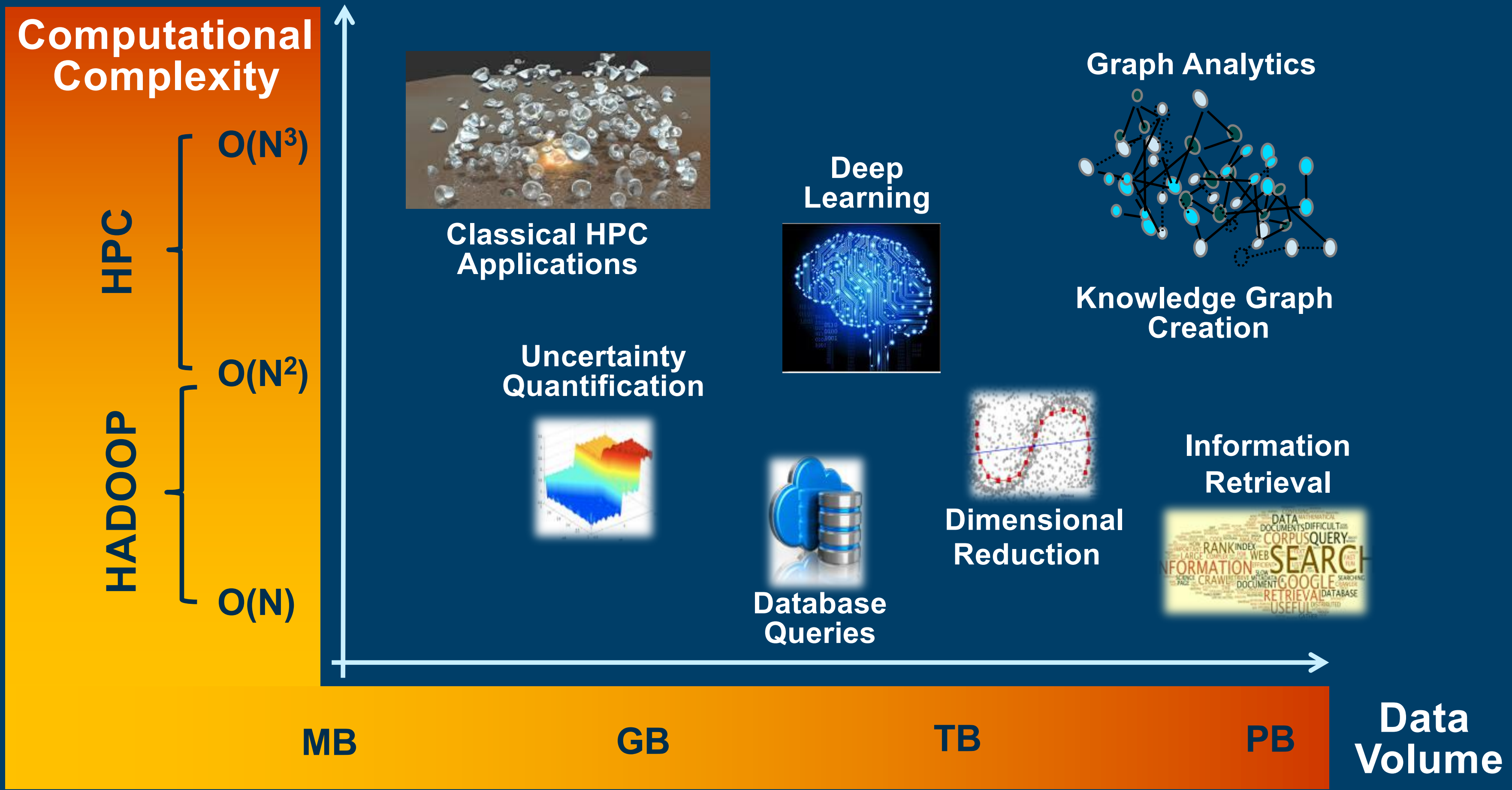
September 11, 2018

Brain-inspired non-von Neumann Computing for AI applications

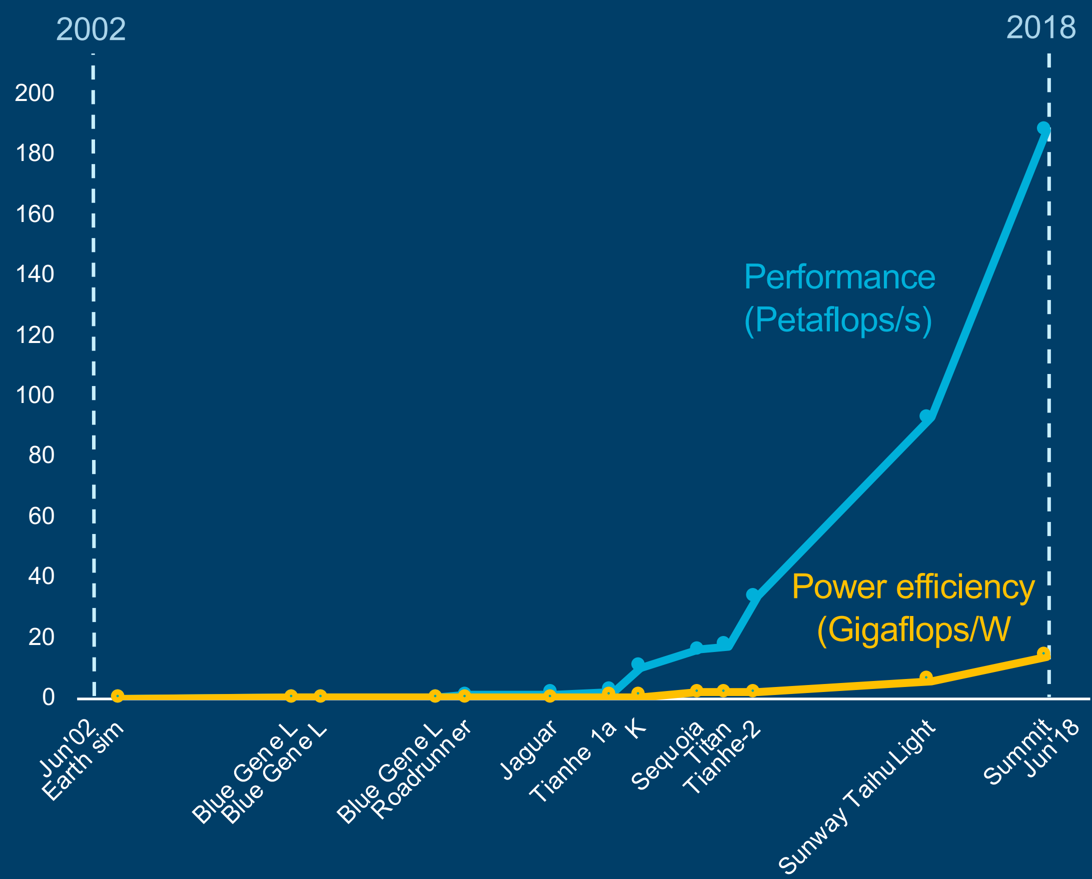
Evangelos Eleftheriou, IBM Fellow

IBM Research - Zurich

Application Trends



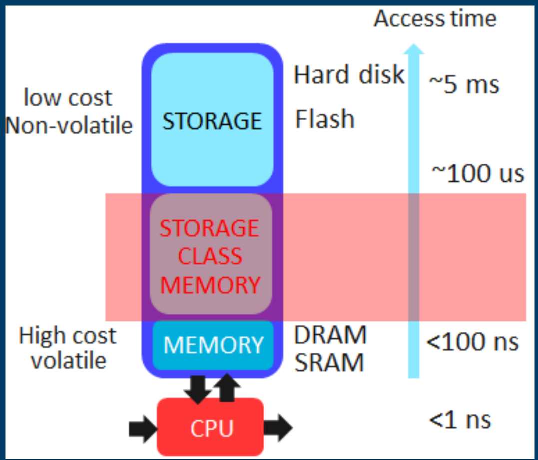
Performance and Power Efficiency Trends



- Increasing gap between performance and power efficiency
- Diminishing performance/power efficiency gains from technology scaling

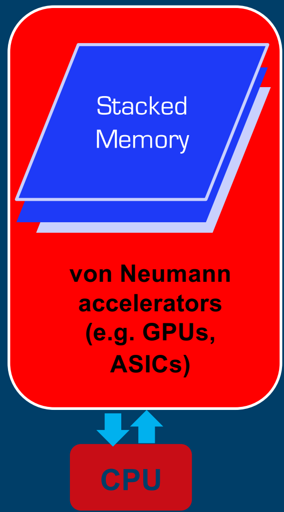
Advances in von Neumann Computing

Storage class memory

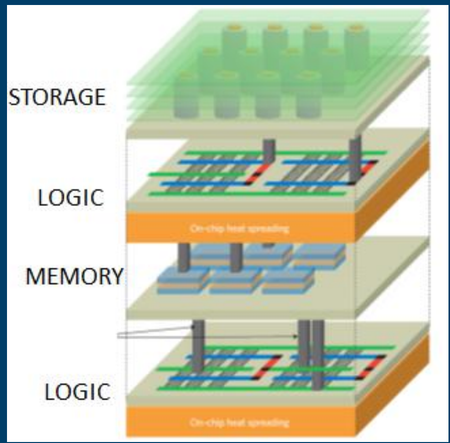


Burr *et al.*, *IBM J. Res. Dev.*, 2008

Near memory computing



Monolithic 3D integration



Wong, Salahuddin, *Nature Nanotechnology*, 2015

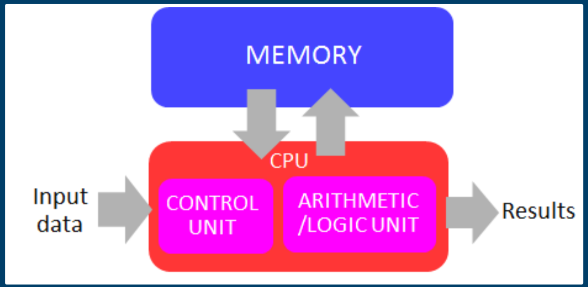
Minimize the time and distance to memory access

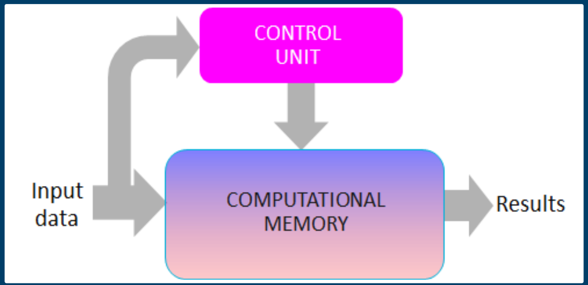
Go beyond von Neumann Computing

Spiking Neural Networks

LeCun, Bengio, Hinton, *Nature*, 2015
Merolla *et al.*, *Science*, 2014
Indiveri, Liu, *Proc. IEEE*, 2015

Computational memory



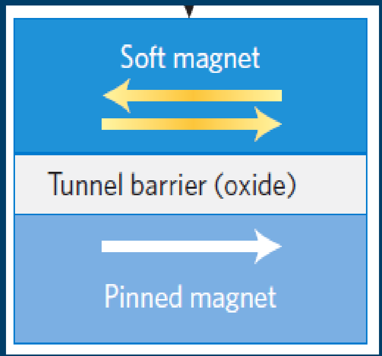


Borghetti et al., *Nature*, 2010
Di Ventra and Pershin, *Scientific American*, 2015
Hosseini et al., *Electron Dev. Lett.*, 2015

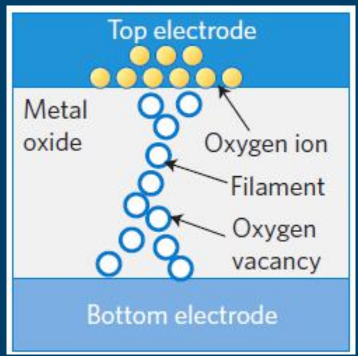
Enabling bio-mimetic *Computation* and *Storage*

Neuromorphic and In-memory Computing:
the Constituent Elements

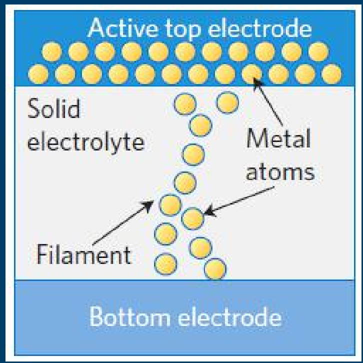
Charge-based memory/storage → resistance-based memory/storage



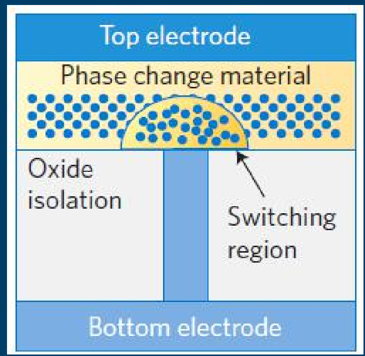
Spin-torque transfer magnetic random access memory (STT-MRAM)



Metal oxide random access memory (ReRAM)



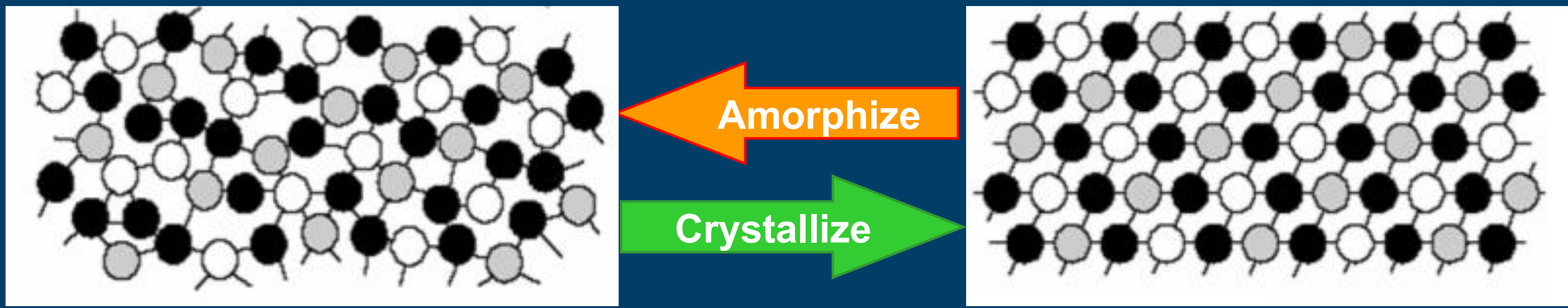
Conductive bridge random access memory (CBRAM)



Phase change memory (PCM)

- Significant impact on memory/storage hierarchy
- Monolithic integration of memories and computation units
- Sufficient richness of dynamics for non-von Neumann computing

Phase-Change Memory (PCM)

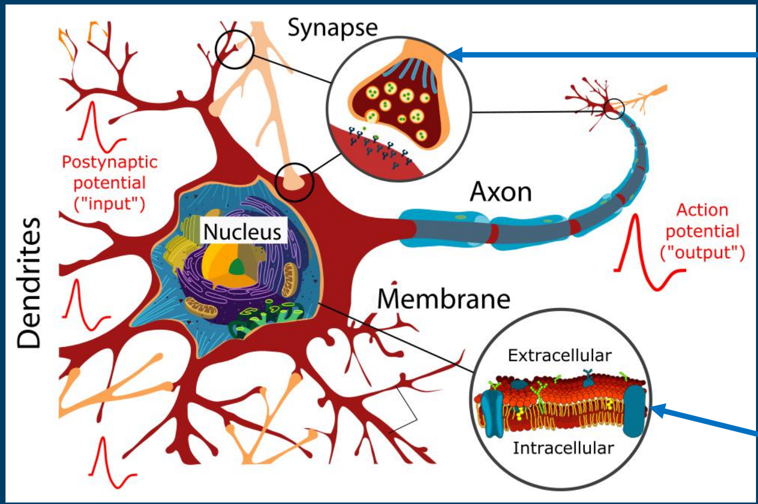


High-resistance state

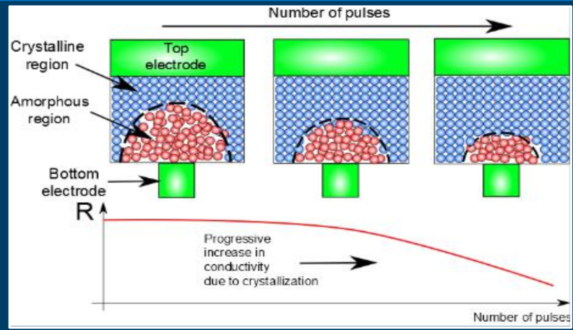
Low-resistance state

- Use two distinct solid phases of a Ge-Sb-Te metal alloy to store a bit
- Use intermediate phases to obtain a continuum of different states or resistance levels
- Transition between phases by controlled heating and cooling

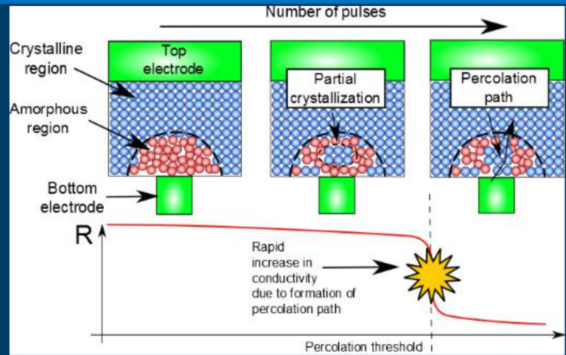
Phase-Change Devices in Spiking Neural Networks



Synapse



Neuron



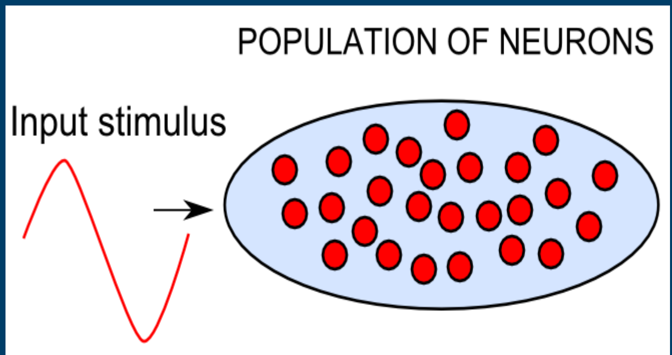
Ovshinsky, *E\PCOS*, 2004
Wright, *Adv. Mater.*, 2011
Kuzum *et al.*, *Nano Lett.*, 2012
Jackson *et al.*, *ACM JETCS*, 2013

Tuma *et al.*, *Nature Nanotechnology*, 2016
Pantazi *et al.*, *Nanotechnology*, 2016
Tuma, *et al.*, *IEEE Electron Dev. Lett.*, 2016

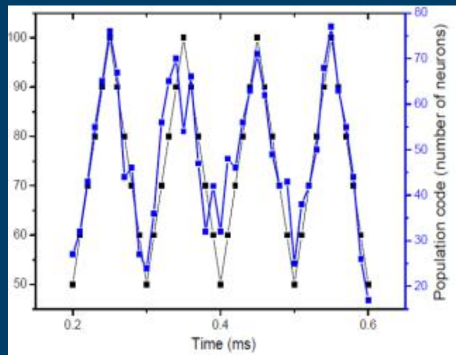
All-PCM non-von Neumann architecture: Areal/energy efficiency

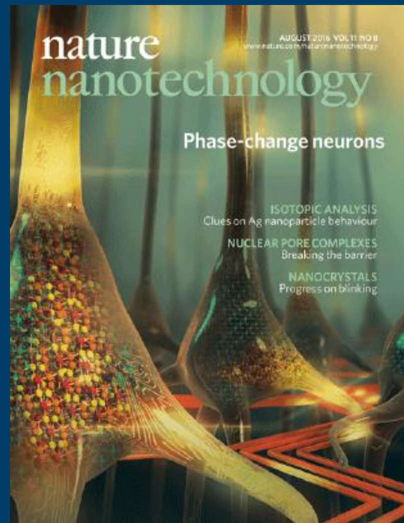
Neuronal Population Coding

High-speed, information-rich stimuli are processed by populations of slow (~10 Hz), stochastic, and unreliable neurons in our brain



Spiking Activity

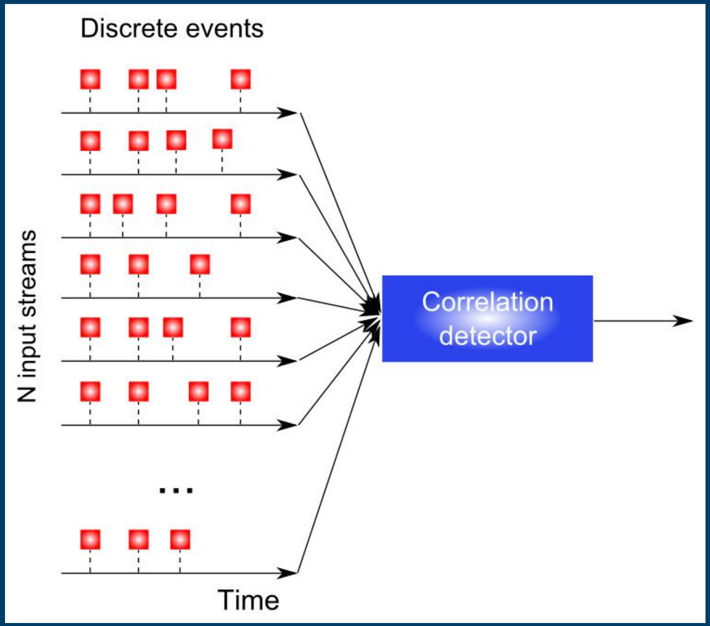




T. Tuma, et. al. *Nature Nanotechnology*, Aug. 2016

- 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**

Application: Temporal Correlation Detection



Algorithmic goals

- Determine whether some data streams are **statistically correlated**
- **Observe variations** in the activity of the correlated input
- Quickly **react to occurrence of correlated inputs**
- Continuously and **dynamically re-evaluate** the learned statistics

FINANCE

SCIENCE

MEDICINE

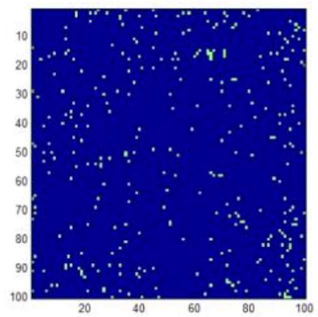
BIG DATA

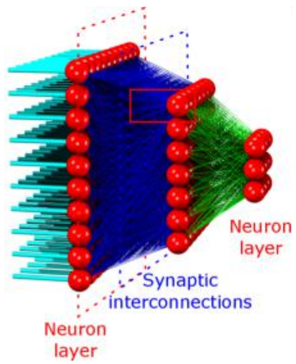
… and more

Learning Patterns with a Spiking Neural Network

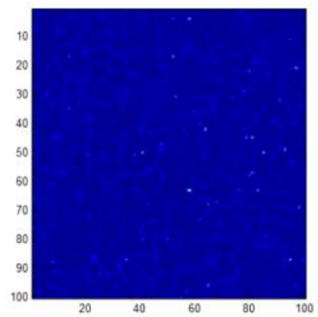
Experiments with 30,000 PCM cells

Input pattern

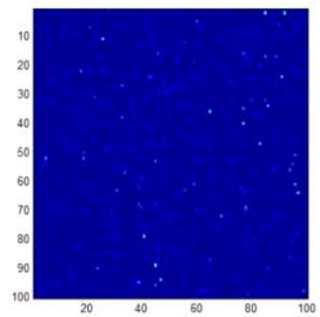




Neuron #1: synaptic weights



Neuron #2: synaptic weights



A. Pantazi *et al.*, *Nanotechnology*, 2016

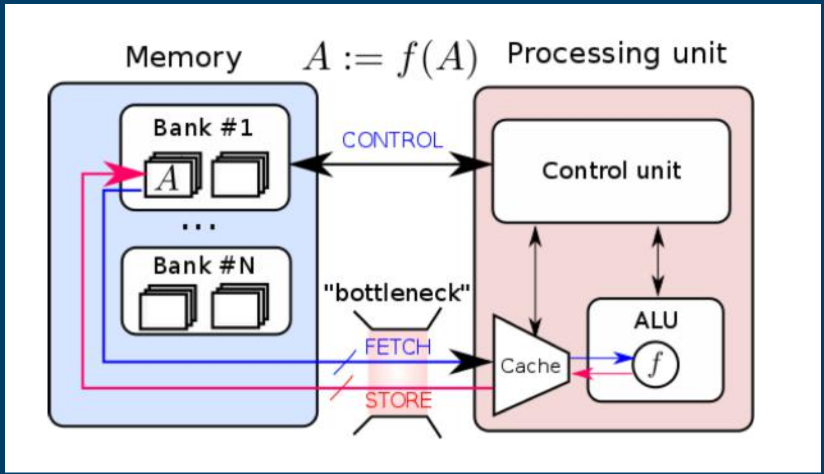
Neuron #1: output

Neuron #2: output

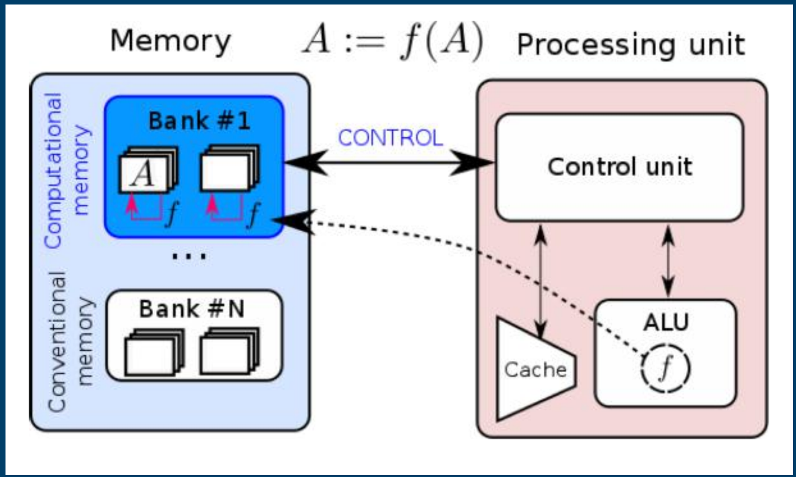
Purely unsupervised neuromorphic computation:
No counting, no transfers between memory and CPU!

In-memory Computing

Processing unit & Conventional memory



Processing unit & Computational memory



Borghetti et al, *Nature*, 2010
Di Ventra and Pershin, *Scientific American*, 2015
Hosseini et al., *Elect. Dev. Lett.*, 2015
Sebastian et al., *Nature Communications* 2017

- 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

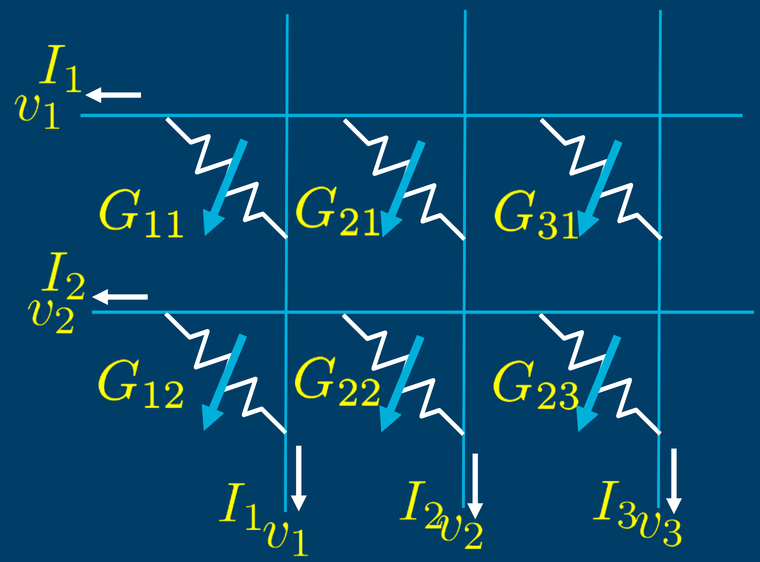
Matrix-vector Multiplication

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

MAP to conductance values

MAP to read voltage

DECIPHER from the current

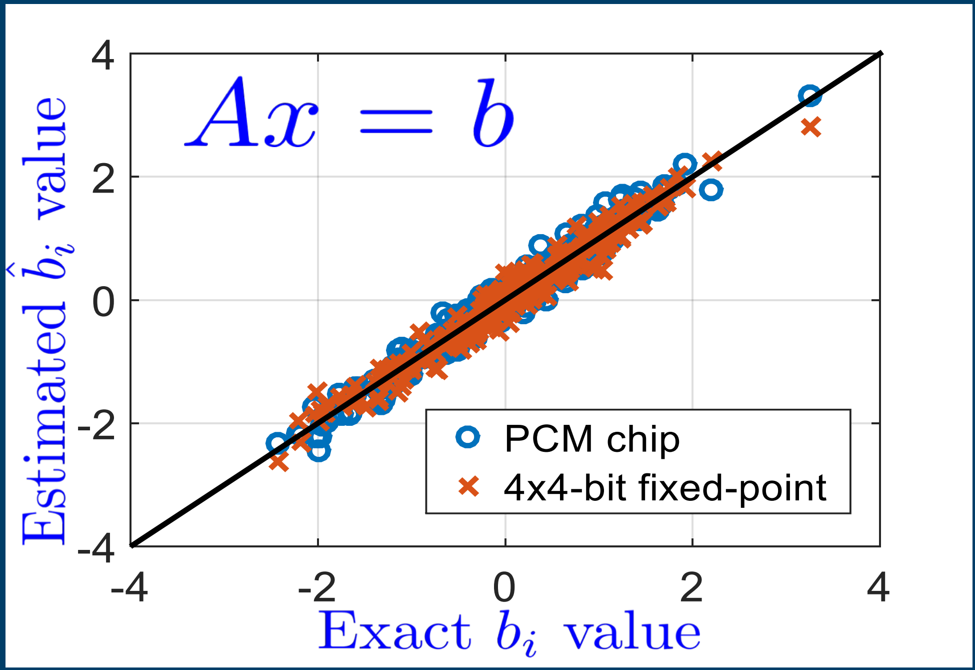


Burr et al., *Adv. Phys: X*, 2017
Zidan et al., *Nature Electronics*, 2018

- Matrix multiplication: Exploits multi-level storage capability and Kirchhoff and Ohm laws
- A crossbar array performs fast matrix-vector multiplication without data movements in O(1)

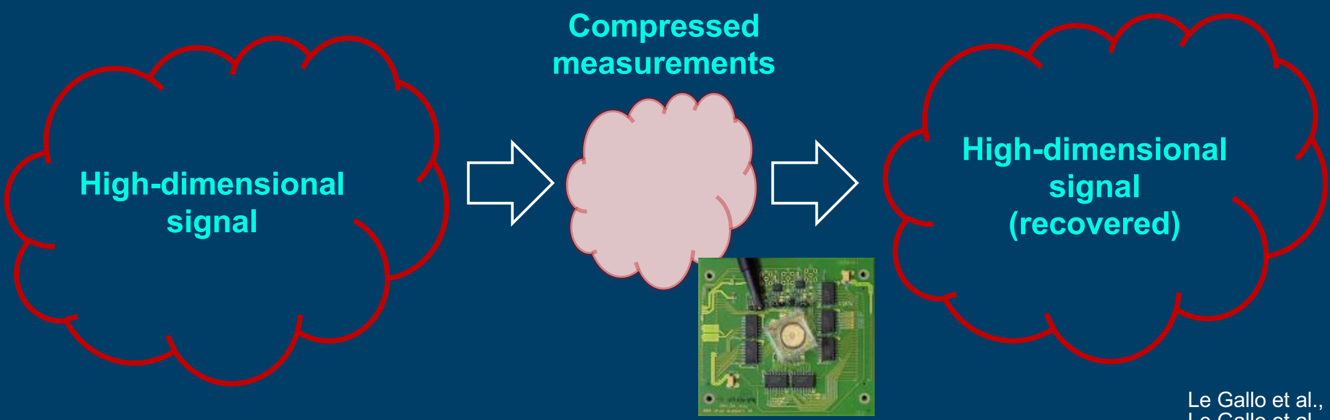
Matrix-vector Multiplication using PCM Devices

IBM



- A is a 256 X 256 Gaussian matrix coded in a PCM chip
- x is a 256-long Gaussian vector applied as voltages

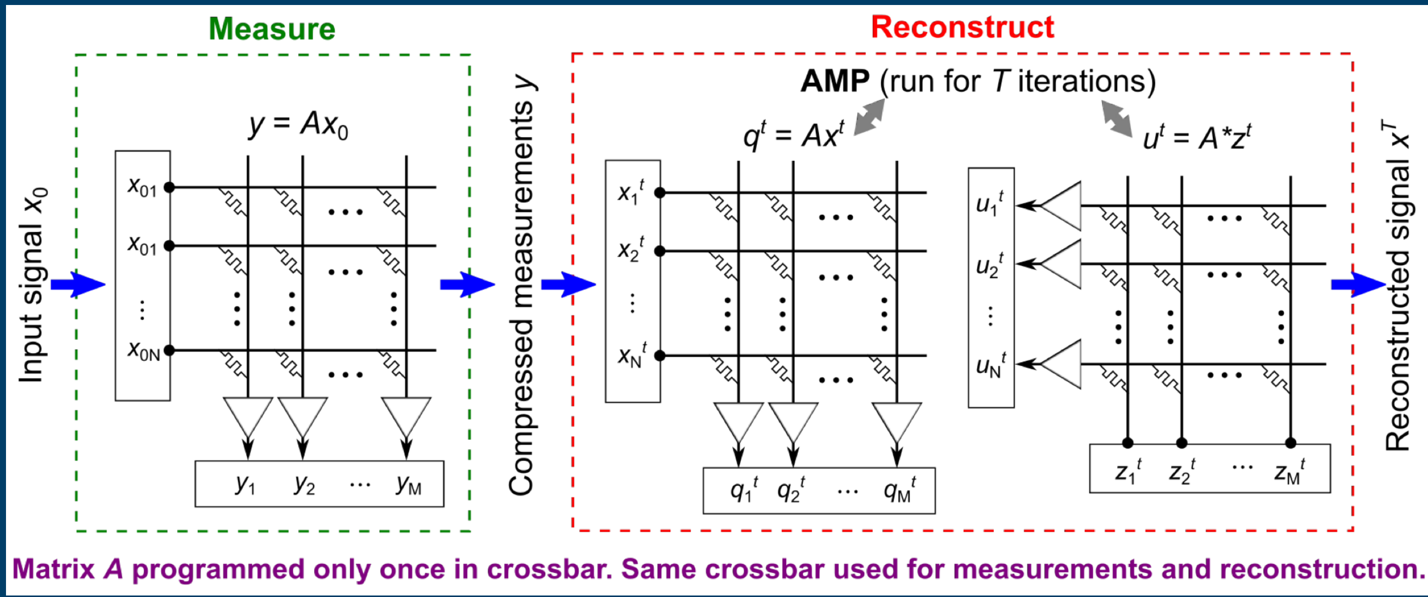
Applications: Optimization Solvers



Le Gallo et al., *IEDM*, 2017
Le Gallo et al., *IEEE TED*, 2018

- Compressed sensing: Acquire a large signal at sub-Nyquist sampling rates and subsequently reconstruct that signal accurately
- Applications in MRI, facial recognition, holography, audio restoration or in mobile phone camera sensors

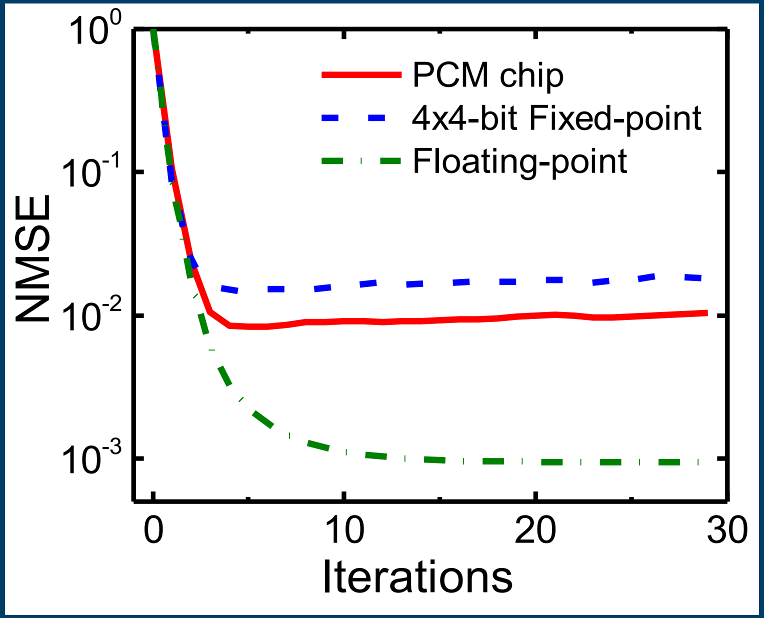
Compressed Sensing and Recovery



Le Gallo et al., *IEDM*, 2017
Le Gallo et al., *IEEE TED*, 2018

Complexity reduction: O(NM) → O(N);
Potential 10^5 speed-up on 1000 x 1000 pixel image with 10-fold compression ratio

Compressive Imaging: Experimental Results



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

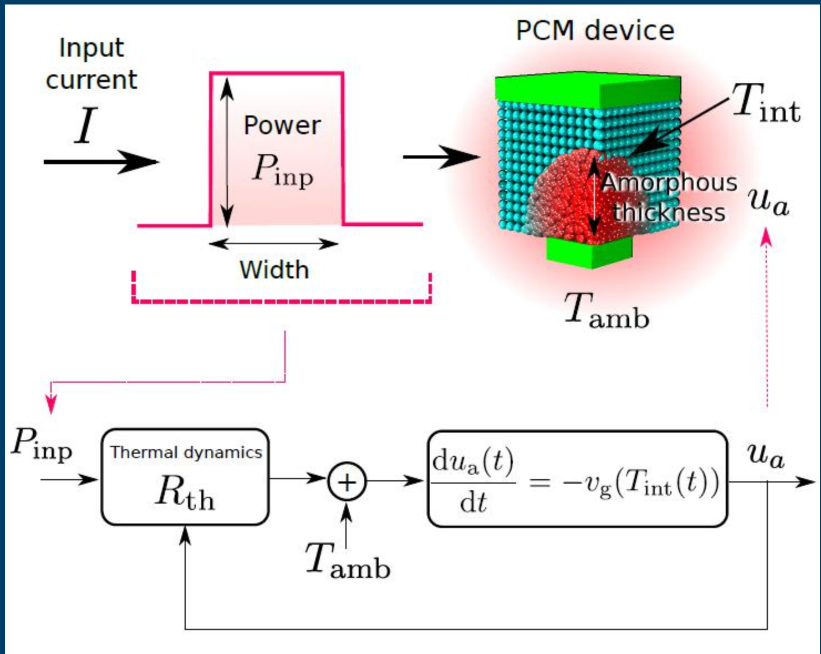
Original image

Reconstructed image

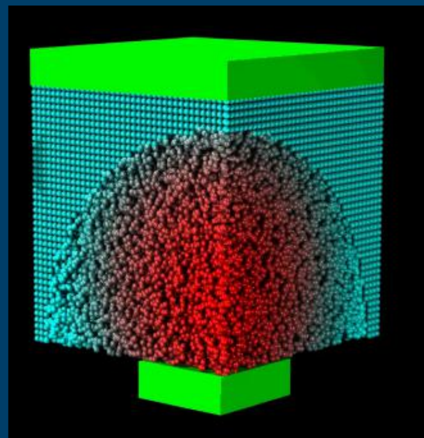
Le Gallo et al., *IEDM*, 2017
Le Gallo et al., *IEEE TED*, 2018

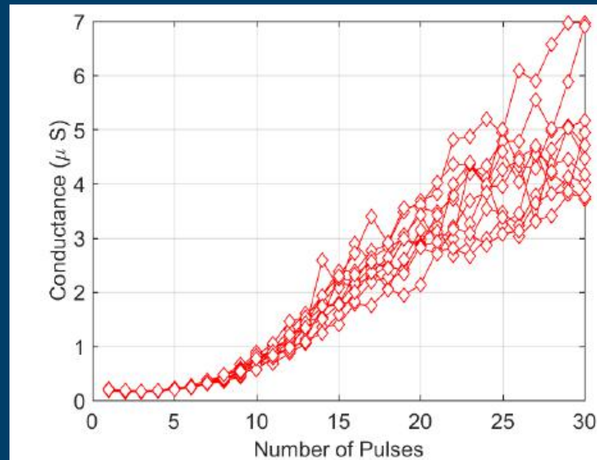
Estimated power reduction of 50x compared to using an optimized 4-bit FPGA matrix-vector multiplier that delivers same reconstruction accuracy at same speed

Can We Compute with the Dynamics of PCM?



A nanoscale non-volatile integrator

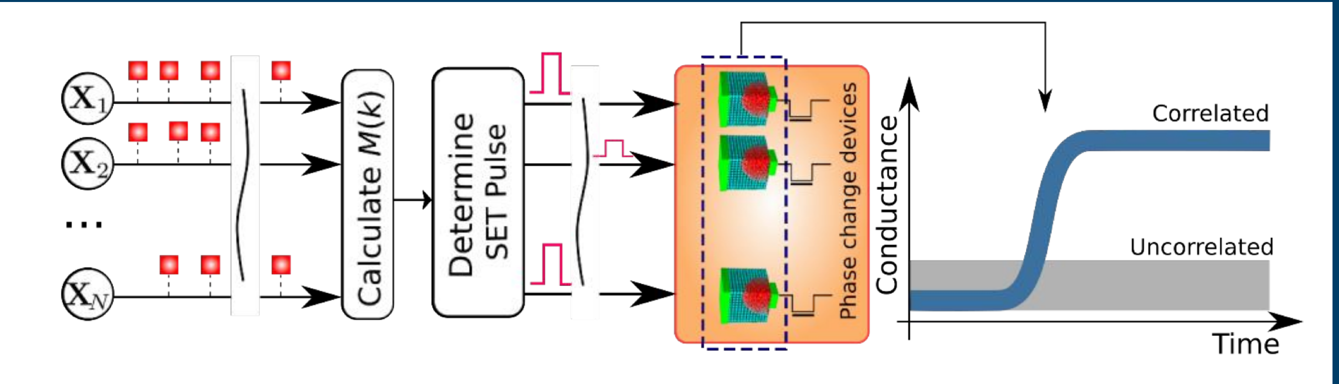




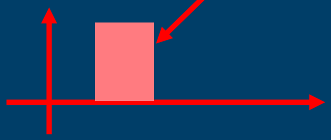
Sebastian et al., *Nature Communications*, 2014

Nonvolatile nanoscale integrator but stochastic and nonlinear

Unsupervised Learning of Temporal Correlations



Modulate the amplitude or width of the applied pulse based on: $M(k) = \sum_{j=1}^{N} X_j(k)$

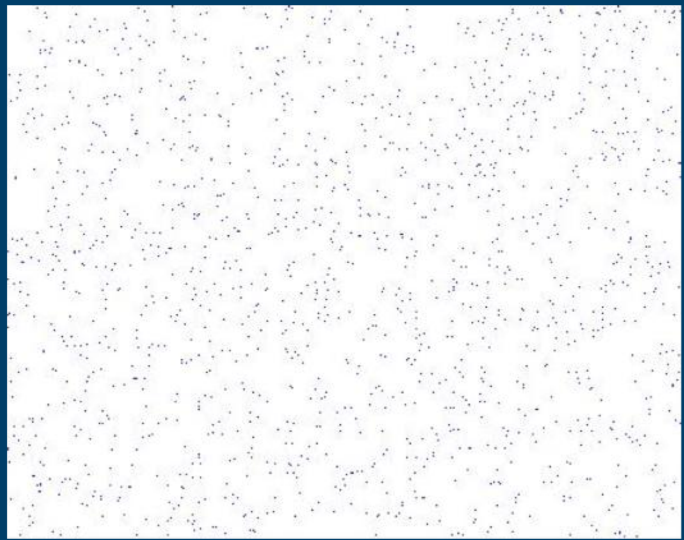


Sebastian et al., *Nature* Communication, 2017

Devices interfaced to the correlated processes go to a high conductance state

Experimental Results (1 Million PCM Devices)

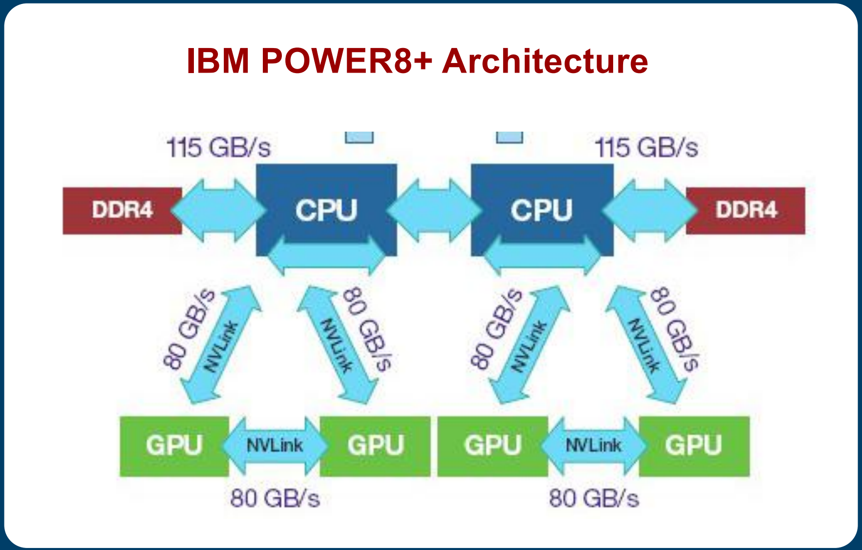
Processes

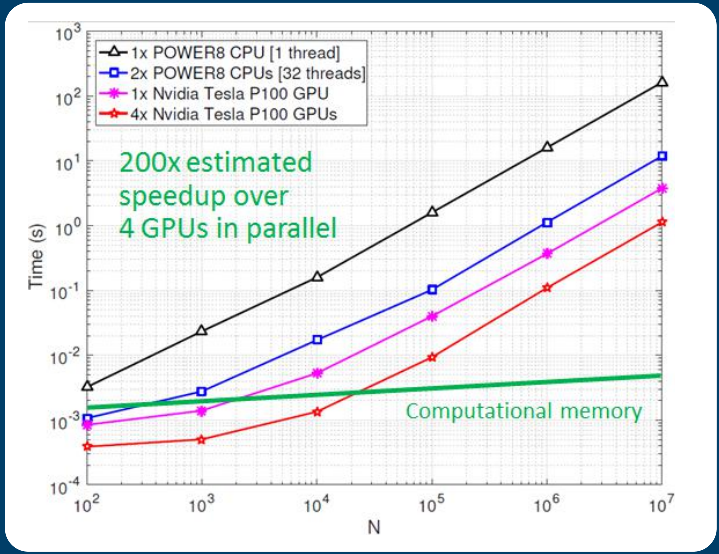


Conductance

- **Very weak correlation of c = 0.01**
- **No shuttling back and forth of data**
- **Massively parallel**

Comparative Study



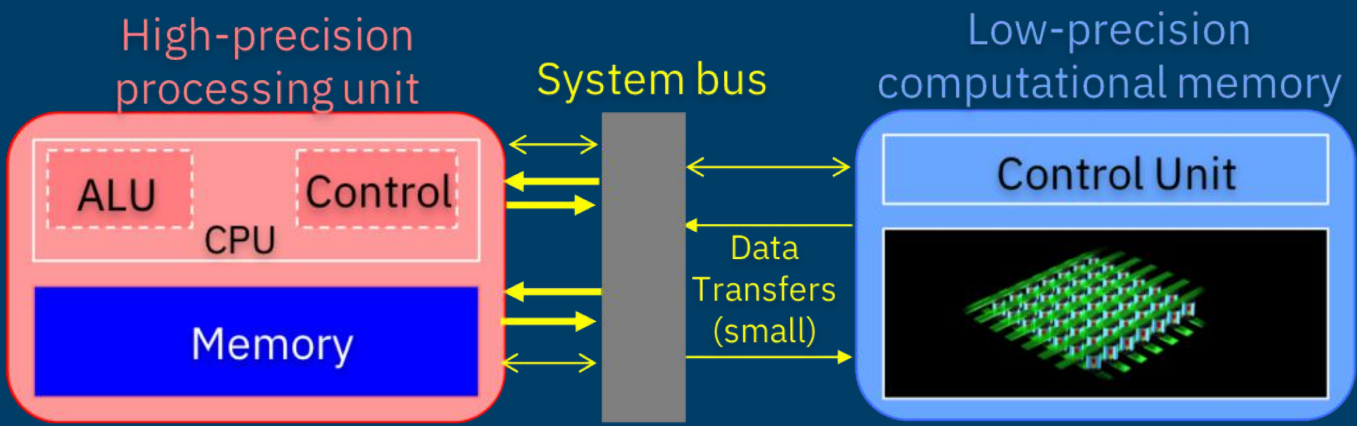


Sebastian et al., *Nature Communications*, 2017

Complexity reduction: O(N) → O(klog(N)).
For 10^7 parallel processes a 200X improvement in computation time is expected !
2 orders of magnitude energy improvement

What if Arbitrarily High-precision is Needed?

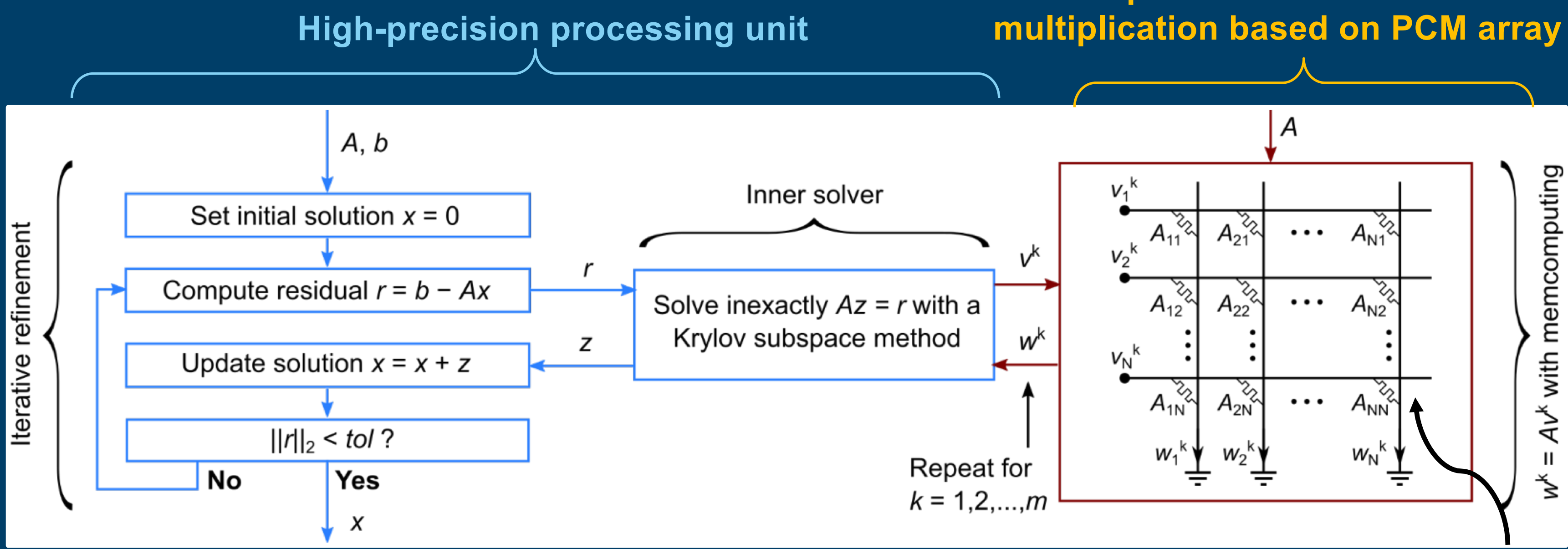
Mixed-precision computing to the rescue!



- **Bulk of computations in low-precision Computational Memory**
- **Refinement in high-precision digital processing engine**

Application: Linear Equation Solver

$$\text{if } Ax = b, \text{ find } x$$

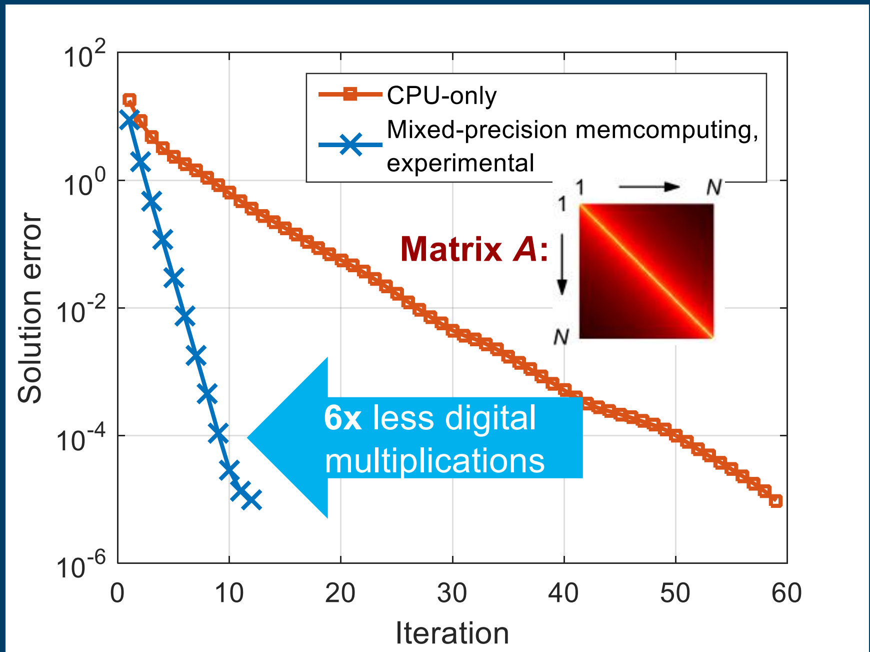


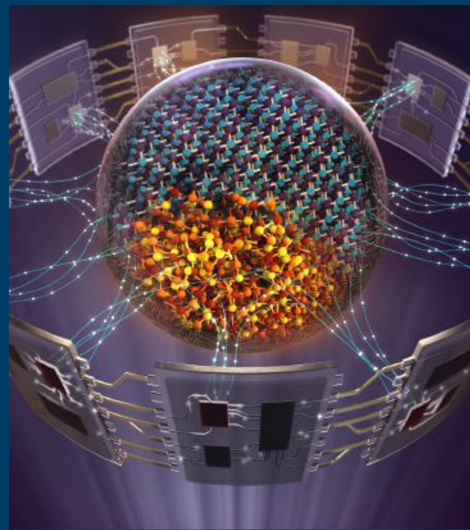
Le Gallo et al., *Nature Electronics*, 2018

- Solution iteratively updated with low-precision error-correction term
- Error-correction term obtained using inexact inner solver
- The matrix multiplications in the inner solver are performed using a PCM array

Linear Equation Solver: Experimental Results

**Experimental result: 10,000x10,000 matrix,
959,376 PCM devices**

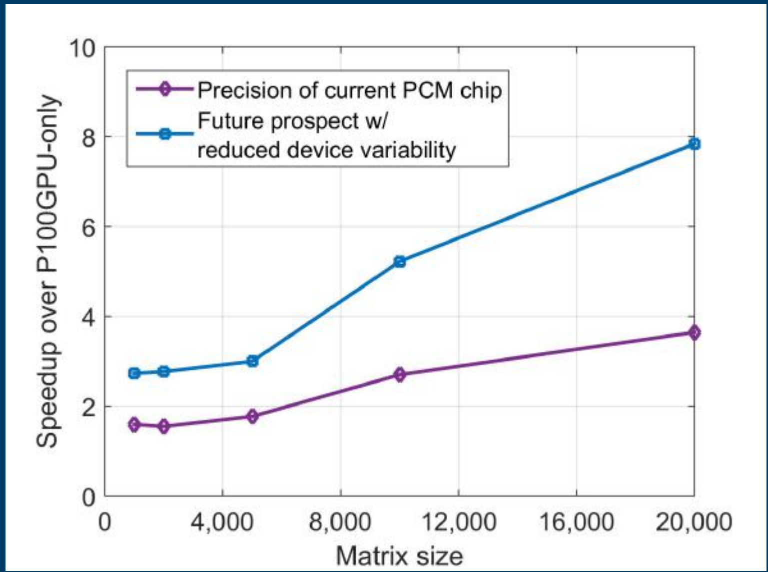


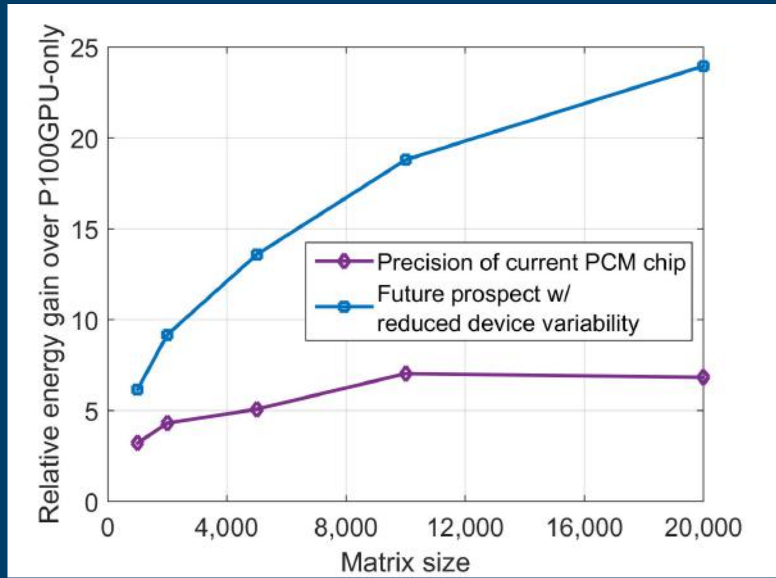


Le Gallo et al., *Nature Electronics*, 2018

Mixed-precision computing provides a pathway for arbitrarily precise computation using computational memory.

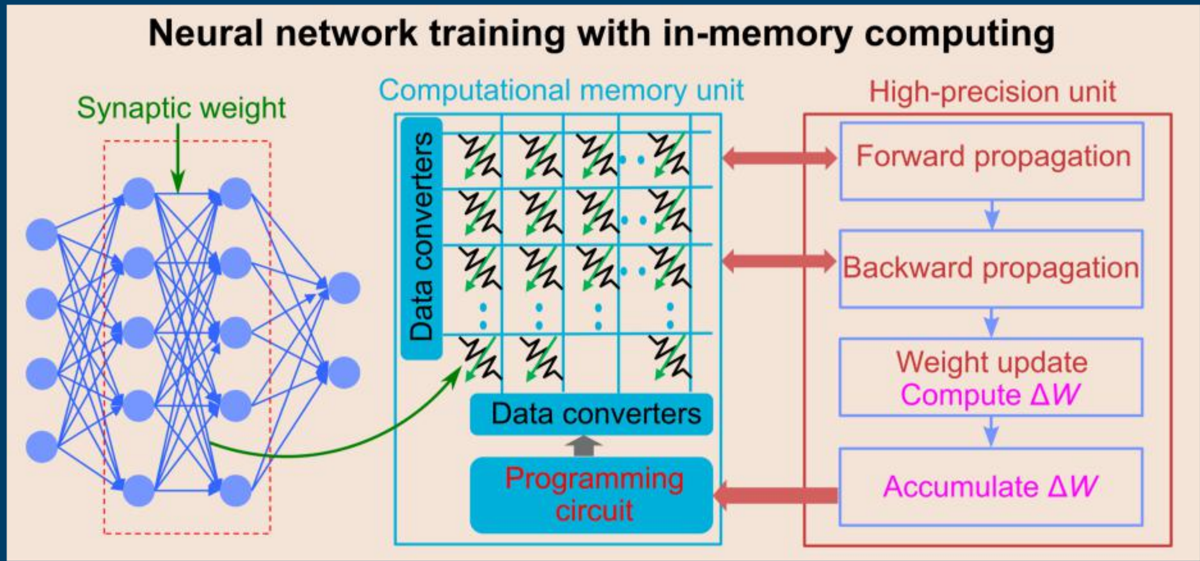
System-Level Performance Analysis





- Significant improvement in the time/energy to solution metrics
- The higher the accuracy of the computational memory, the higher the gain

Application: Mixed-Precision Deep Learning

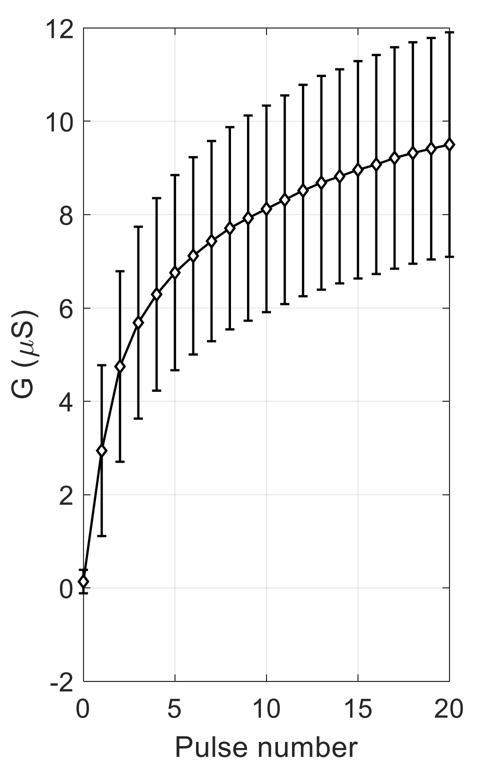


Nandakumar et al., *arXiv*:1712.01192, 2017
Nandakumar et al., *ISCAS*, 2018

- **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**

Mixed-precision DL: Simulations

PCM model



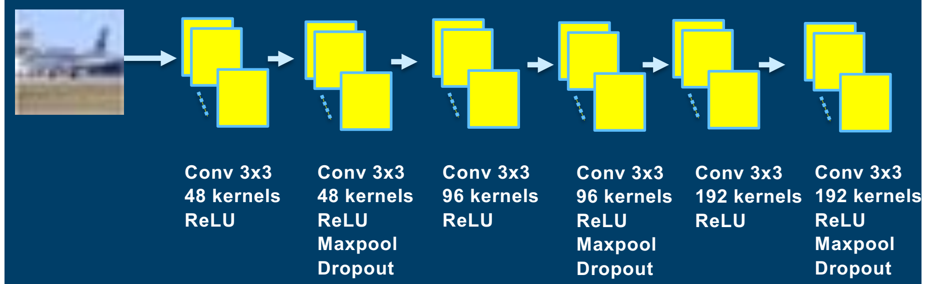
- CIFAR-10 classification problem
- Network: 6 convolution layers, 3 fully connected layer
- Training 400 epochs@ batch size = 100
- Model used: Most realistic device model known

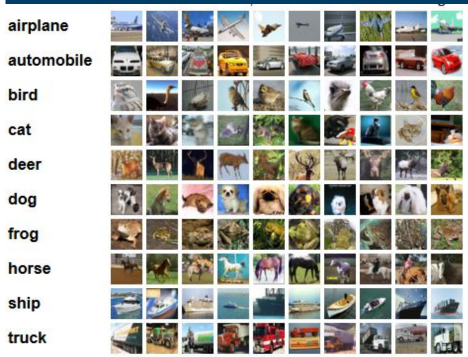
Test accuracies:

float32 Precision
86.02%

Mixed-Precision
86.89%

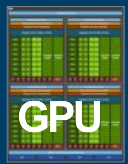
⇒ ***Better than floating point precision accuracy due to regularization effect!***

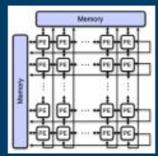


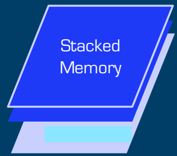


In-memory Computing: Future perspectives

Orders of magnitude improvements in speed and efficiency are possible





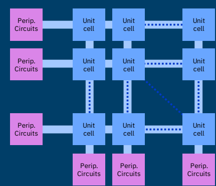


Algorithms and Architectures for approximate computing

Traditional CMOS

1x — 10x — 50-100x — *Speed up*

Mixed-precision in-memory computing

RPU

Arrays of analog memory elements, mixed precision

In-memory Computing

10x — 100x — 1000x — 10,000x — *Speed up*

Brain-inspired Computing

Spiking Neural Networks (SNNs) ?

Acknowledgements

- Exploratory memory and cognitive technologies, IBM Zurich
  - Irem Boybat
  - Benedikt Kersting
  - Iason Giannopoulos
  - Riduan Khaddam-Aljameh
  - Manuel Le Gallo
  - Christophe Piveteau
  - Vinay Manikrao
  - Timoleon Moraitis
  - Stanislaw Wozniak
  - Varaprasad Jonnalagadda
  - Angeliki Pantazi
  - Giovanni Cherubini
  - Abu Sebastian
- Foundations of cognitive solutions, Cloud storage and analytics
- M. BrightSky, IBM T.J. Watson Research Center
- G. W. Burr, IBM Research - Almaden
- University of Patras, RWTH Aachen, NJIT, Oxford, Exeter, EPFL, ETH