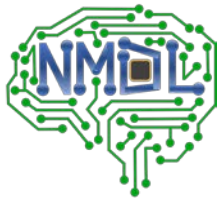
Neuromorphic Materials and Devices Lab.



IBM Research

Neuromorphic computing using phase-change memory devices

[Introduction to SNU class]

2021. 04. 27.

Uicheol Shin

Neuromorphic Device, Science & Technology, IBM Research - Tokyo

Department of Materials Science and Engineering, Seoul National University

Outlines

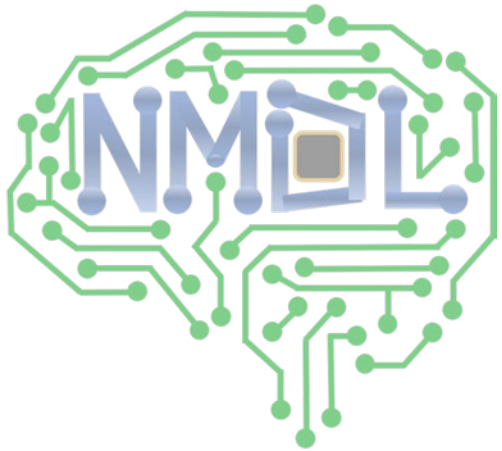
1. Introduction

2. Computational memory

3. Deep learning co-processors

4. Spiking neural networks (SNN)

5. Summary



Outlines

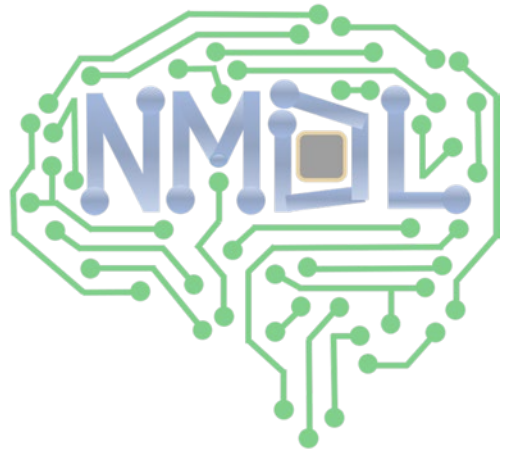
1. Introduction

2. Computational memory

3. Deep learning co-processors

4. Spiking neural networks (SNN)

5. Summary



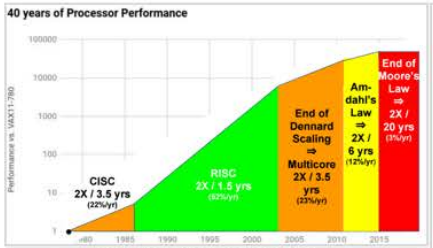
Introduction

nature electronics

Editorial | Published: 17 April 2018

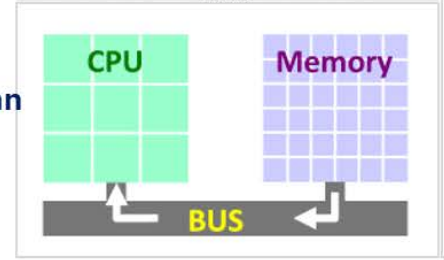
Does AI have a hardware problem?

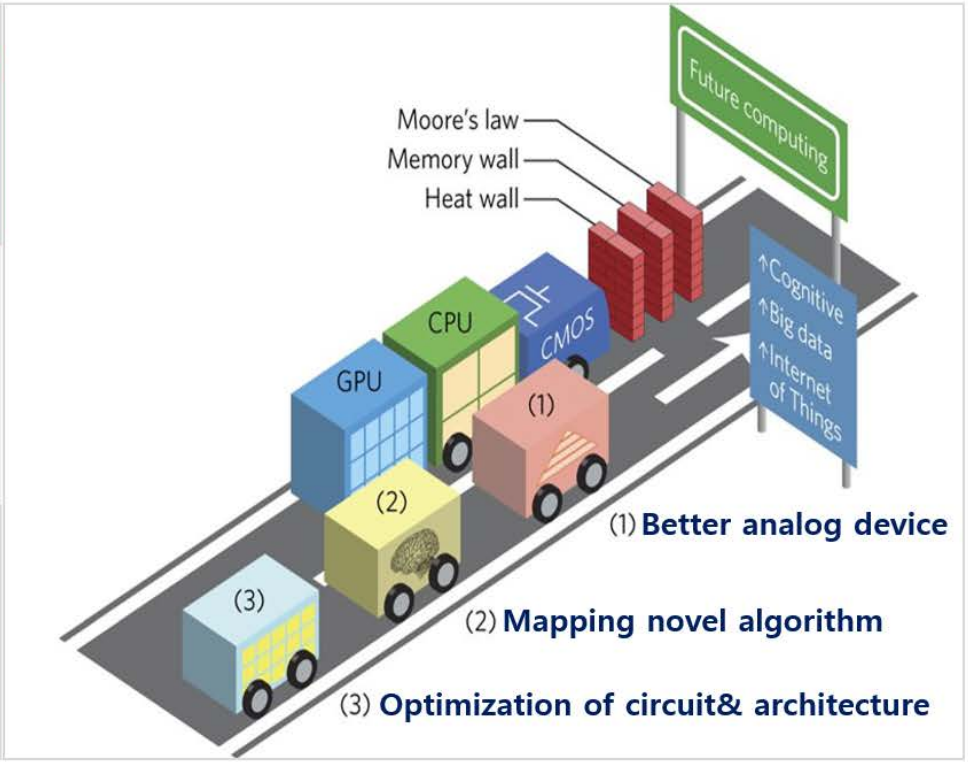
End of Moore's Law



Explosion of data usage

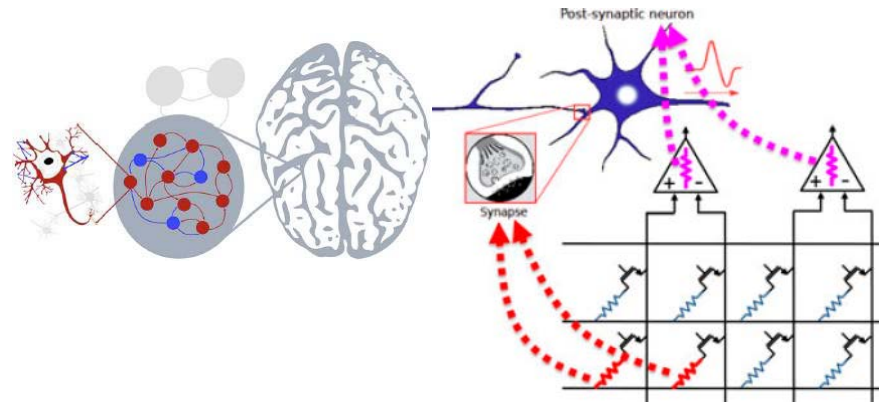
Von Neumann bottleneck





Neuromorphic computing system

→ One of the non-von Neumann approach,



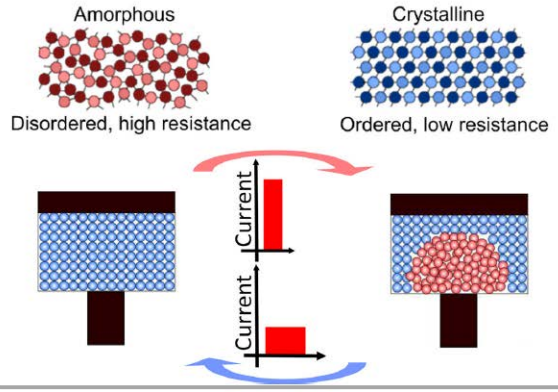
G. Burr, Adv. in Phys.:X, **2017**, *Vol. 2 No.1* , 89-124

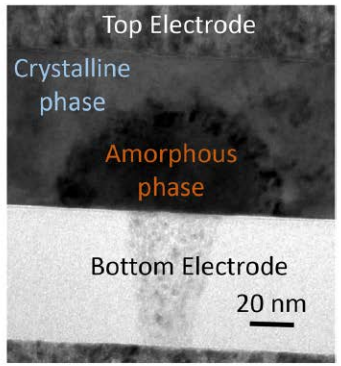
Novel architectures where memory and processing are collocated can perform efficiently in area/energy

Phase-change memory devices are used to build computational memory.

This novel architecture can accelerate the training of deep neural networks(DNN), and also it can play a key role in spiking neural networks(SNN).

Introduction – PCM device



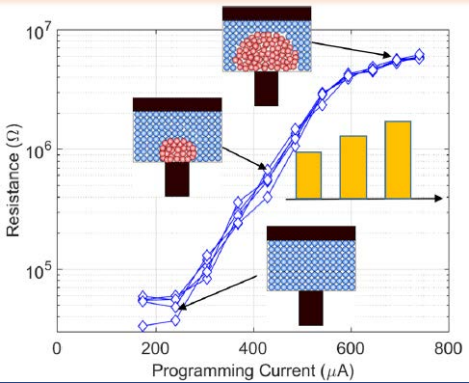


Phase change memory is one of the 'mature' non-volatile memory devices.

Typically, using Ge, Sb, Te chalcogenide compounds to program as memory.

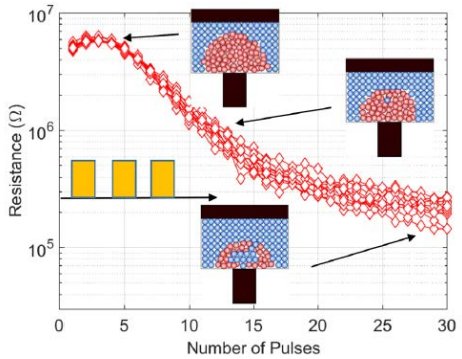
These certain materials exhibit drastically different electrical characteristics depending on their atomic arrangement.

Key properties for brain-inspired computing



1. Achieve not just two levels, but a continuum of resistance or conductance values

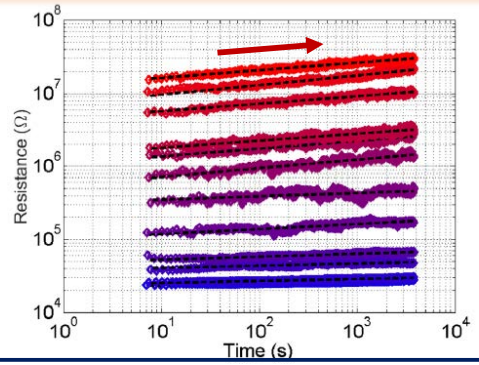
→Analog synaptic weight



2. Accumulative behavior, to achieve a linear increase in G as a function of the # of SET pulses

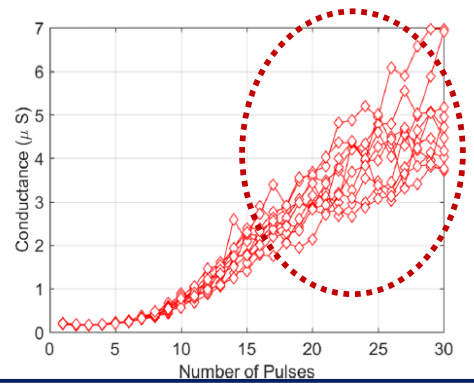
→ Controllable weight change

Drawback characteristics of PCM devices



- Resistance drift,

Temporal fluctuations of R from a spontaneous 'structural relaxation' of the amorphous phase



- Cycle-to-cycle randomness,

Due to the inherent stochasticity associated with the crystallization process

Outlines

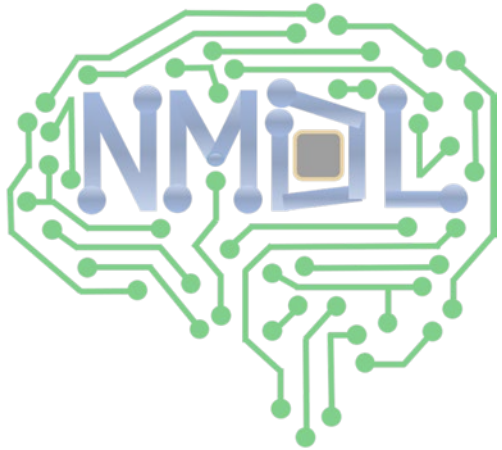
1. Introduction

2. Computational memory

3. Deep learning co-processors

4. Spiking neural networks (SNN)

5. Summary

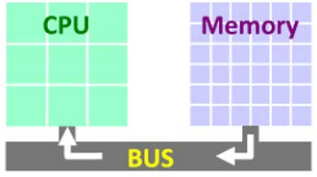


Novel computer architecture for AI

Samsung AI Forum, **2019**

SAMSUNG

SK hynix



H.S. Shin, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, VOL. 37, NO. 11, NOVEMBER 2018

In-memory computing,

Near-memory computing

NPU(Neural processing unit)

⋮

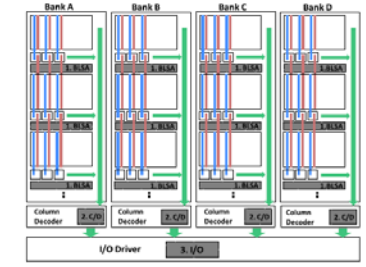


Fig. 1. MAC locations (BLSA, column decoder, and I/O driver).

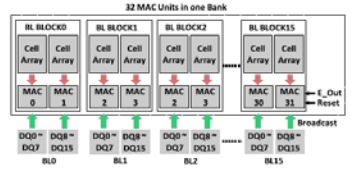
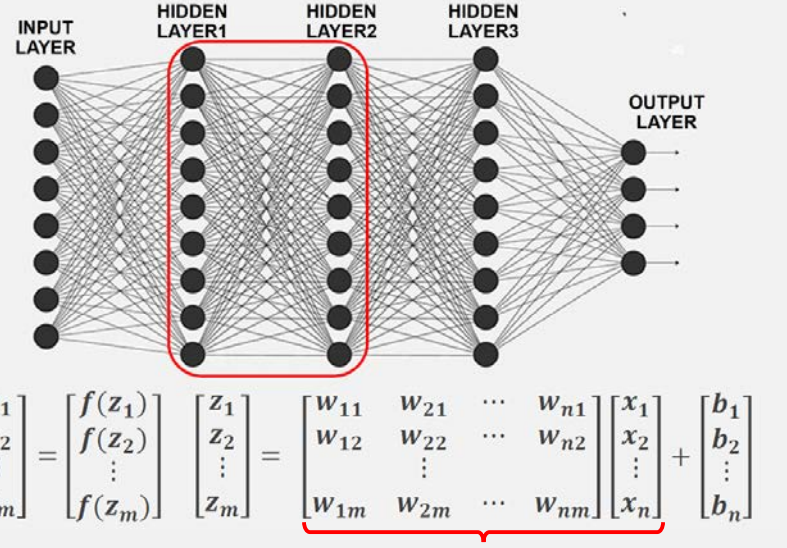


Fig. 4. 32 MAC units in a bank.

Why do we need a new computing architecture?

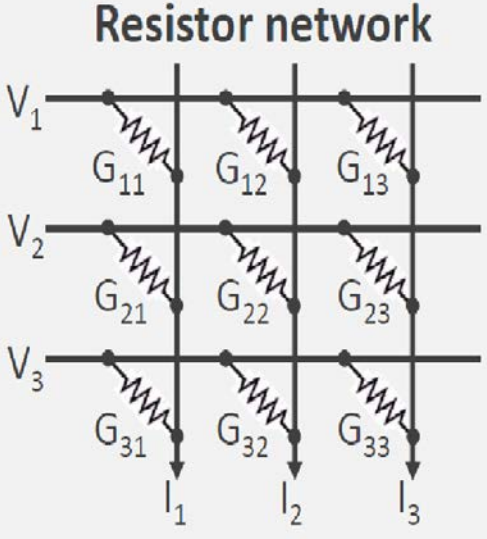
Computation in Artificial Neural Network(ANN)



$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_m) \end{bmatrix} \quad \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ & \vdots & & \\ w_{1m} & w_{2m} & \cdots & w_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Key Operation : Multiply-Accumulate, "MAC"

Key ideas in analog neuromorphic hardware

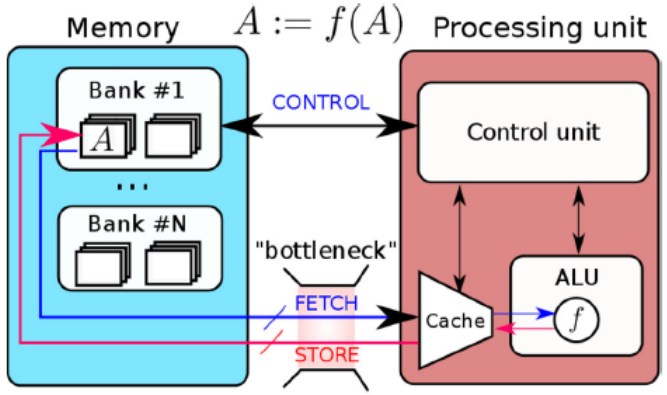


$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{21} & G_{31} \\ G_{12} & G_{22} & G_{32} \\ G_{13} & G_{23} & G_{33} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} G_{11}V_1 + G_{21}V_2 + G_{31}V_3 \\ G_{12}V_1 + G_{22}V_2 + G_{32}V_3 \\ G_{13}V_1 + G_{23}V_2 + G_{33}V_3 \end{bmatrix}$$

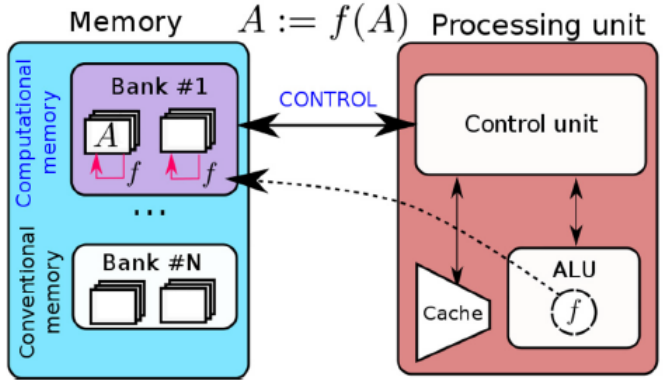
The physics of Ohm's law and Kirchhoff's current law allow the implementation of analog MAC operations in parallel.

In-memory computing with PCM devices

In-memory computing can overcome memory-processing unit bottleneck



Conventional Von Neumann architecture



Non-Von Neumann architecture

f(A) can be performed in memory which stores data A

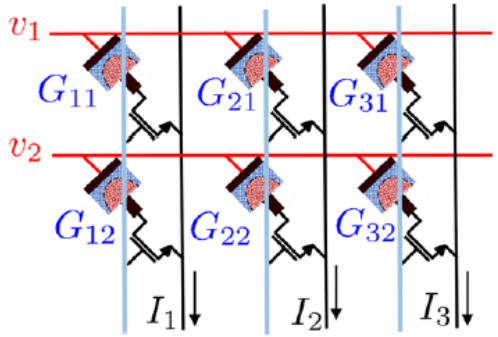
Saves energy and time!

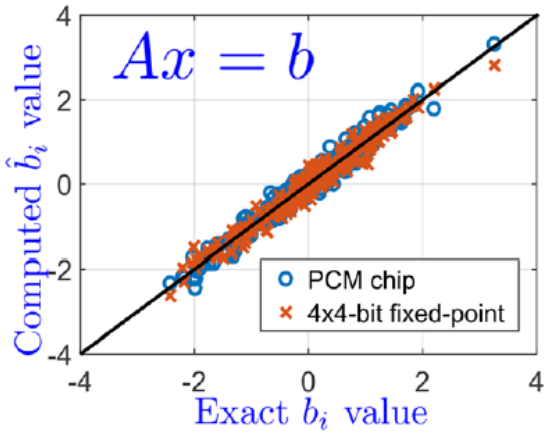
Especially, in pattern recognition, real-time recognition, and intuitive functions!

Nature Comms., **2017**, *8*, 1115

- Matrix-vector multiplication using PCM devices

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





Accuracy of the computation with PCM is comparable to that of fixed-point digital.

Outlines

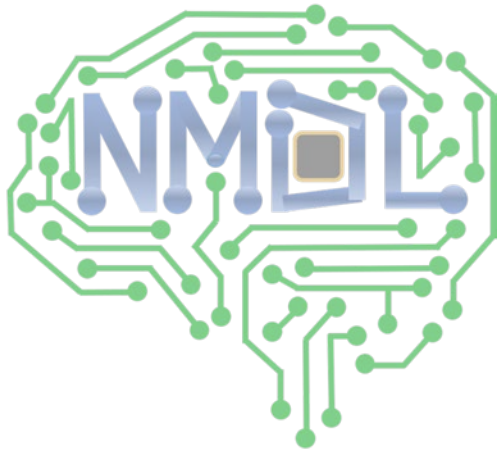
1. Introduction

2. Computational memory

3. Deep learning co-processors

4. Spiking neural networks (SNN)

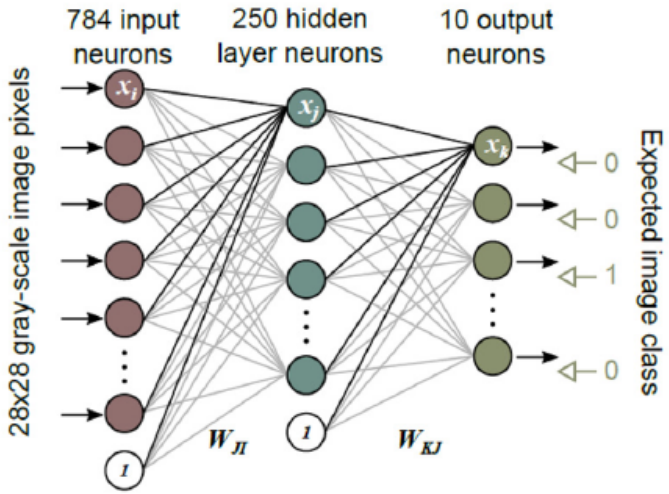
5. Summary



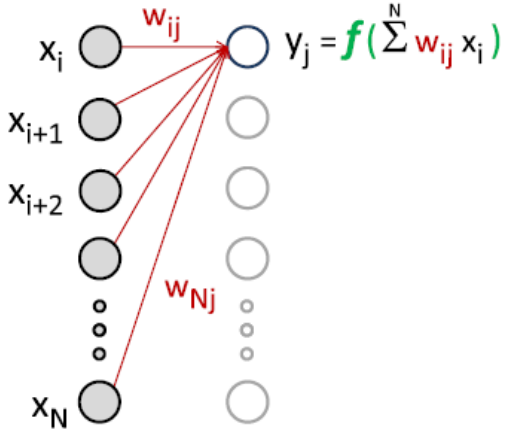
Analog memory-based DNN process

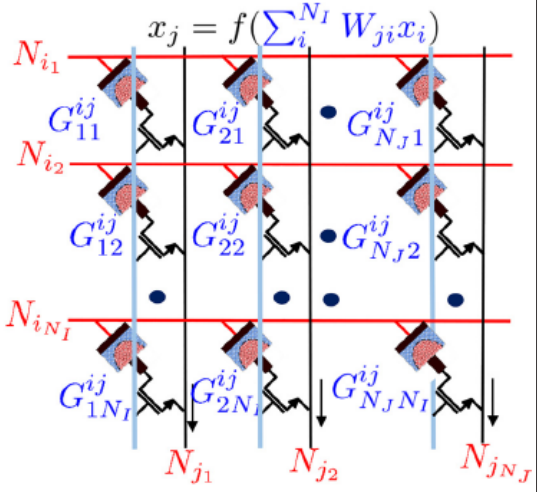
- Deep Learning (Deep neural network, DNN) on PCM hardware(cross-bar array)

J. Appl. Phys., **2018**, *124*, 283011

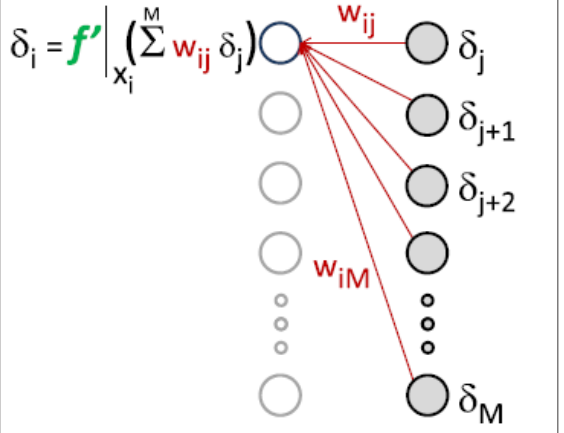


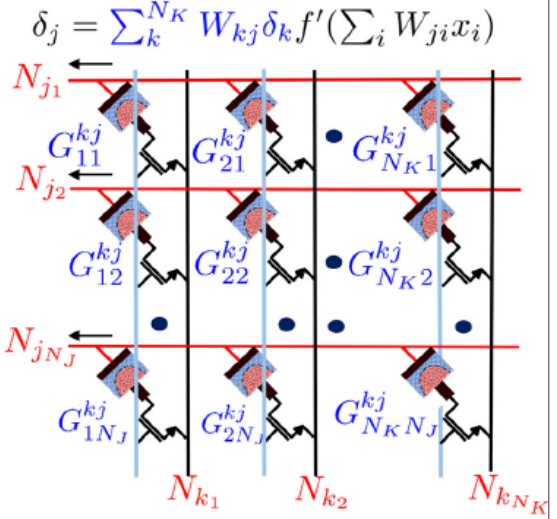
Forward-inference



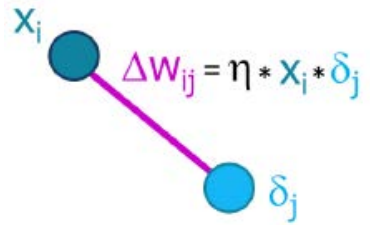


Back-propagation

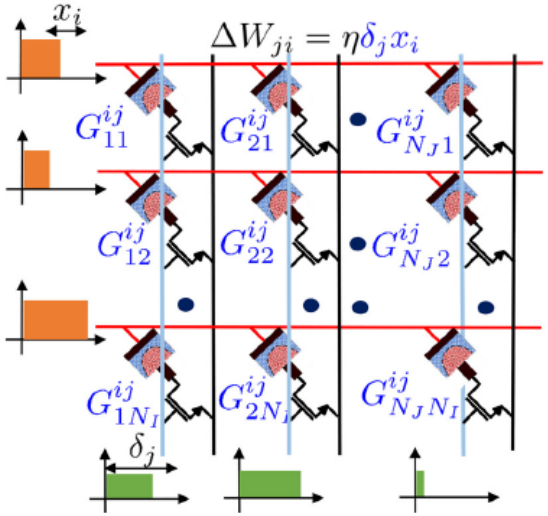




Synaptic weight update

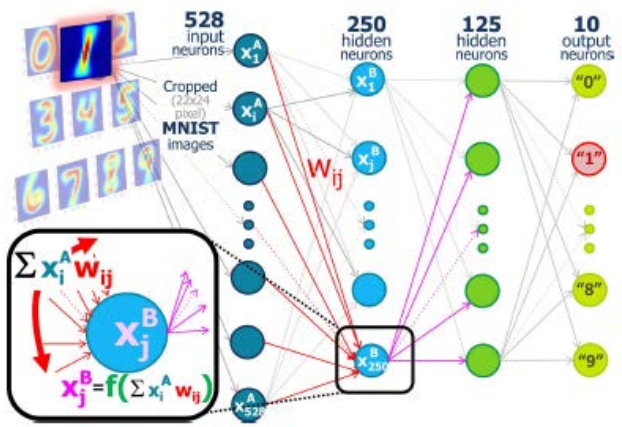


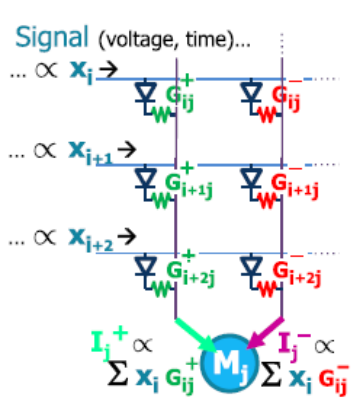
Cost(error) can be minimized according to learning rule, such as gradient decent rule

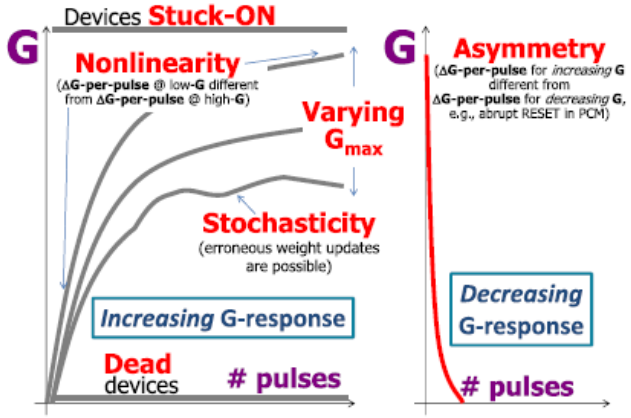


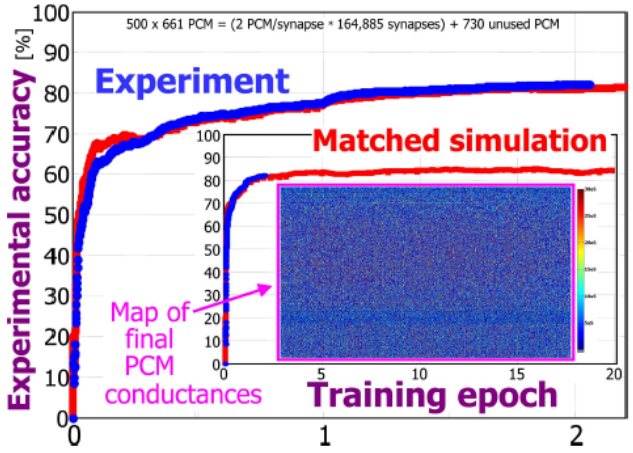
Demonstration of DNN using PCM devices

- **MNIST handwritten image recognition**



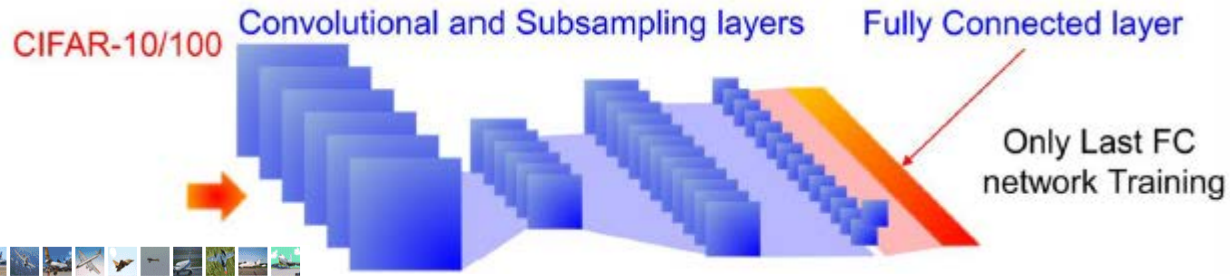


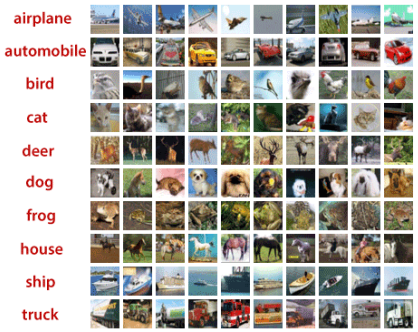


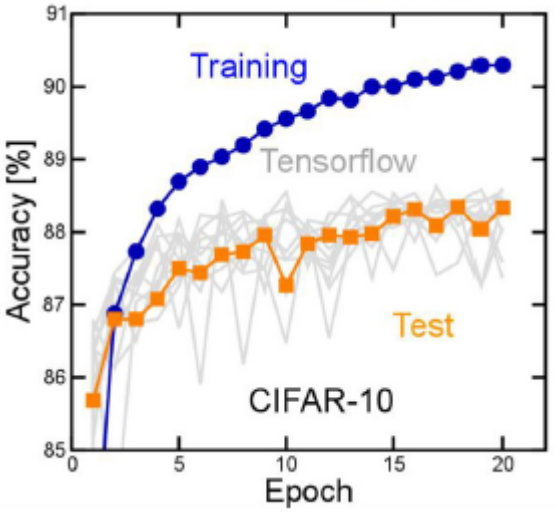


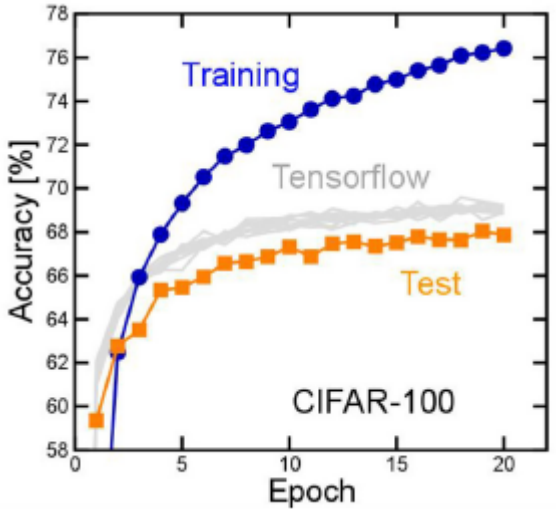
G^+, G^- approach used to overcome PCM's inherent shortcomings

- **CIFAR-10/100 image recognition (Using Convolutional layer)**



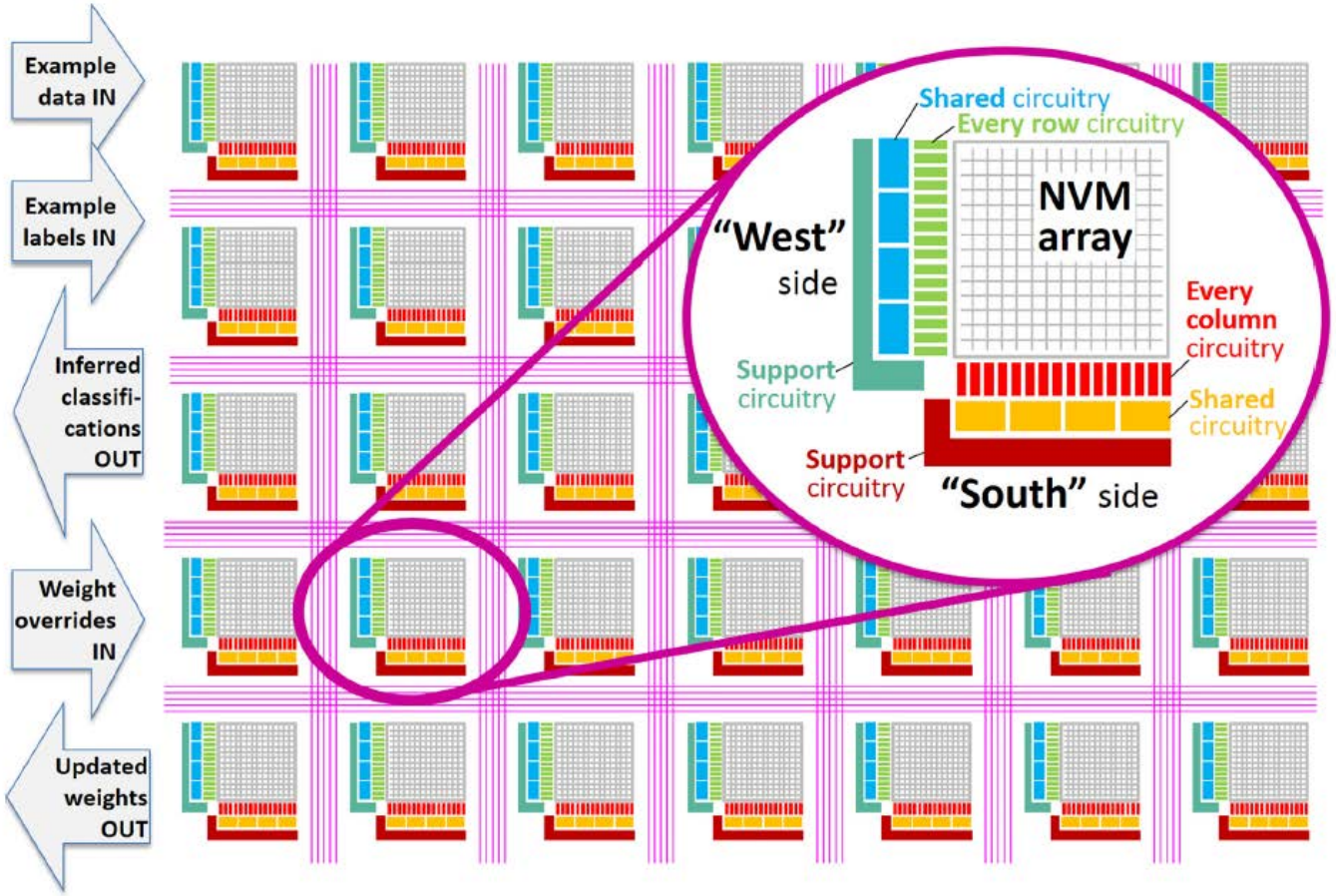


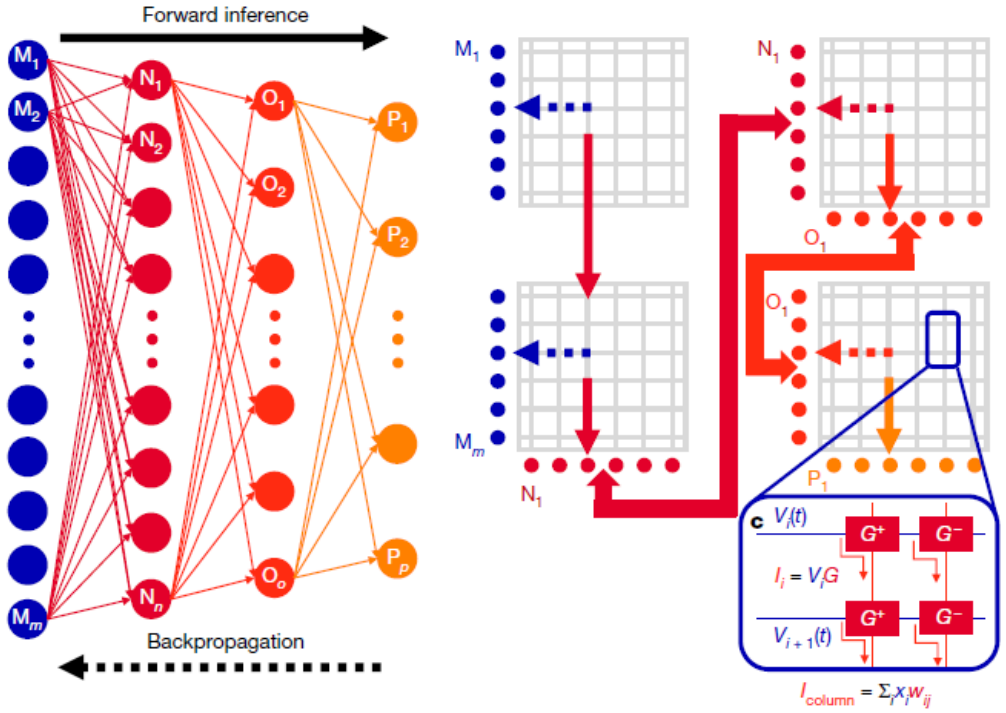




A proposed chip architecture

- Chip architecture for a co-processor for deep learning based on PCM arrays





G. Burr et al.,Nature, **2018**, vol *8*, 60-67

Outlines

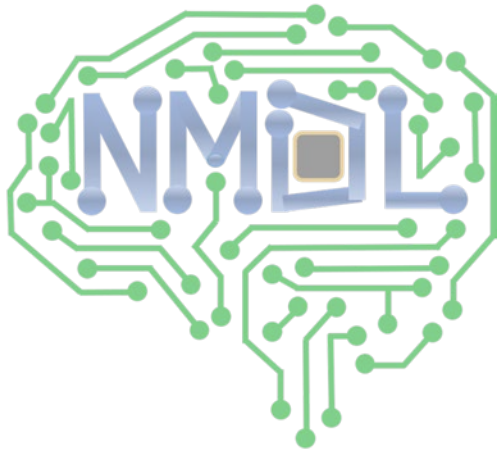
1. Introduction

2. Computational memory

3. Deep learning co-processors

4. Spiking neural networks (SNN)

5. Summary



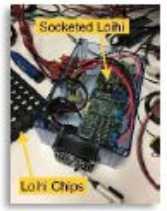
Spiking Neural Networks (SNN)

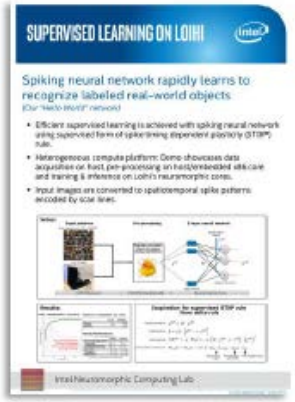
POINT OF VIEW

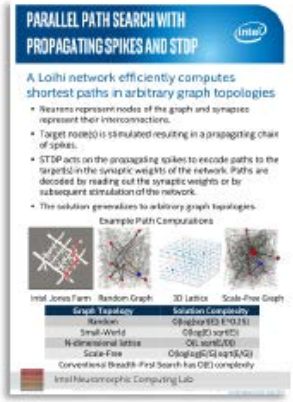
To Spike or Not to Spike: That Is the Question

By WOLFGANG MAASS
Institute for Theoretical Computer Science, Graz University of Technology, Graz 8010, Austria

Vol. 103, No. 12, December 2015 | PROCEEDINGS OF THE IEEE

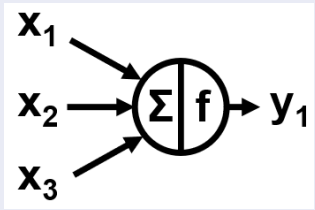
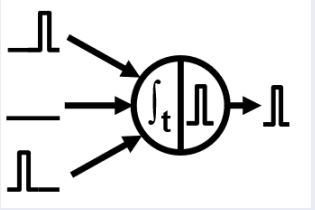


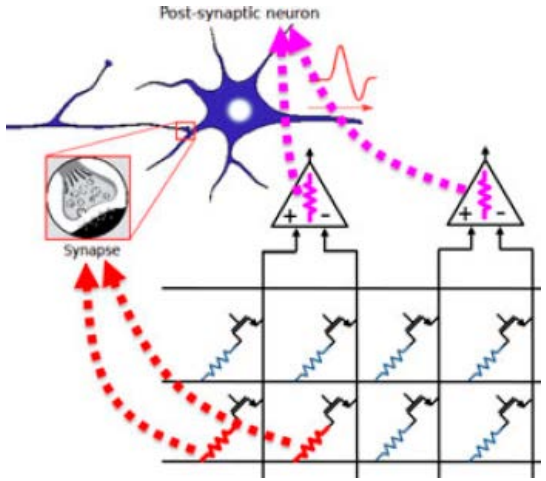




M. Davies, NICE, **2018**, Loihi intro talk

- **Third generation of neural network?**

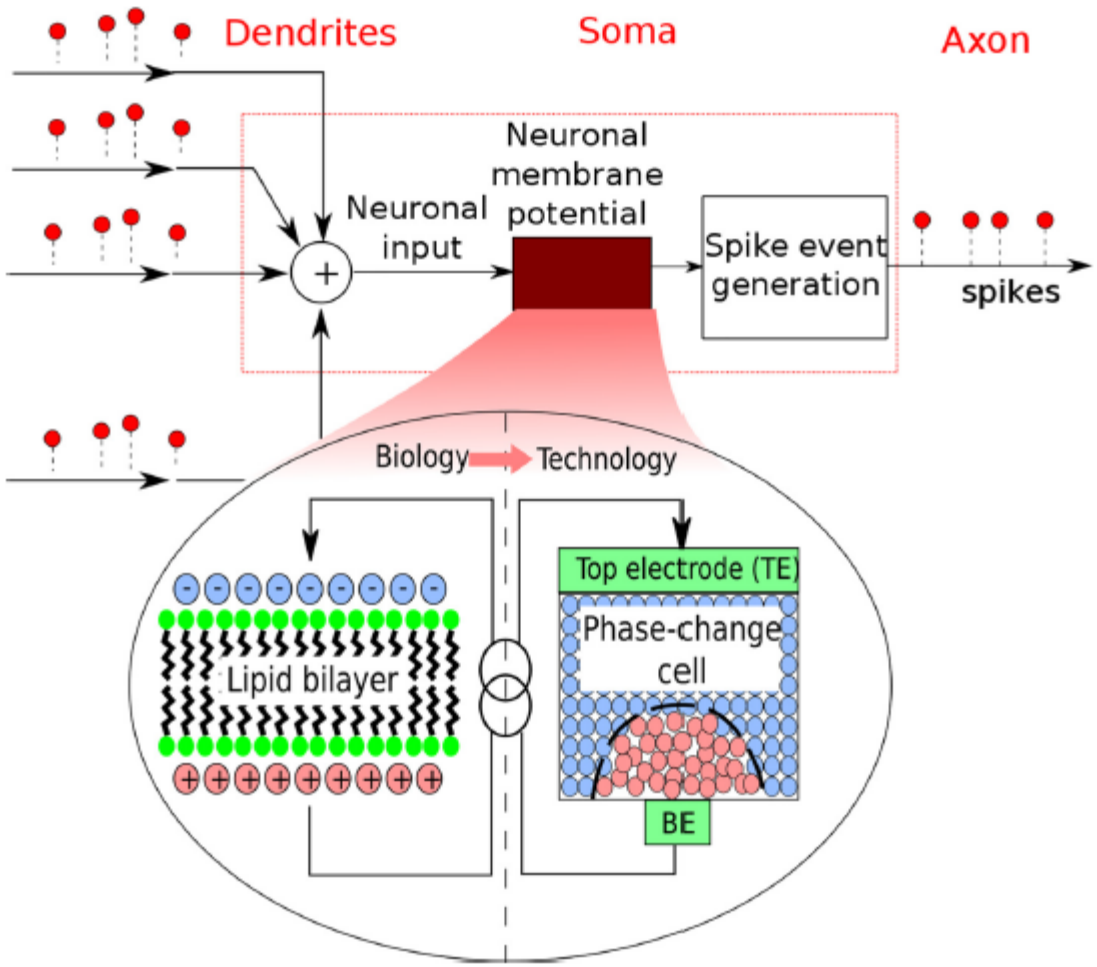
	ANN : Non-spiking NN	Brain : Spiking NN
Inputs & Outputs	Real-valued numbers	Spikes (encoded)
Neuron Operation		

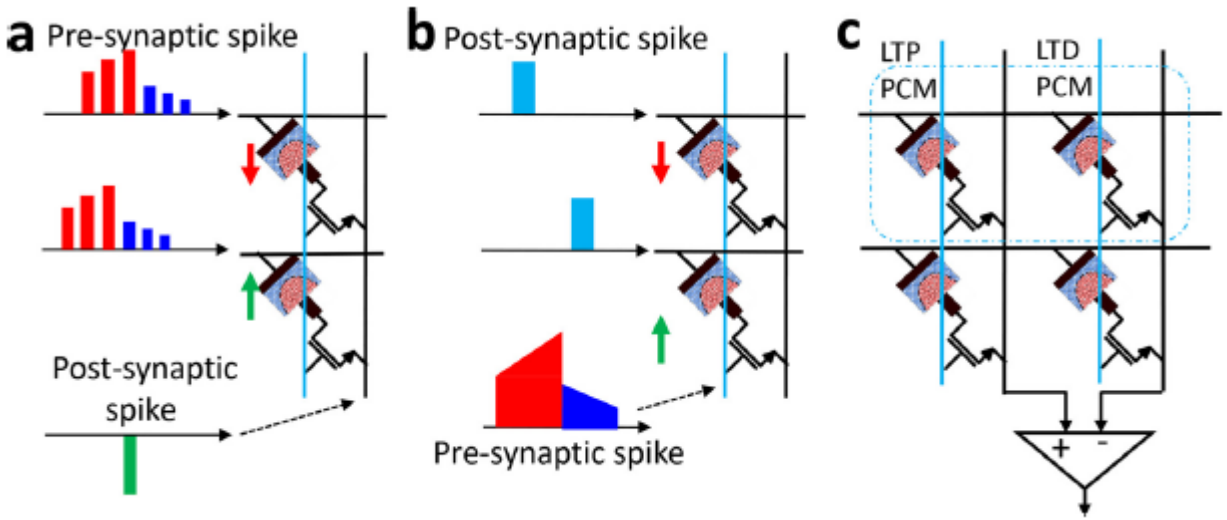


G. Burr, Adv. in Phys.:X, **2017**, *Vol. 2 No.1* , 89-124

Neuron?
Synapse?

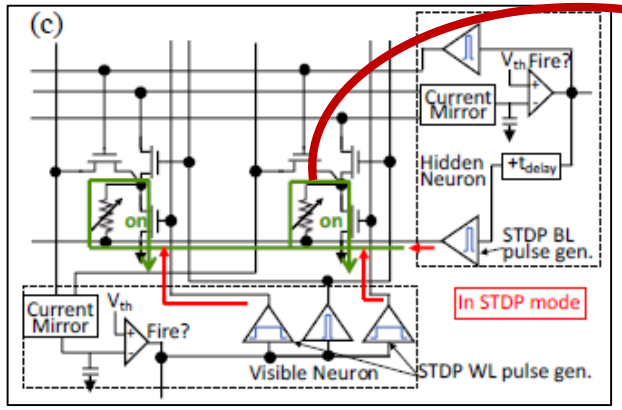
Neuronal dynamics implementation using PCMs



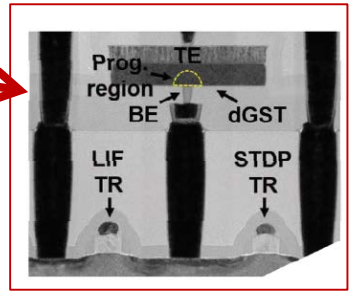


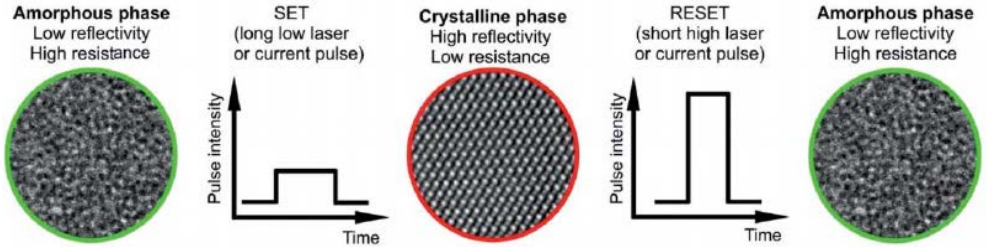
Hardware concept for PCM programming & LIF

[Phase change memory]

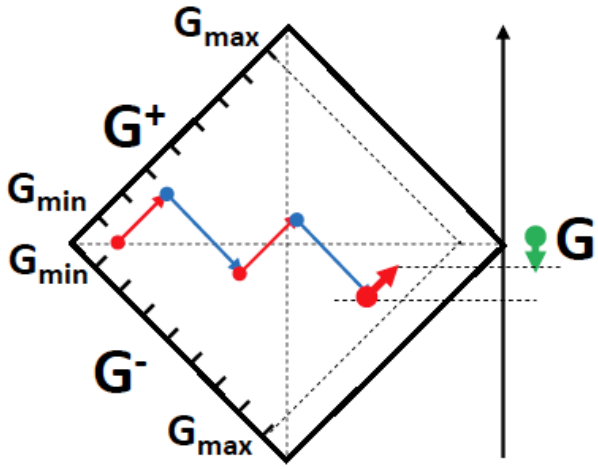


M. Ishii *et al., in IEDM*, **2019**, pp. 14.2.1-4.

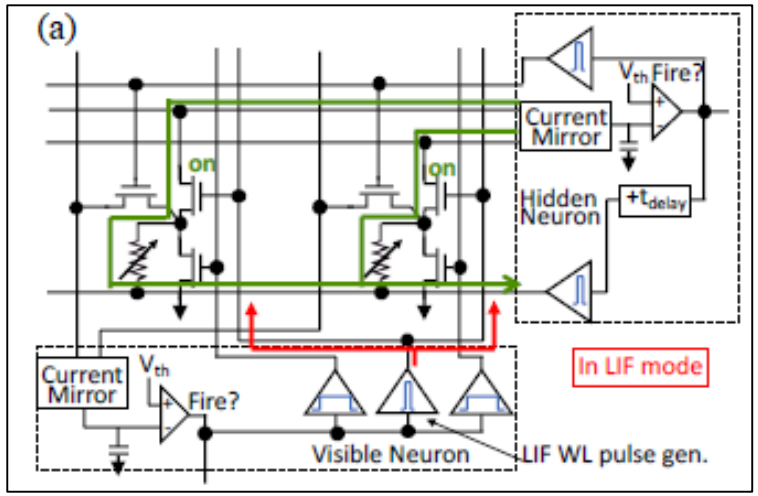




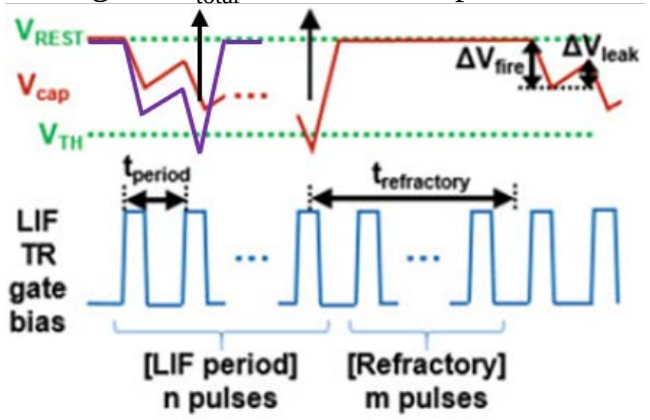
[Bipolar synaptic weights]



[LIF (Leaky-integrate & fire) process]

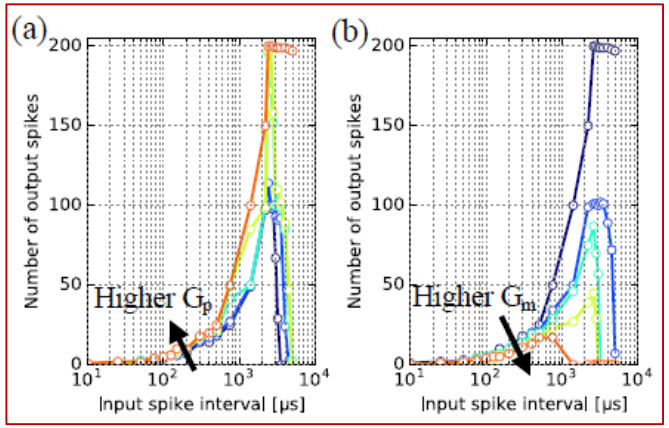


Higher G_{total} makes more spikes!



{ SET G_p → G_{total} increase → LIF outputs ↑
G_m → G_{total} decrease → LIF outputs ↓ }

($G_{total} = G_p - G_m$)



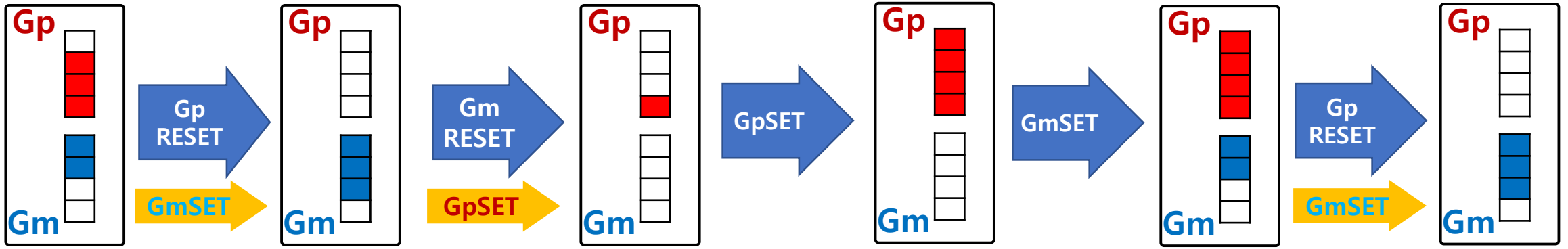
Gradual PGM & LIF results

[PGM & LIF test flow]

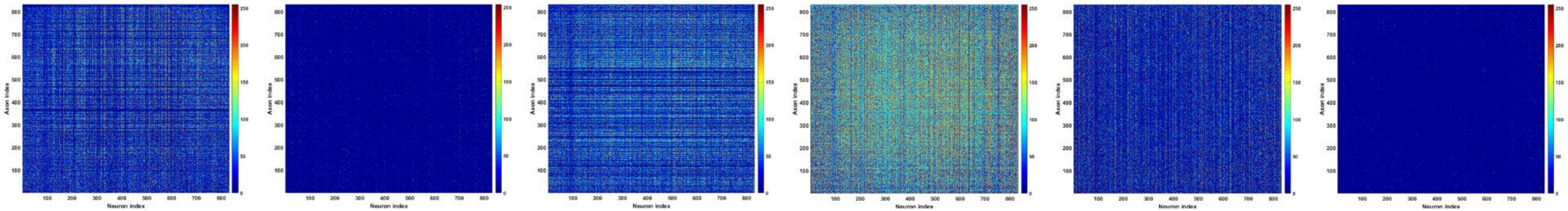
Virgin (Initial state?) → All RESET x3 (Clean up before the test) → Gp SET × 10 (Increased?) → Gm SET × 5 (Decreased?) → All RESET x3 (Clean up after the test)

[Schematics of the programming] SET | RESET

SET : SET processes occur inevitably during the RESET process.

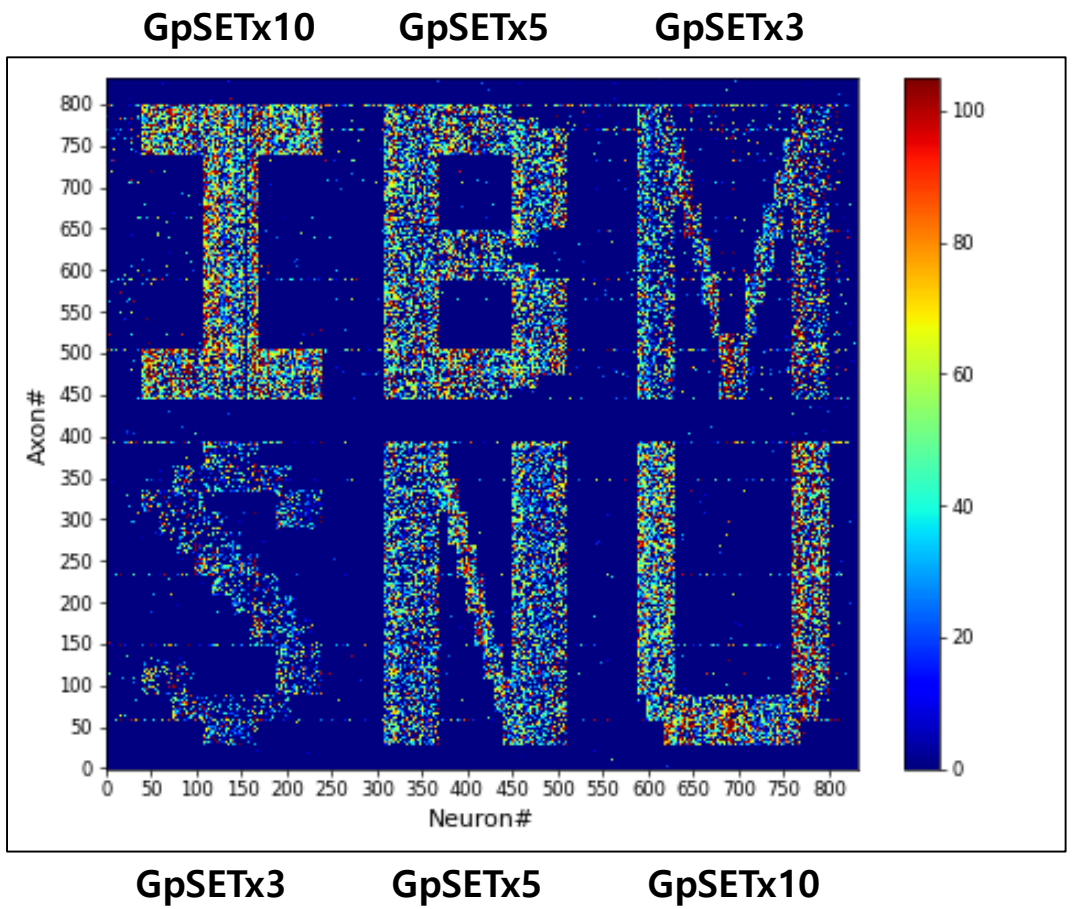


[LIF results of the programming]



From the results, we can program all PCM cells gradually on both Gp & Gm

Controllable weight update



We can program the synaptic array with accurate position so that able to do 'pixel art'.

And we are able to confirm that how the conductance of each cell affects hardware with basic operation, LIF.

Even we don't know accurate conductance of each synaptic cell, we can expect the 'actual weight'!

This programming technology can be applied to implementation of 'weight transfer', which is necessary for 'off-chip learning'.

Summary

Neuromorphic computing using phase-change memory devices

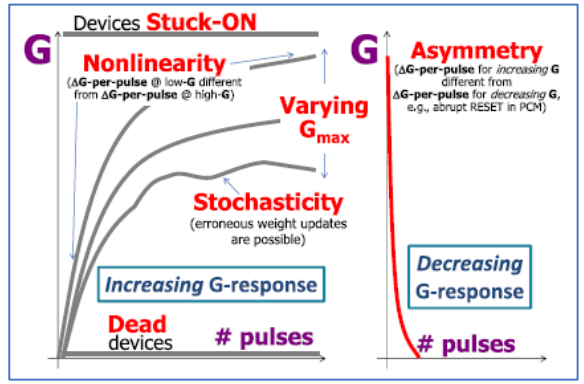
Brain-inspired computing schemes is promising technology to overcome 'von Neumann architecture' in AI computing.

Phase-change memory device, one of the 'mature' non-volatile memory technologies
can achieve significantly higher performance compared to conventional computation with
ⅰ) Multi-conductance level, ⅱ) Accumulative behavior.

Computational memory, In-memory computing, Near-memory computing
are novel approaches to mitigate 'von-Neumann bottleneck'.

These massively parallel computing units with PCMs can perform in application of

ⅰ) Deep-learning, Deep Neural Network(DNN)
ⅱ) Spiking Neural Network(SNN)



But there are challenges towards the adoption of PCM-based computing systems in future AI hardware,
ⅰ) Phase-change memory device – conductance fluctuation, non-linearity & non-symmetric…
ⅱ) Spiking Neural Network(SNN) – novel learning algorithms…

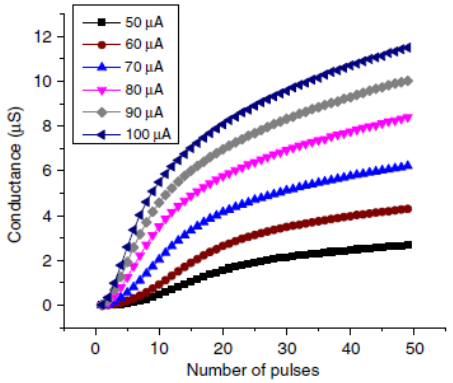
Thank you for listening!

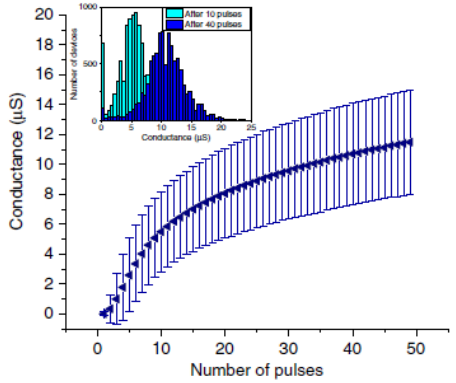
Demonstration of in-memory computing with PCM

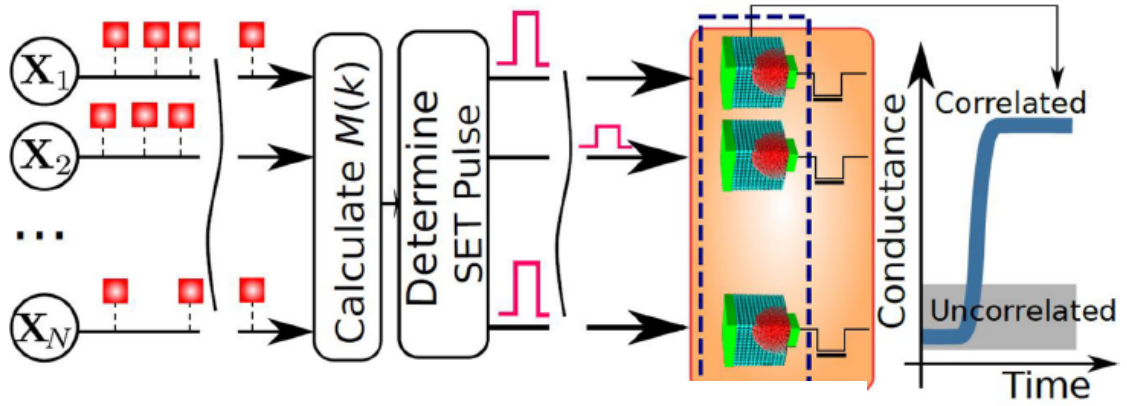
- **Unsupervised learning of temporal correlations**

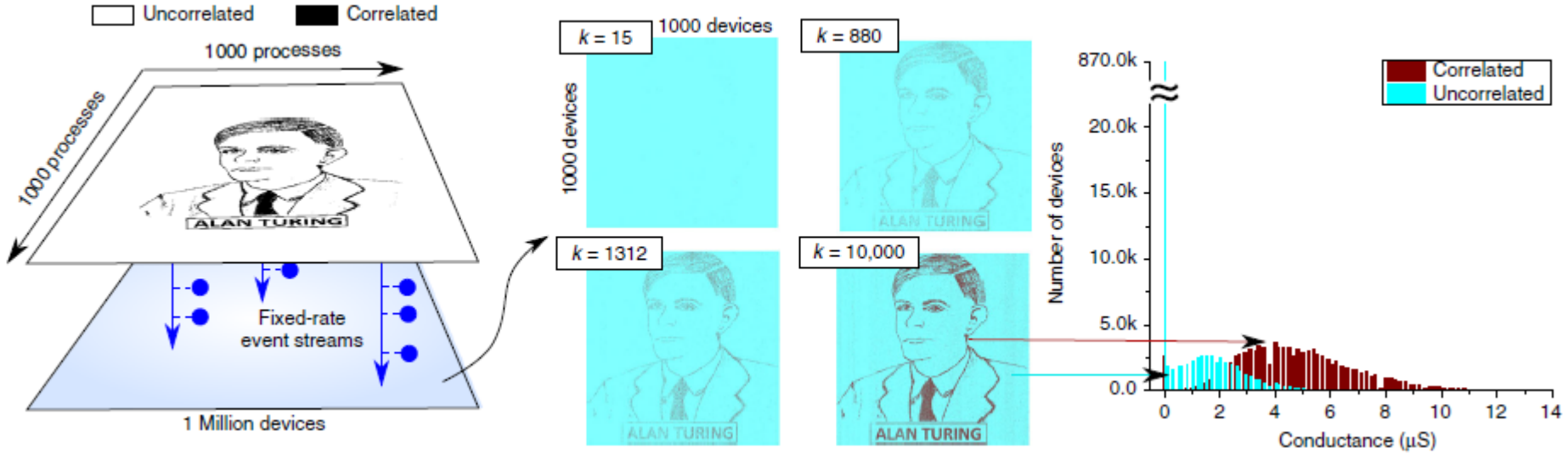
Nature Comms., **2017**, *8*, 1115

➡ Using the accumulative behavior of the PCM devices,









➡ Computation can be accelerated by a factor of 200 relative to using 4 GPU devices,
And also energy saving is over two orders of magnitude. (Assumed the PCM write latency of 100ns and SET programming energy of 1.5pJ