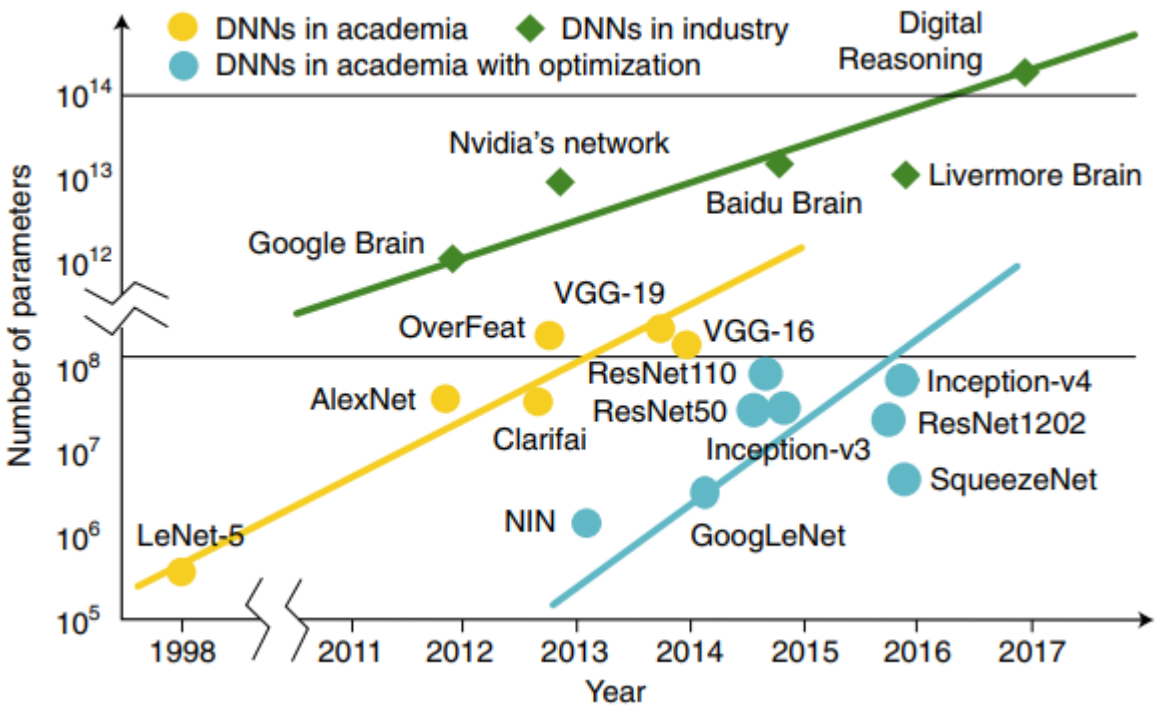
Device-Algorithm Co-Optimization for Analog in-Memory Deep Learning

Sangbum Kim
Associate professor
Department of Materials Science and Engineering
Seoul National University

sangbum.kim@snu.ac.kr

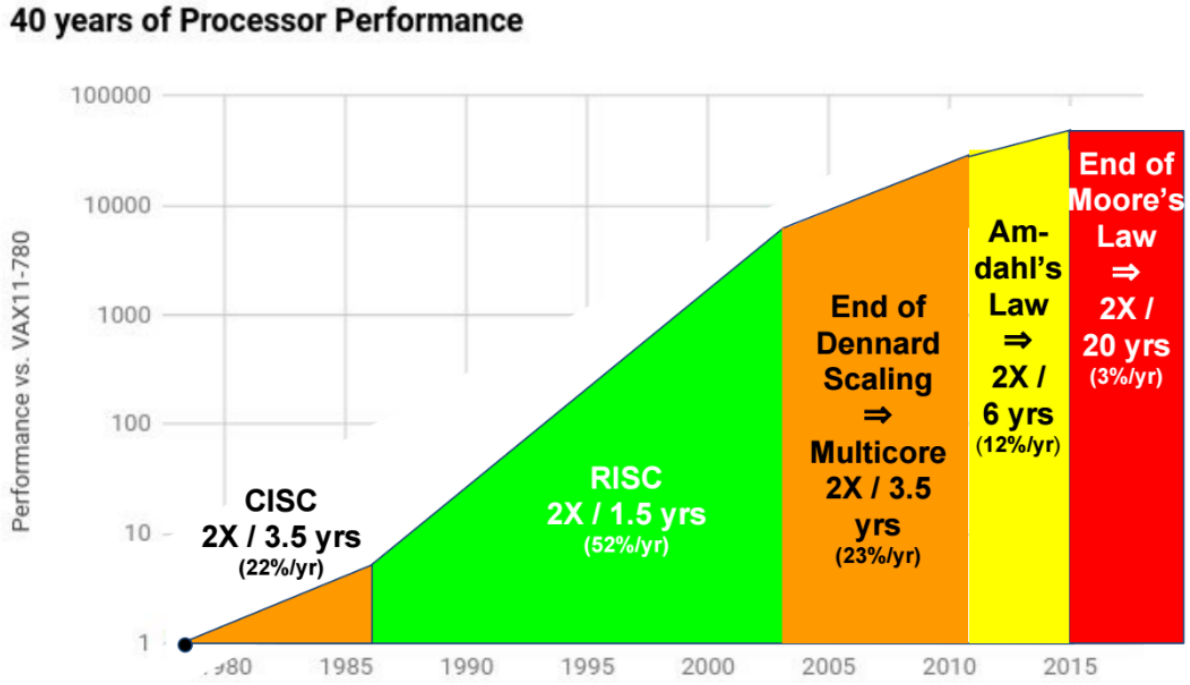
Need for Deep Learning Accelerator

SW: 2X growth every 3.5 months



X. Xu, Nature Electronics, 2018.

HW: 2X growth every 18 months



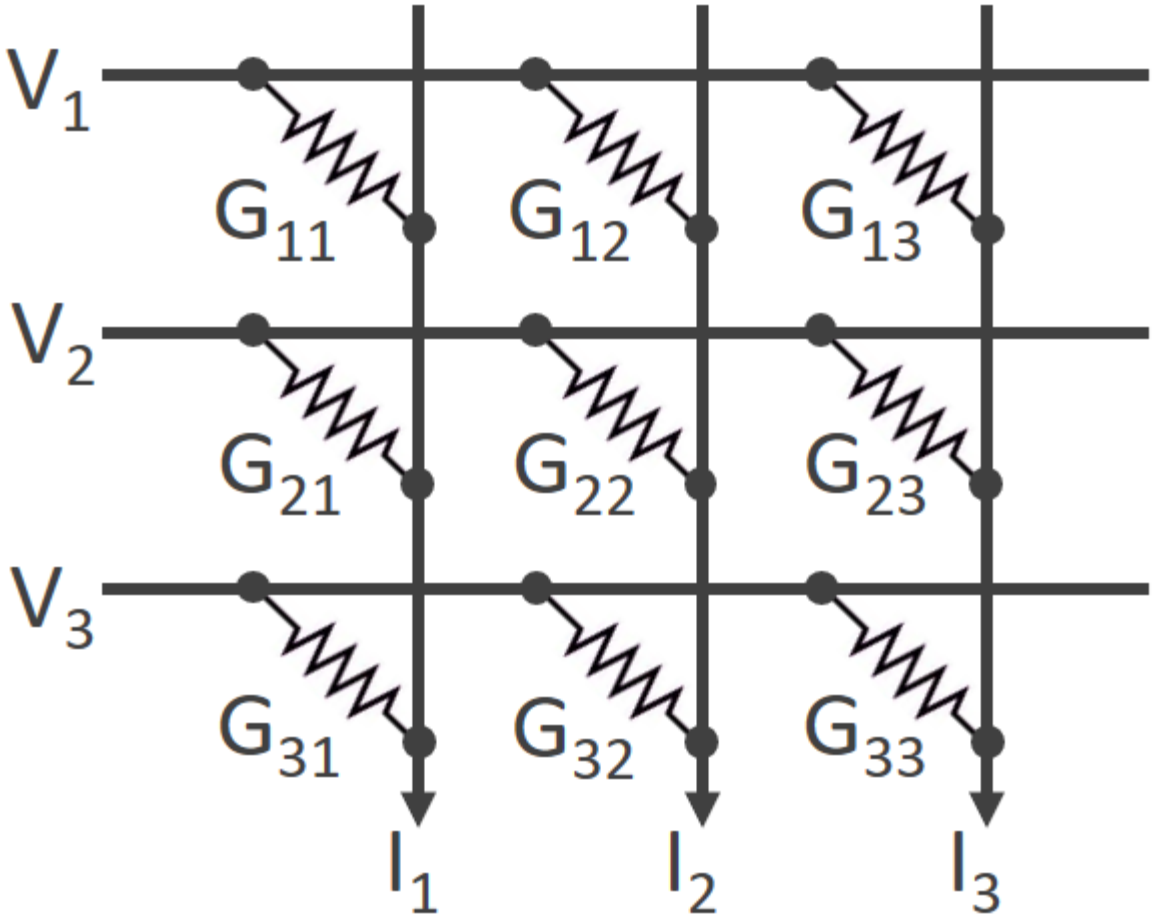
D. Patterson, NAE Regional Meeting, 2017.

- The number of parameters in DL neural network is increasing exponentially.
  - So does the required computing power.

⇔

- Moore's Law is ending (or already ended): No more processor performance growth.

Resistor network



We need synaptic devices with weight update symmetry and linearity.

Matrix-vector multiplication using Kirchhoff's law

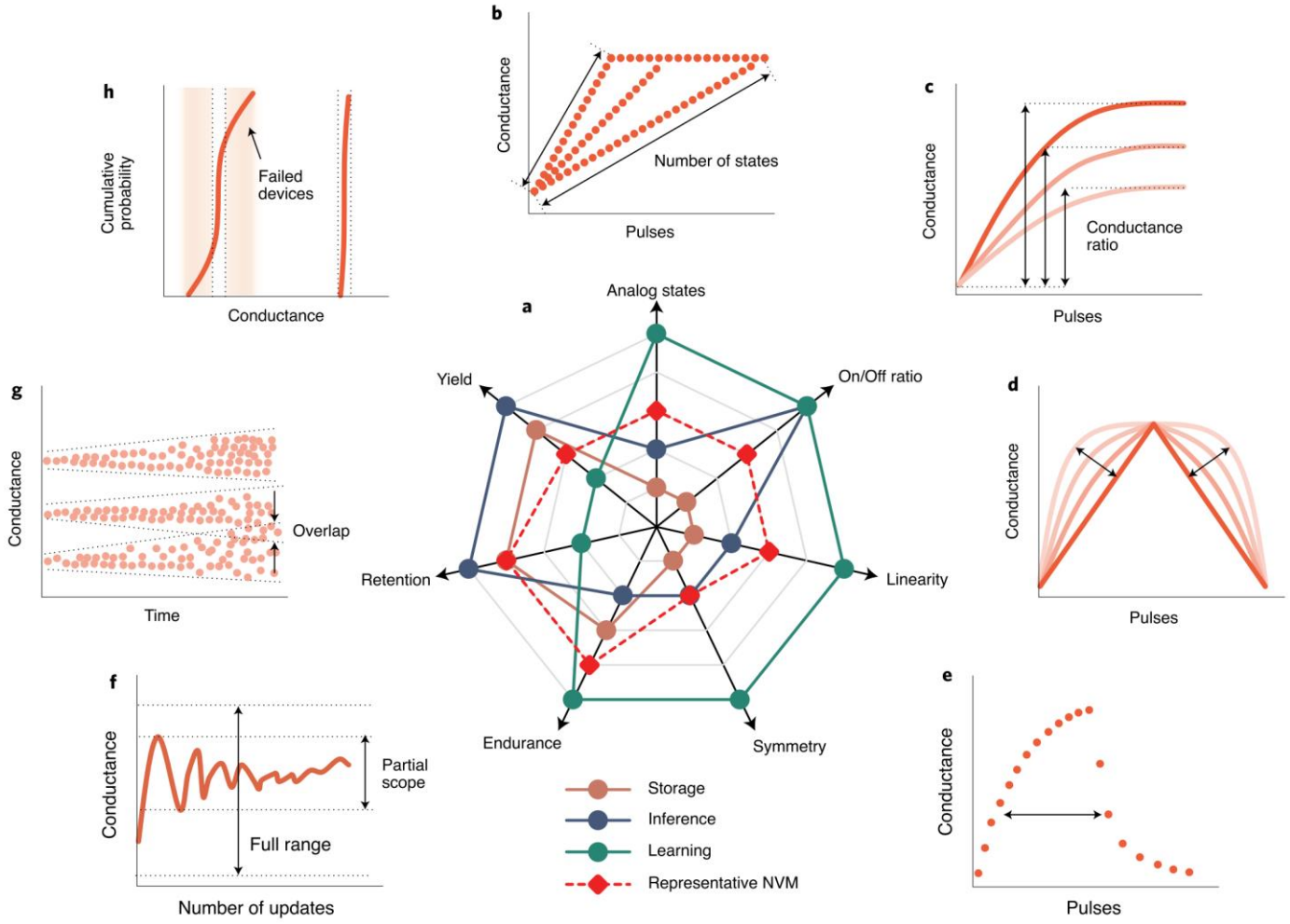
$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{21} & G_{31} \\ G_{12} & G_{22} & G_{32} \\ G_{13} & G_{23} & G_{33} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$$

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

Back-propagation weigh update operations in one layer ($n \times m$ synapse)

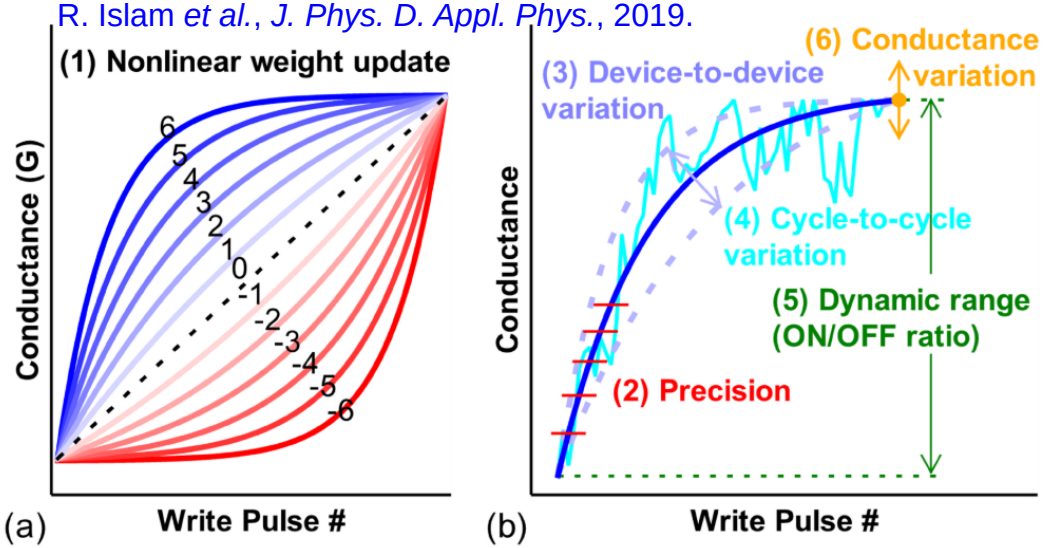
$$\begin{bmatrix} \Delta w_{11} & \cdots & \Delta w_{n1} \\ \vdots & \ddots & \vdots \\ \Delta w_{1m} & \cdots & \Delta w_{nm} \end{bmatrix} = \eta \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \times \begin{bmatrix} q_1 \\ \vdots \\ q_m \end{bmatrix}$$

Various Challenges at the Synaptic Device Level



W. Zhang *et al.*, "Neuro-inspired computing chips," *Nat. Electron.*, vol. 3, pp. 371–382, 2020, doi: 10.1038/s41928-020-0435-7.

R. Islam *et al.*, *J. Phys. D. Appl. Phys.*, 2019.



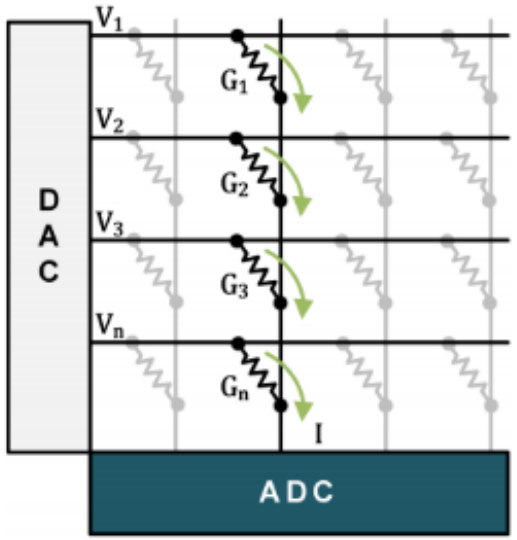
Specs	Parameter	Value	Tolerance
Average device resistance	R_{device}	24 $M\Omega$	7 $M\Omega$
Resistance on/off ratio	$\max(g_{ij})/\min(g_{ij})$	8	
Resistance change at $\pm V_S$	$\Delta g^{\pm}_{min}$	100 $K\Omega$	30 $K\Omega$
Resistance change at $\pm V_S/2$		10 $K\Omega$	
Storage capacity	$(\max(g_{ij}) - \min(g_{ij}))/\Delta g_{min}$	1000 levels	
Device up/down asymmetry*	$\Delta g^{+}_{min}/\Delta g^{-}_{min}$	1.05	2%

T. Gokmen *et al.*, Front. Neurosci., 2016.

- Various requirements: It is very challenging to meet them all.

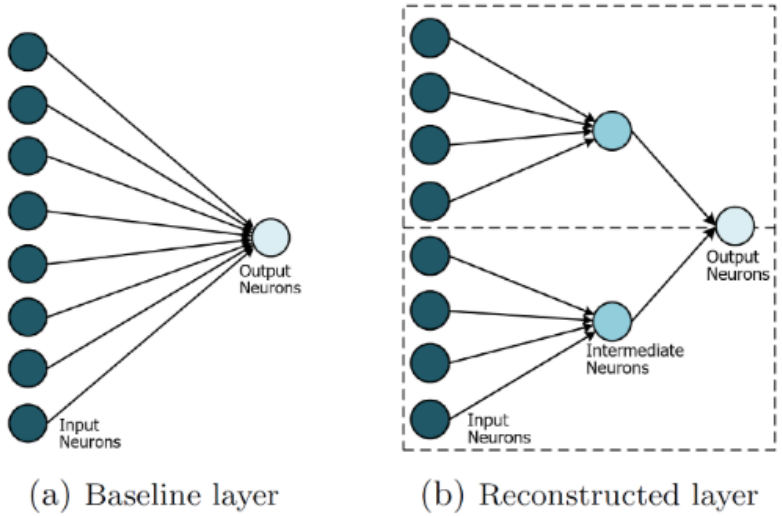
Various Challenges at the Circuit and Architecture Level

- Conversion to/from analog
  - ADC and DAC consume large area and power.



Naive solution: If you have to share one DAC (ADC) among multiple word lines (bit lines), you lose the massive parallelism.

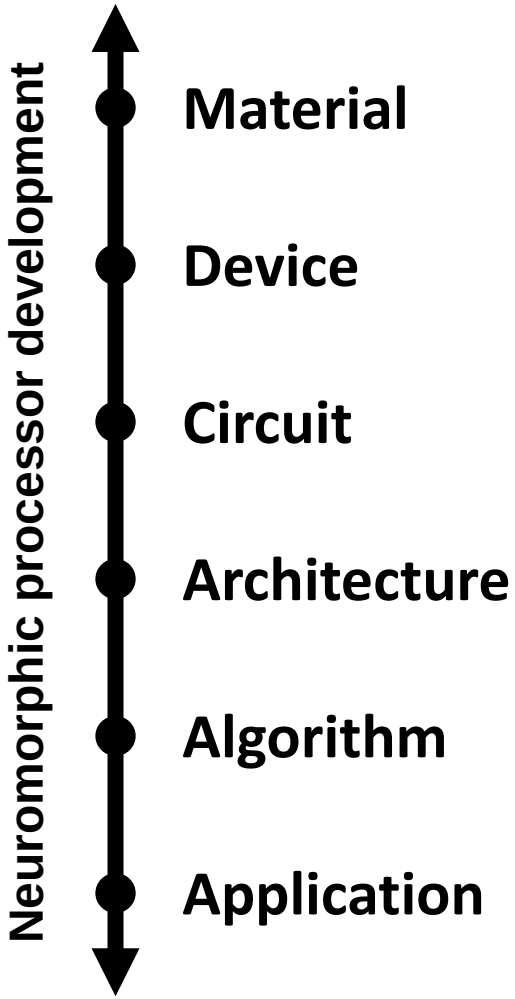
- Array size mismatch
  - Weight matrix size ≠ Synaptic array size



Y. Kim et al, "Neural Network-Hardware Co-design for Scalable RRAM-based BNN Accelerators".

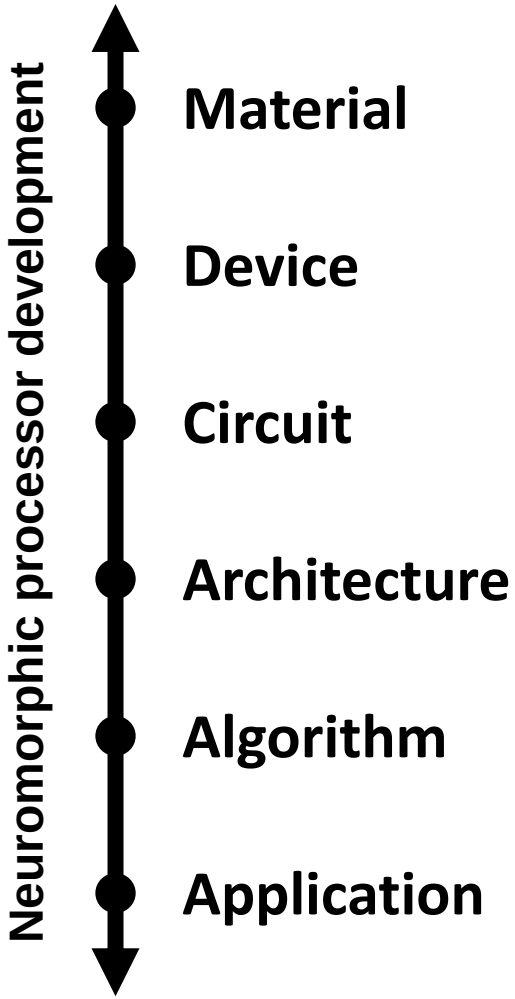
Input to the neural network should be split with weight retraining. Or vast number of synapses will be unused.

Importance of Co-optimization



- *Multidisciplinary research* ranging from materials to machine learning applications is required!

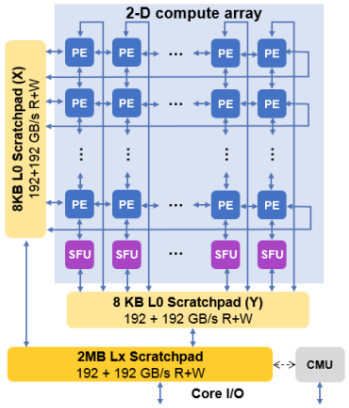
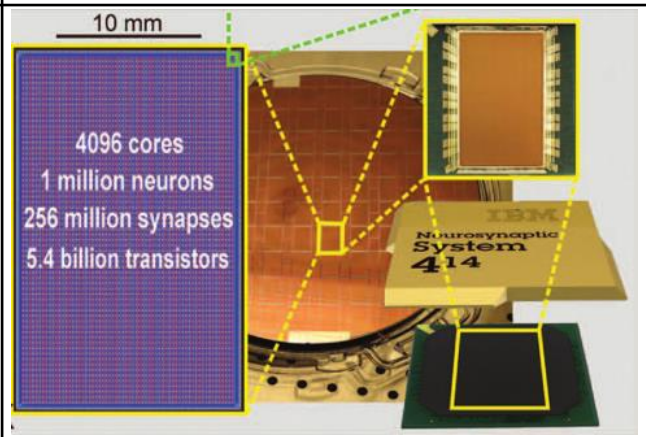
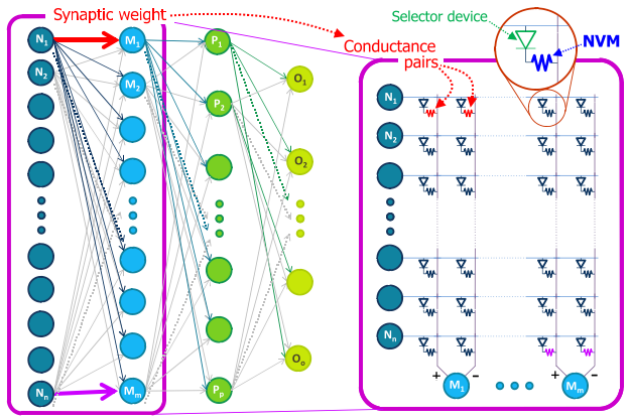
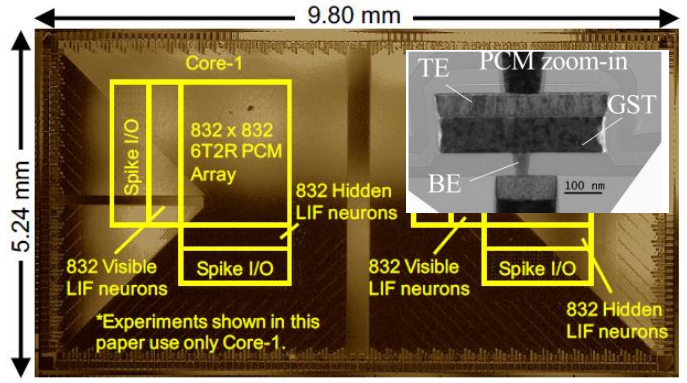
Importance of Co-optimization



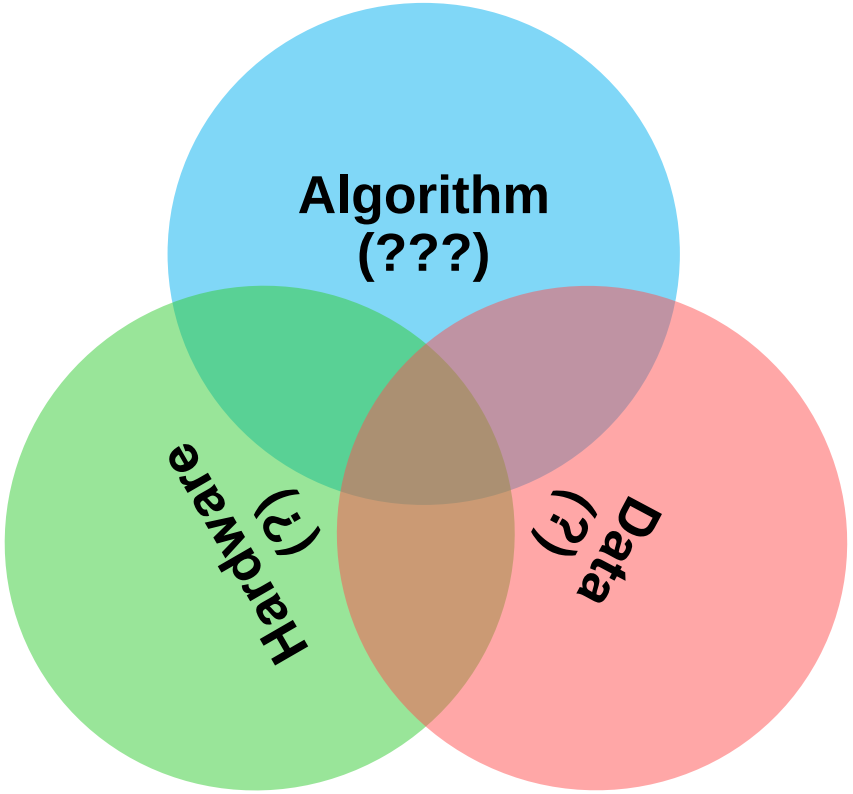
Can we take advantage of non-ideal characteristics of devices?

- *Multidisciplinary research* ranging from materials to machine learning applications is required!

Neuromorphic processor landscape

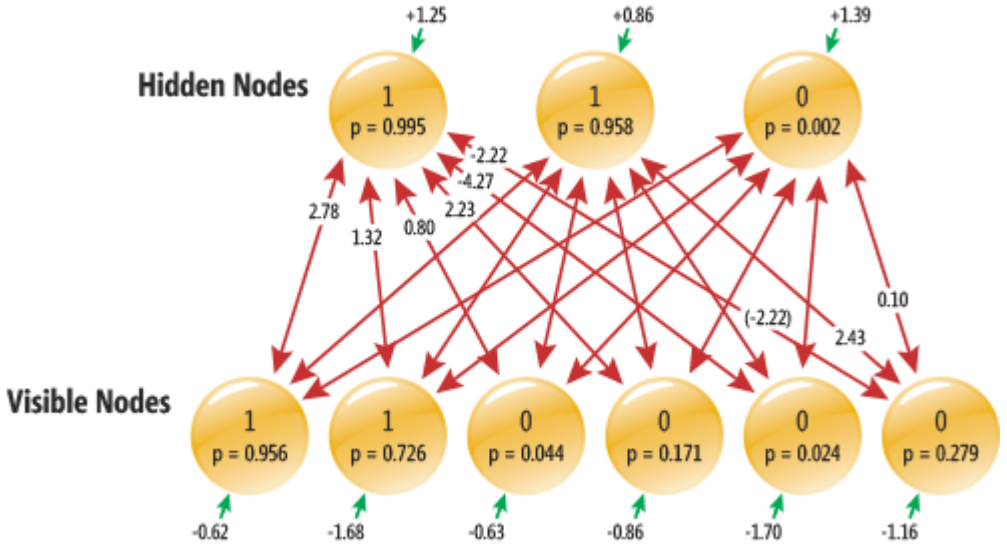
	Artificial neural network (ANN)	Spiking neural network (SNN)
Conventional digital memory device	 ❑ A scalable Multi-TeraOPS Deep Learning Processor ○ SRAM ○ Deep learning training and inference ○ Programmable architecture and custom ISA *B. Fleischer et al., VLSI Circuits, 2018.*	 ❑ TrueNorth ○ SRAM ○ Off-chip learning *P. Merolla et al., Science, 2014.*
Analog non-volatile memory (NVM) device	 ❑ NVM ML accelerator ○ PCM ○ On-chip learning *G. Burr et al., IEDM, 2014.*	 ❑ NVM low-power ML ○ PCM ○ On-chip learning *M. Ishii, S. Kim et al., IEDM, 2019.*

Can Spiking NN take off?



- All three key components are currently missing in SNN.
  - Some initial efforts have shown some promises.

RBM implementation on SNN



E. Neftci et al., Frontiers in Neuroscience, 2014.

- Restricted Boltzmann Machine
  - Originally developed for ANN
  - It is feasible to map RBM onto the "spike"-based ML because of the following RBM characteristics.
    - Binary neuron output : 1 or 0

      → Similar to spikes
    - Stochastic neuron

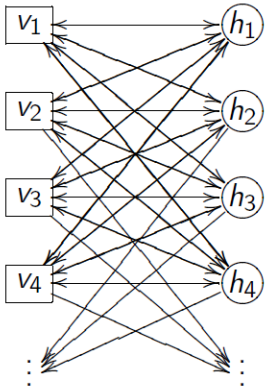
      → Mitigates resistance instability (noise, drift, temperature dependence) issues in PCM
    - Weights are updated by local neuron firing activity (STDP-like update rule)

      → No need to back-propagate the error

→ These similarities between RBM and SNN enable *efficient circuit implementation*.

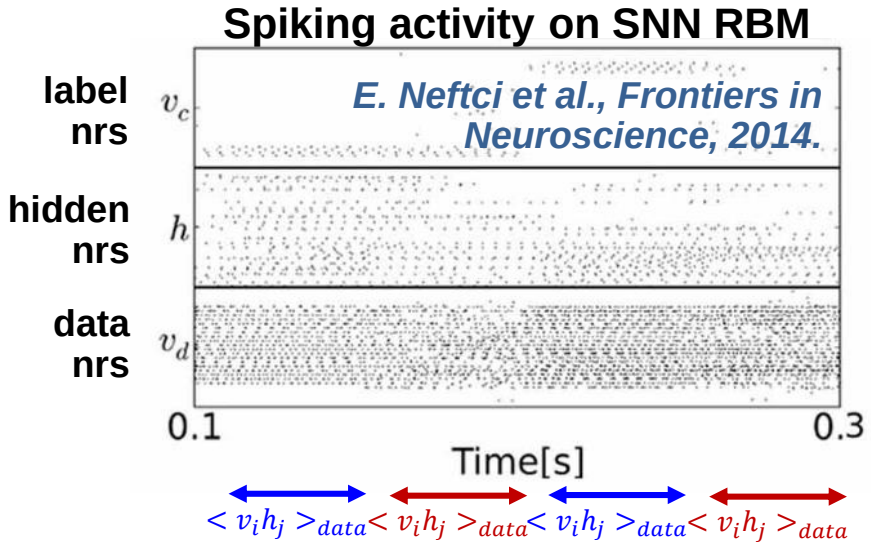
Mapping key operations in RBM to a neuromorphic core

Key operations in the algorithm (RBM)	SNN HW implementation
Forward propagation	Forward LIF
Backward propagation	Backward LIF
Positive Weight update	STDP (positive update)
Negative weight update	STDP (negative update)



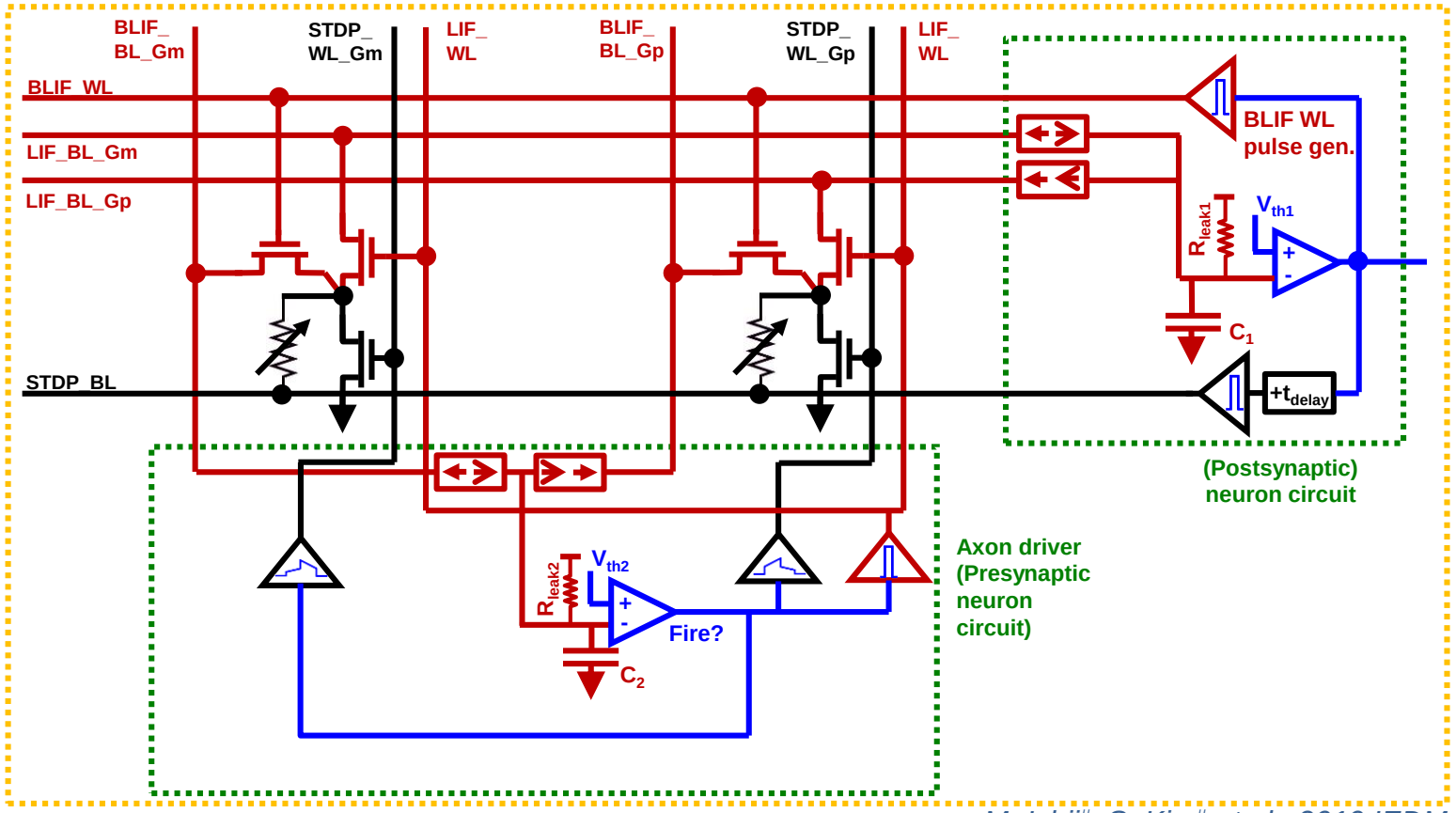
contrastive divergence (CD)

$$\Delta w_{ij} \propto < v_i h_j >_{data} - < v_i h_j >_{recon}$$

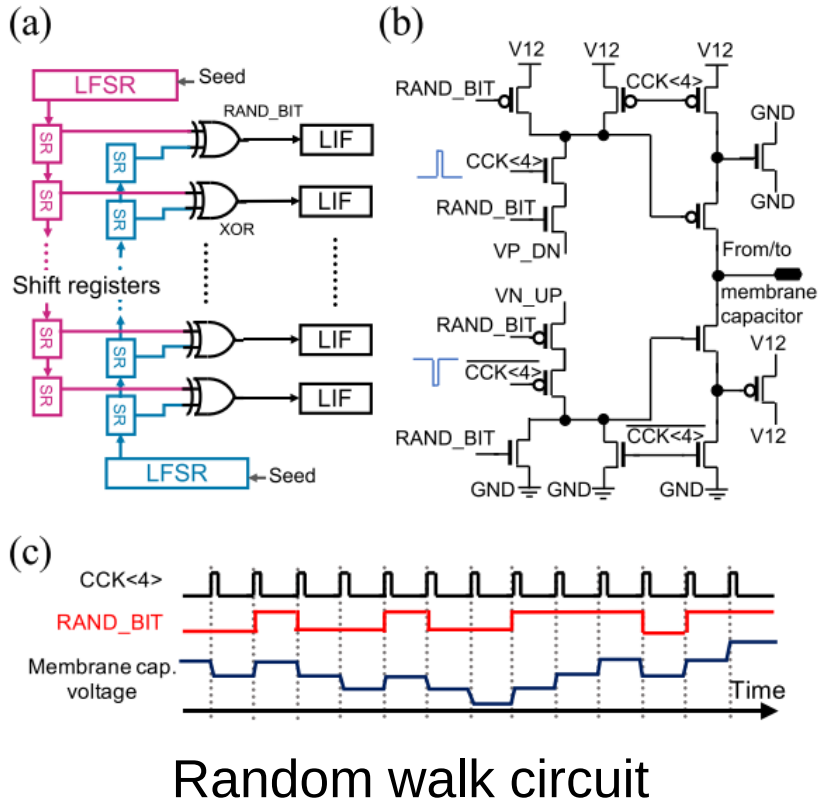


6T-2R for RBM operation with on-chip learning

Multiply / accumulate	Update	Activation



M. Ishii#, S. Kim# et al., 2019 IEDM

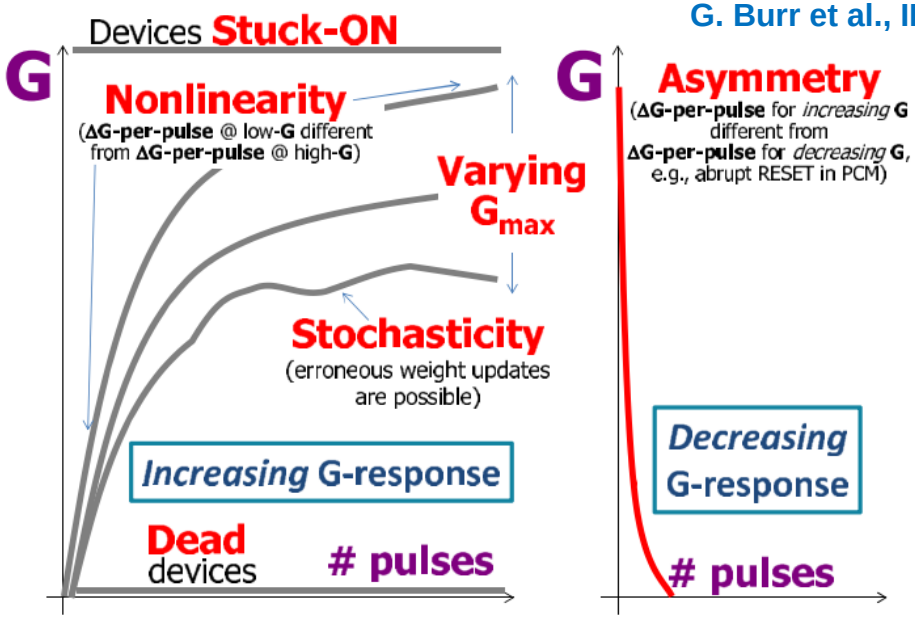


Random walk circuit

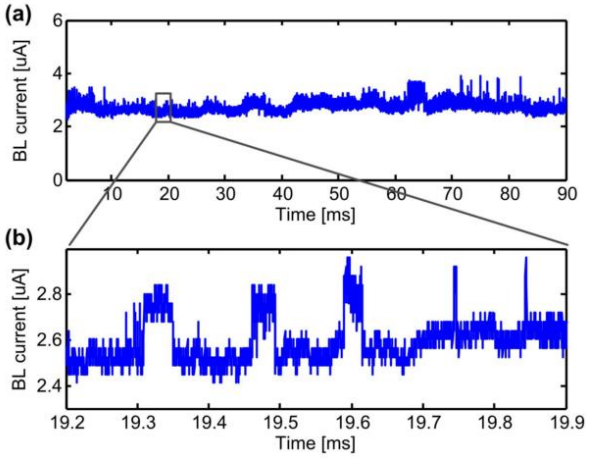
- In addition to synapses, there are many components required for neuromorphic computing.

PCM weaknesses as a synaptic device

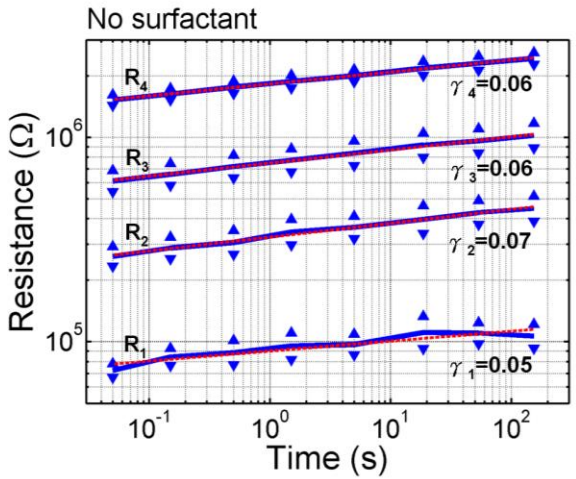
1. One-sided update
2. Non-linear update
3. 1/f noise
4. Resistance drift



G. Burr et al., IEDM 2014

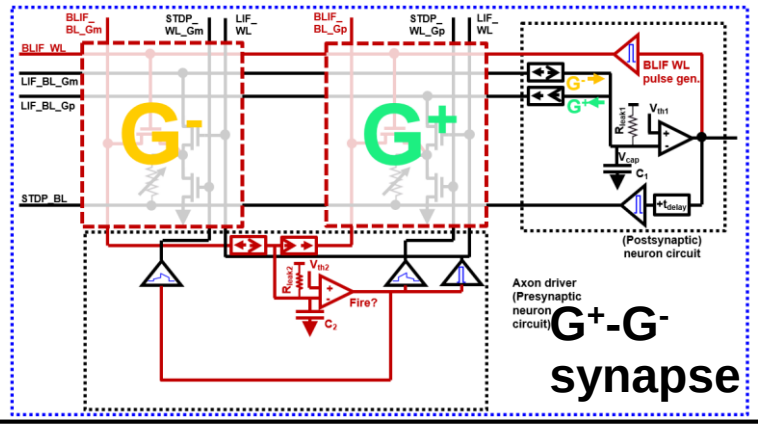


G. Close et al., IEDM 2010

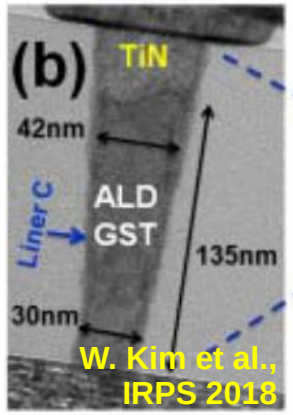


S. Kim et al., IEDM 2013

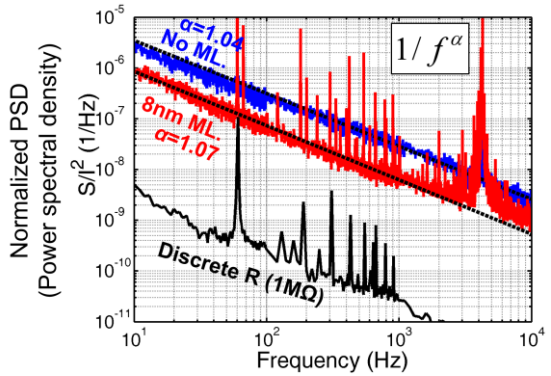
Proposed solutions



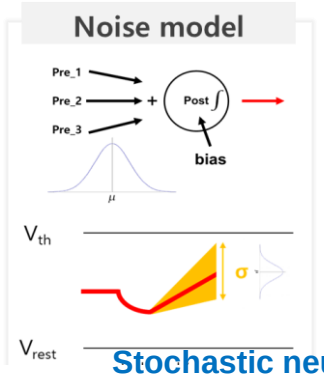
G^+-G^- synapse



W. Kim et al., IRPS 2018



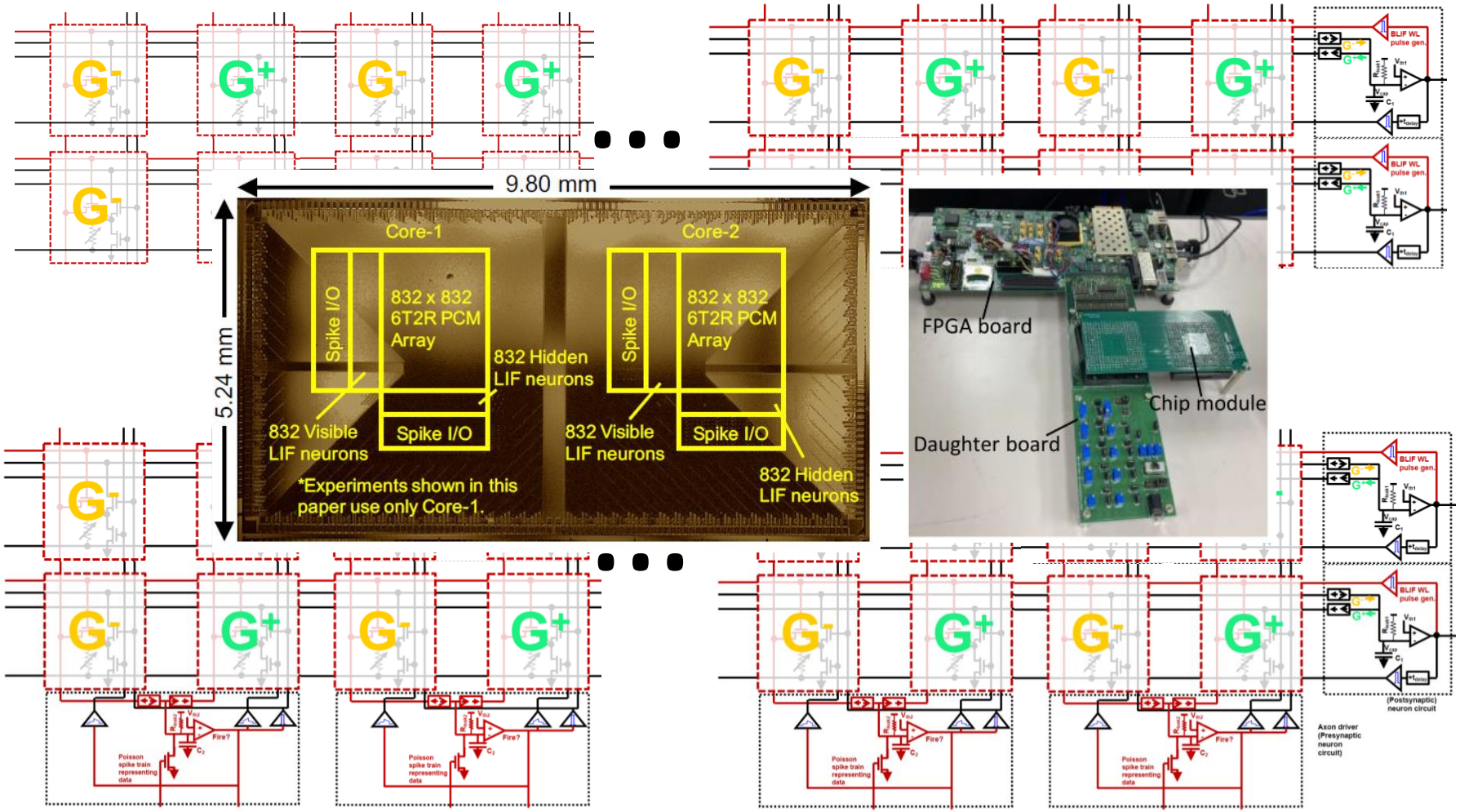
S. Kim et al., IEDM 2013

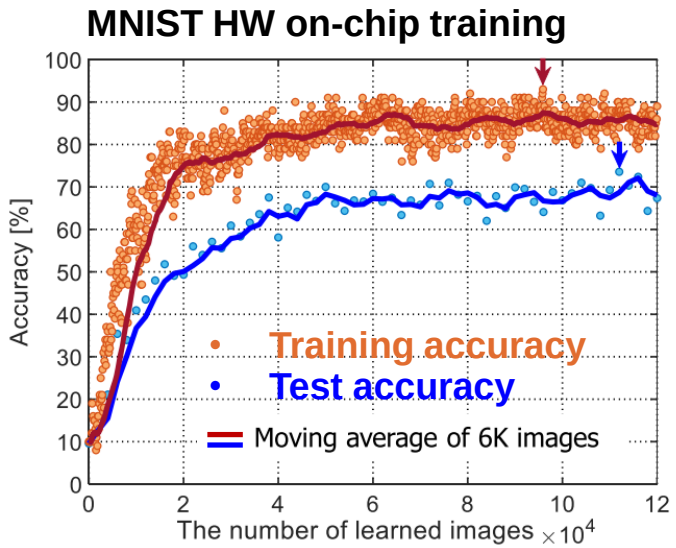


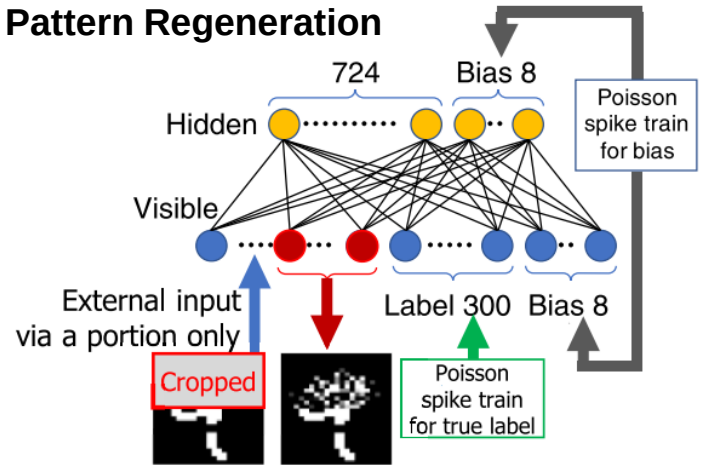
Stochastic neuron implemented by PCM noise

Neuromorphic array for MNIST demonstration

Fully Si-integrated large PCM array size (692K synaptic cells) with on-chip learning

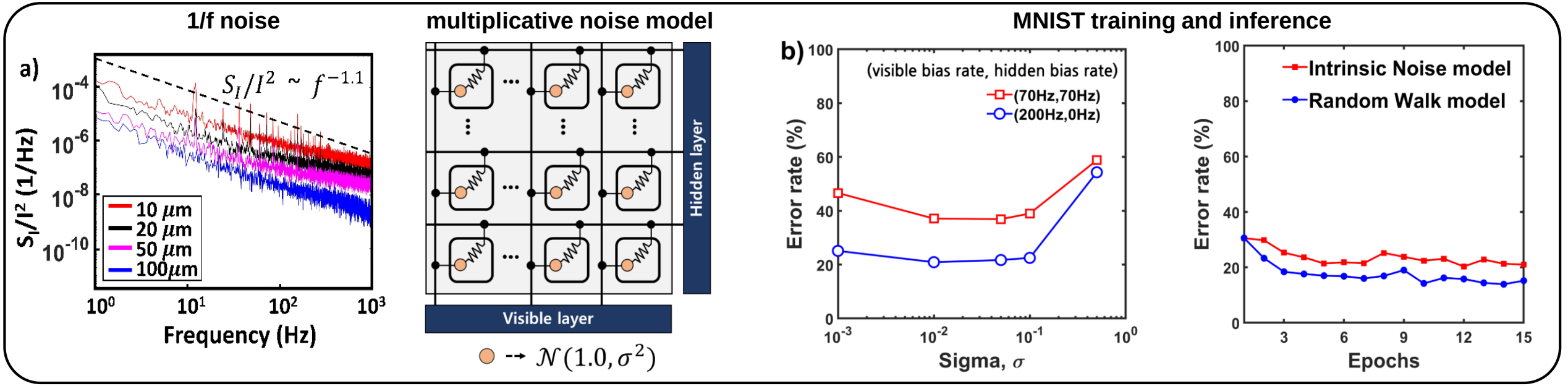






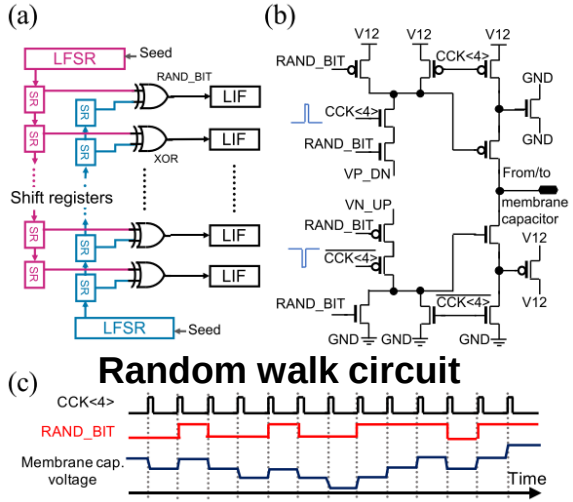
M. Ishii[#], S. Kim[#], et al., IEDM (2019)
U. Shin et al., Adv. Intell. Syst (2022)

PCRAM Noise for Neuromorphic Computing



D. Kang et al., "1/f noise in amorphous Sb_2Te_3 for energy-efficient stochastic synapses in neuromorphic computing," Semicond. Sci. Technol., v.36, p.124001, 2021

- 1/*f* noise of PCRAM devices can improve the efficiency of neuromorphic computing
  - 1/*f* noise can replace random walk circuits leading to ~60 times better energy efficiency.
  - The optimal amount of noise (standard deviation) is ~1-5 %, which is readily available with nanoscale PCRAM cells.



M. Ishii#, S. Kim# et al., 2019 IEDM

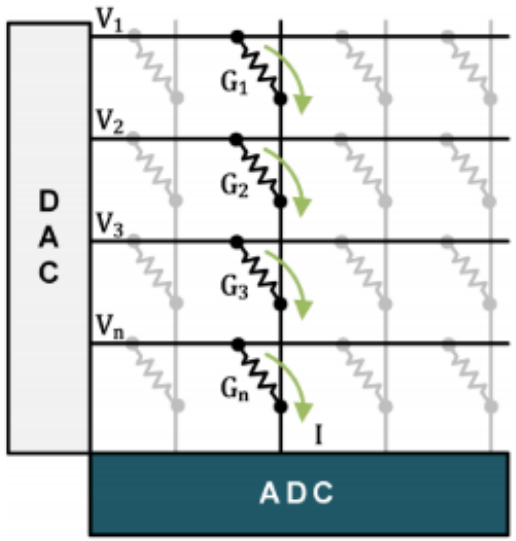
Importance of Co-optimization

Neuromorphic processor development

- **Material**
- **Device**
- **Circuit**
- **Architecture**
- **Algorithm**
- **Application**

Training and inference algorithms for end-to-end analog neural network?

- **Can we build ANN without power-hungry ADC/DAC?**

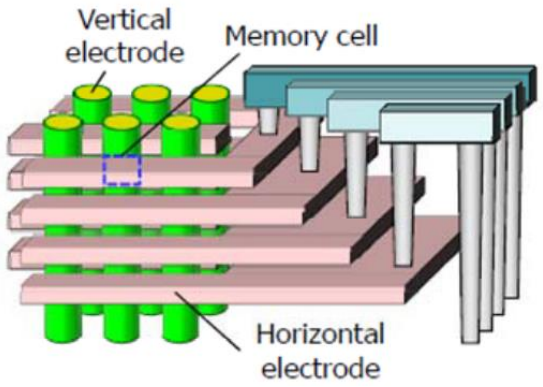


Y. Kim et al, “Neural Network-Hardware Co-design for Scalable RRAM-based BNN Accelerators”.

❑ *Multidisciplinary research* ranging from materials to machine learning applications is required!

“Massively Parallel” AiMC using 3D PRAM/RRAM

3D VPRAM/RRAM

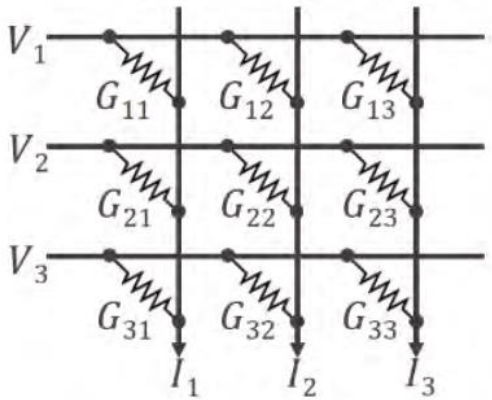


S.-G. Park (Samsung) et al, 2012 IEDM.

- Low cost per bit
- Potential solution for high-density storage class memory

+

Analog in-Memory Computing (AiMC)



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

- Massively parallel operation: O(1)

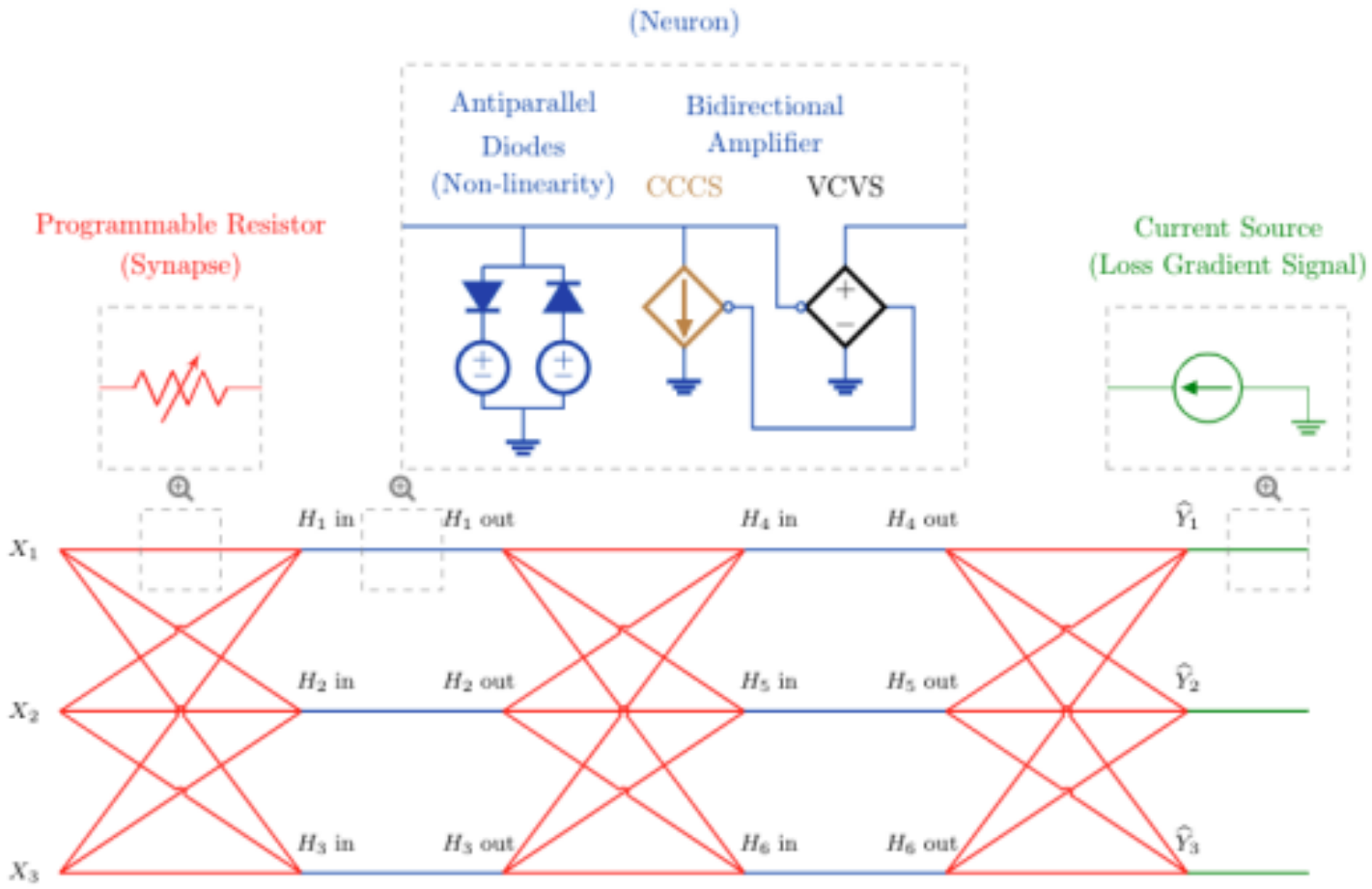
=

AiMC using 3D PRAM/RRAM

?

- Many challenges
  - → Complicated neuron circuit (e.g. ADC/DAC)
  - → Unable to run massively parallel operation

End-to-End Analog Neural Network Algorithm

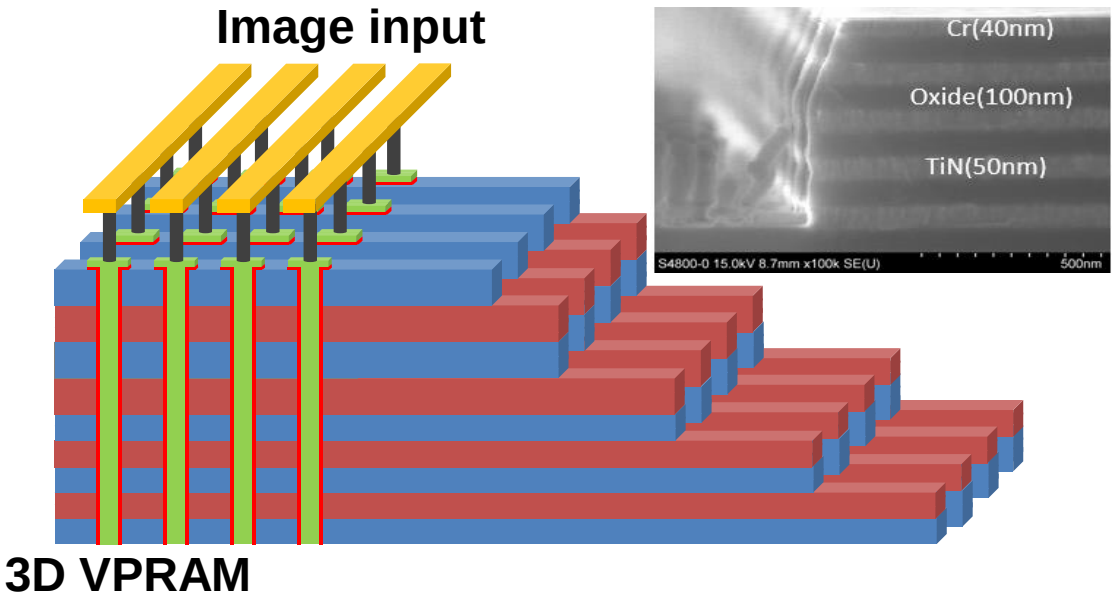


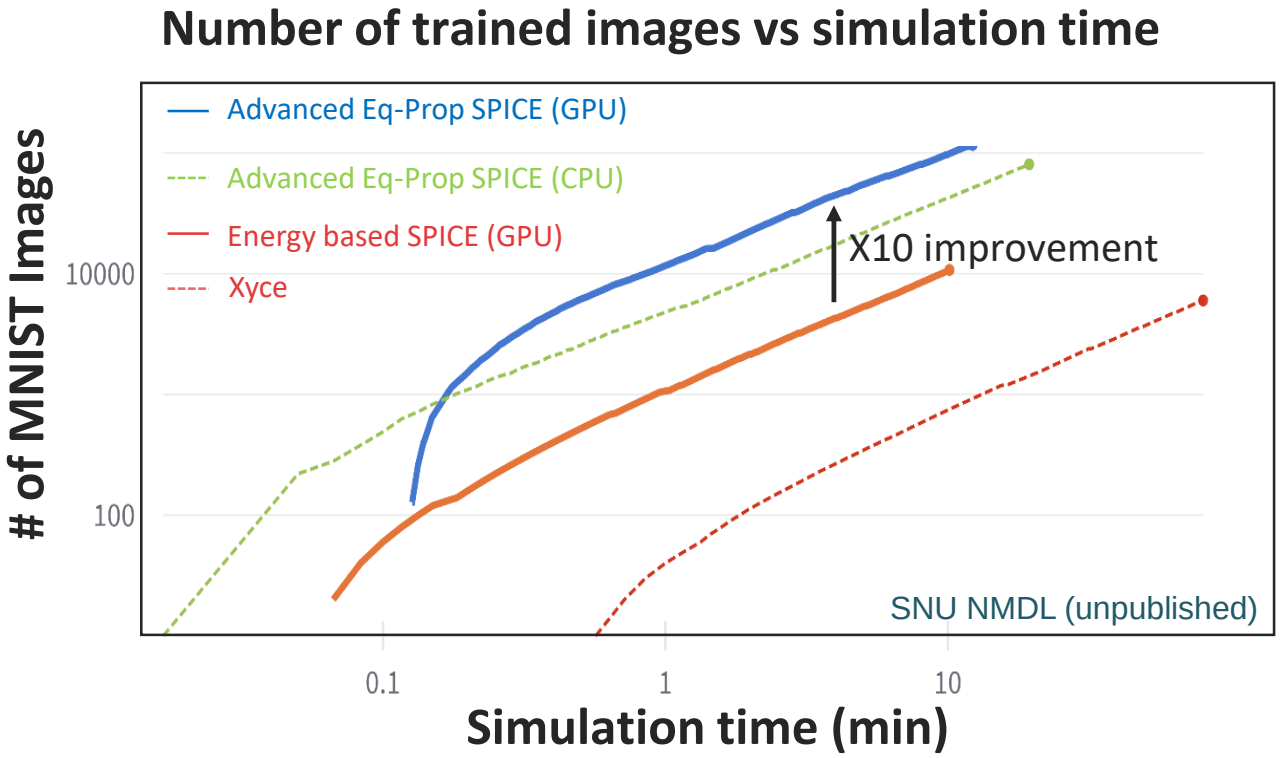
J.Kendall et al, ArXiv 2006.01981

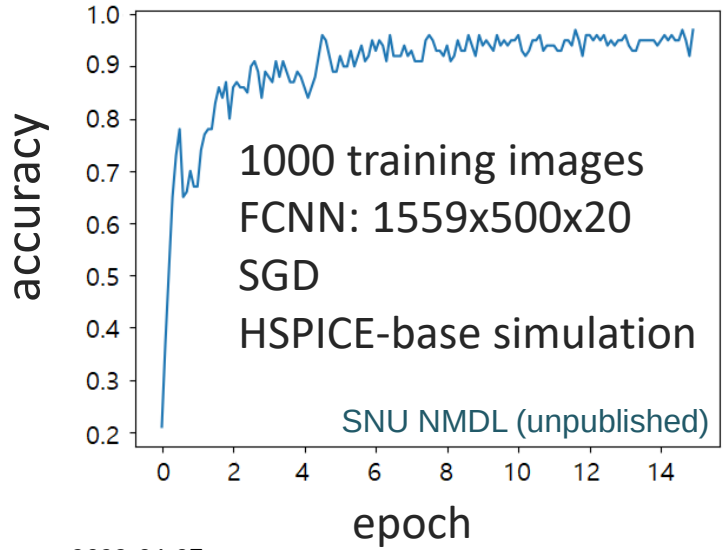
- Key components
  - (3D) crossbar array
  - Antiparallel diode
  - Bidirectional amplifier (Si CMOS)
  - Current source (Si CMOS)
- Major benefits
  - End-to-end analog: No ADC/DAC
  - Weight update based on locally available information

- Equilibrium Propagation has been recently proposed to enable end-to-end analog neural network.

Fabricated 3D VPRAM and Simulation Results

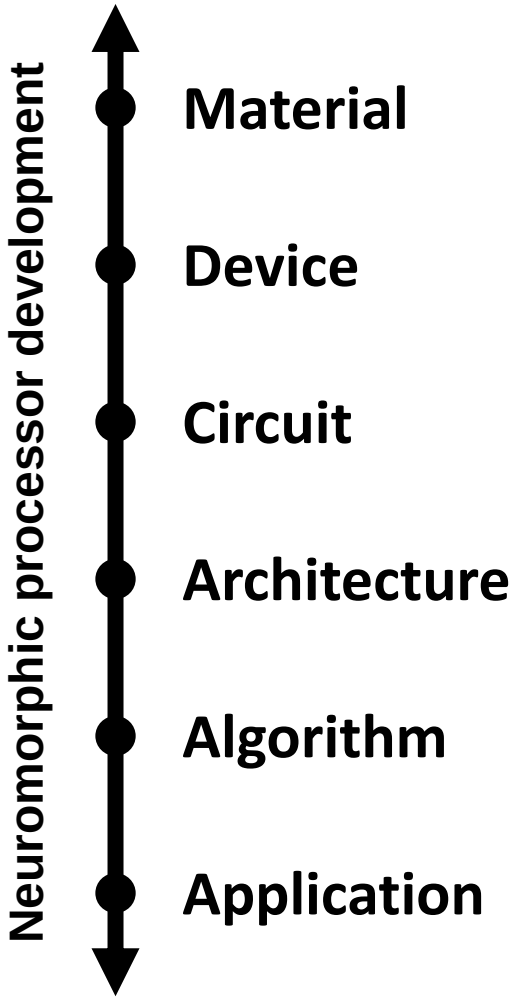






- Using simulations, 3D PRAM/RRAM array shows good inference accuracy using end-to-end analog scheme with O(1) running time.

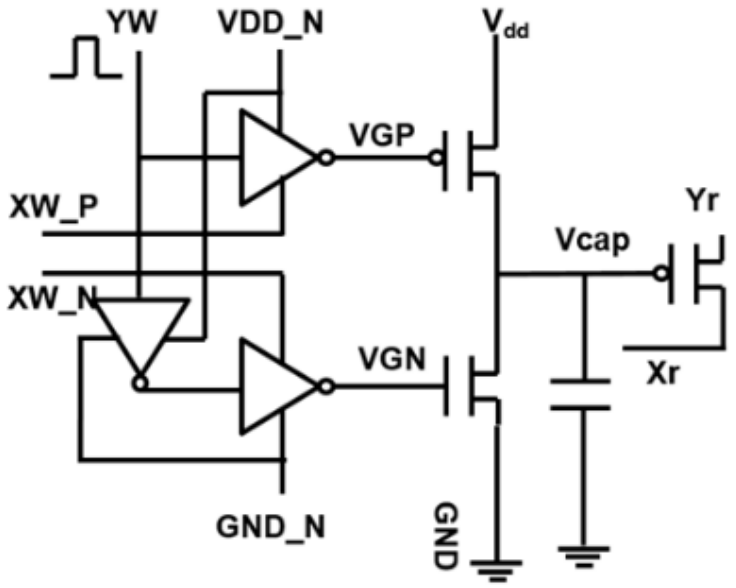
Importance of Co-optimization

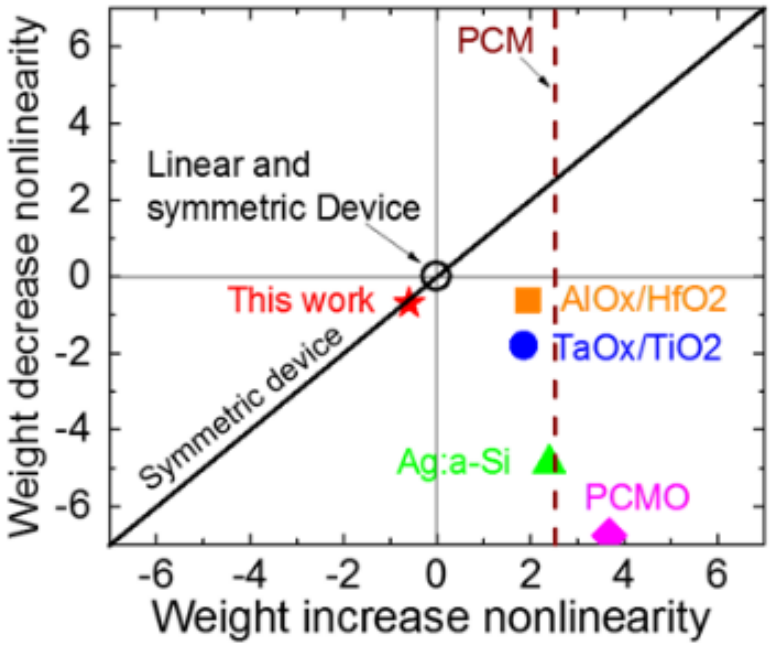


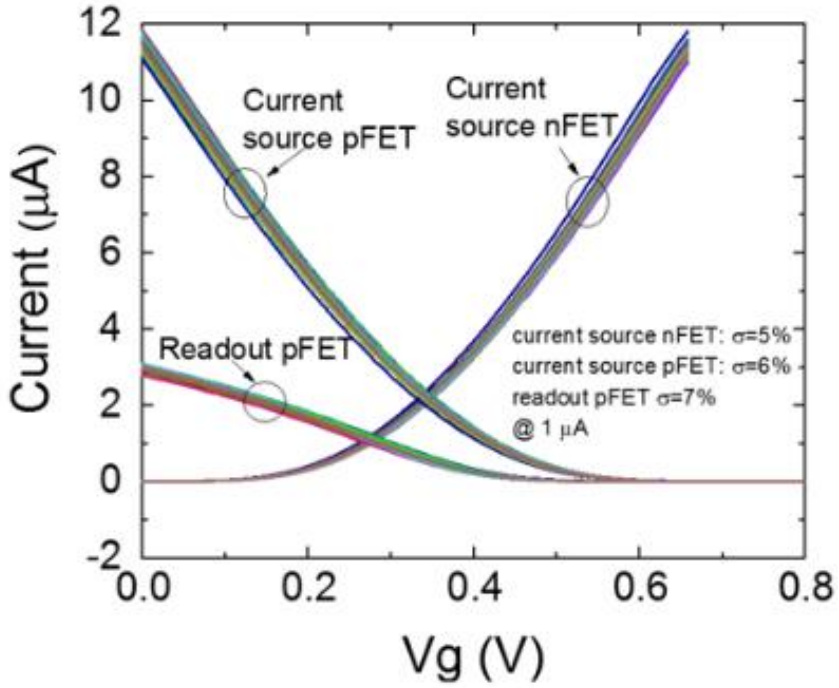
Can we efficiently compensate the device imperfections in the weight update linearity and symmetry?

- *Multidisciplinary research* ranging from materials to machine learning applications is required!

Capacitor-Based Synaptic Device



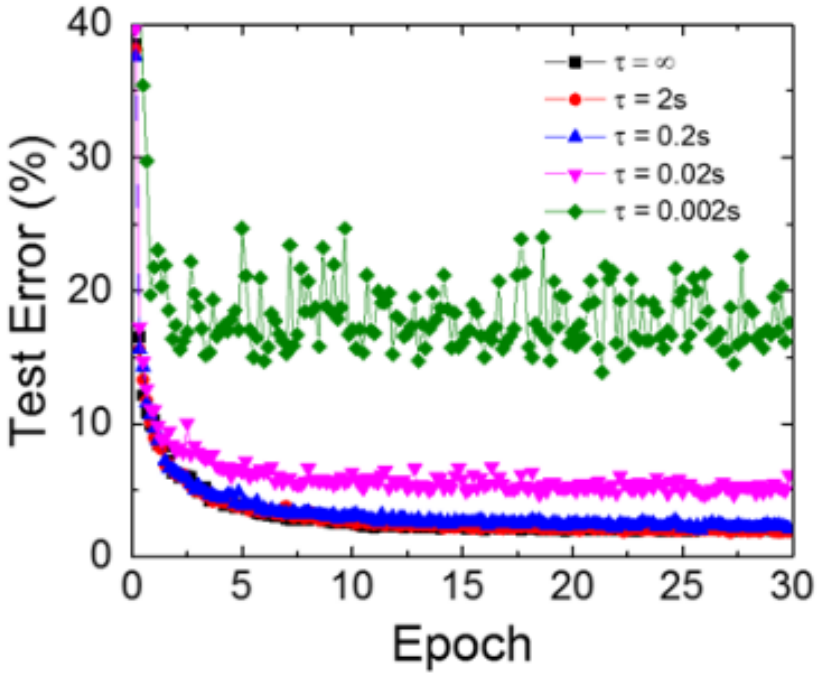


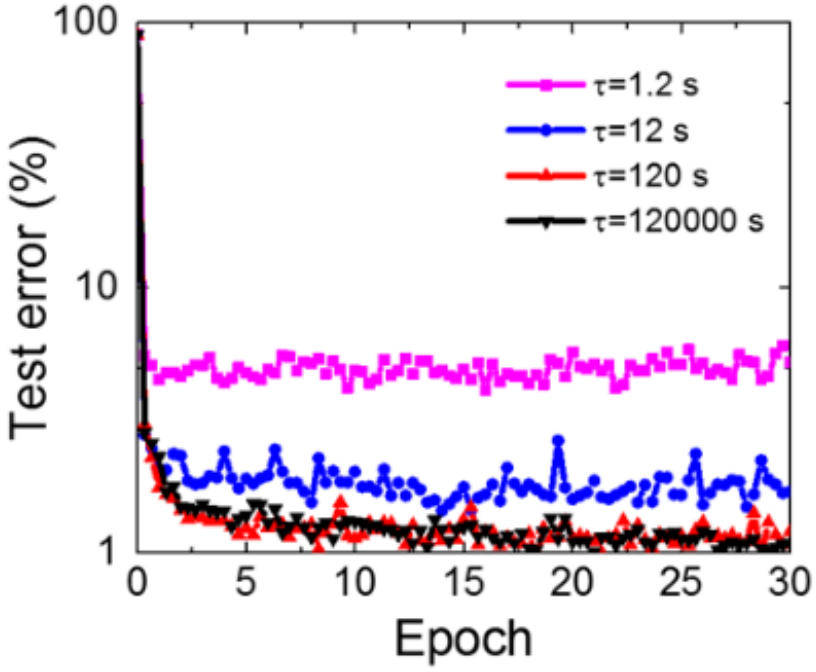


Y. Li et al., VLSI Tech. Symp. (2018)

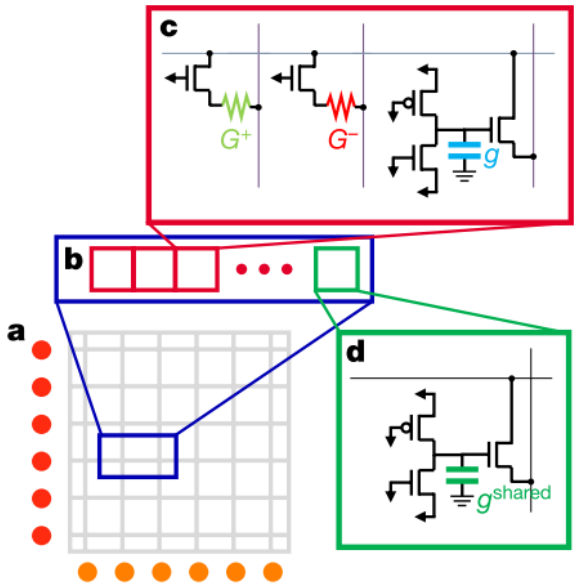
- IBM has demonstrated capacitor-based synaptic devices in 2018.
  - Near ideal symmetry and linearity.
  - Minimal device-to-device variation
- However, how about retention?

Capacitor-Based Synaptic Device





Y. Li et al., VLSI Tech. Symp. (2018)



S. Ambrogio et al., Nature (2018)

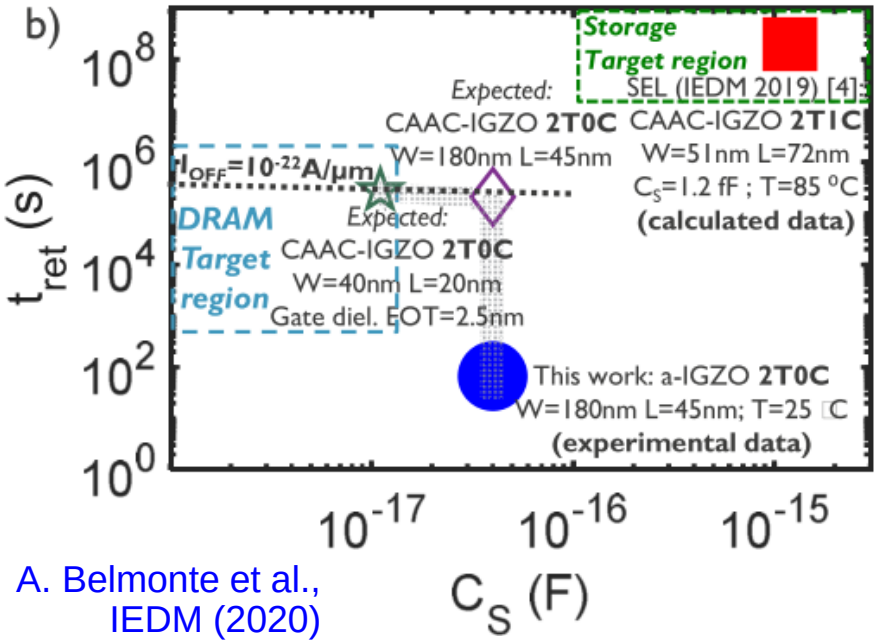
- Infinite retention is not required due to on-chip training.
  - However, deeper NN with larger dataset requires longer retention
- Proposed solution: Online/offline Weight transfer
  - But, how to transfer weights efficiently?

Capacitor-Based Synaptic Devices with IGZO TFT

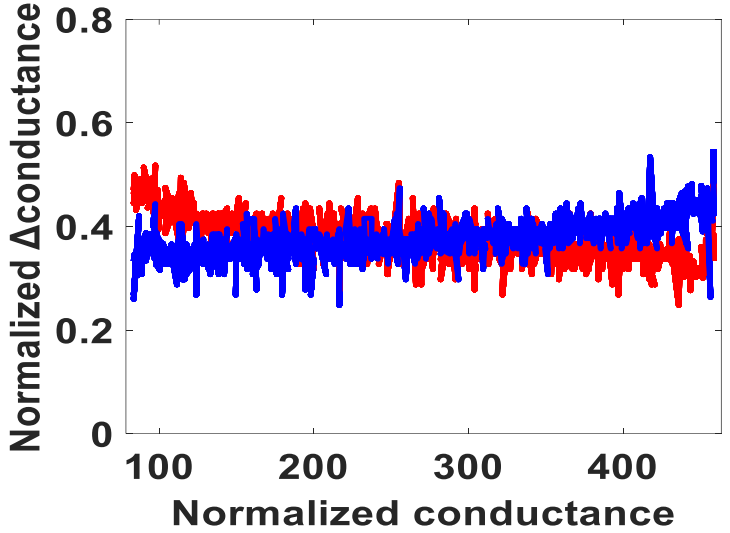
a)

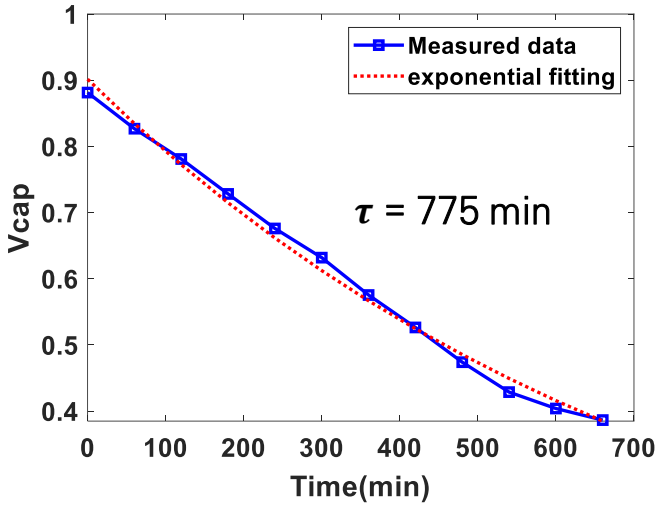
$W_{Wtr,Rtr}$	$L_{Wtr,Rtr}$	I_{OFF}	Median t_{ret}	$C_{ox,Rtr}$
180nm	45nm	~3x10^{-19}A/μm	~60s	38 aF

b)



A. Belmonte et al., IEDM (2020)





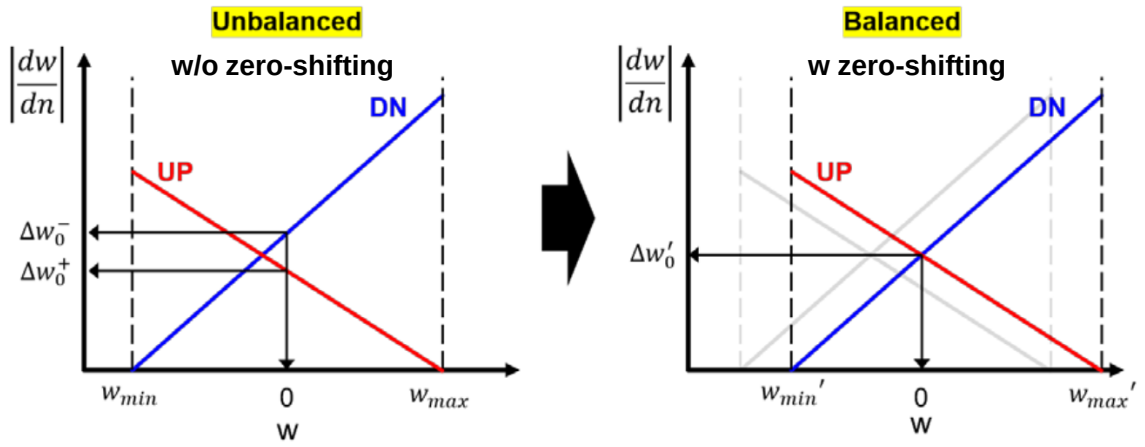
Capacitor-Based Synaptic Devices with IGZO TFT (Unpublished)

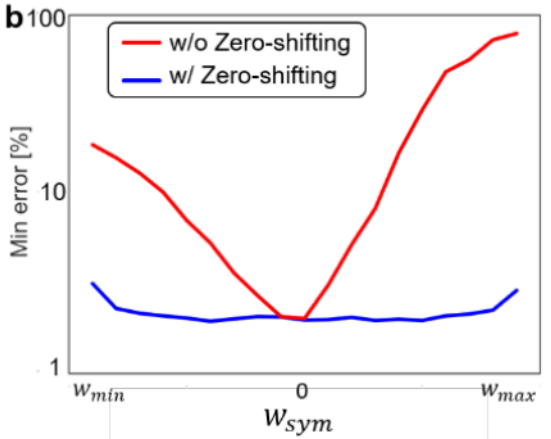
- Using IGZO TFT as an access device, capacitor-based synaptic device can demonstrate
  - Not only good linearity and symmetry
  - But also long retention time ($\tau = 775$ min)

Zero-shifting / Tiki-Taka algorithms

Zero-shifting Technique

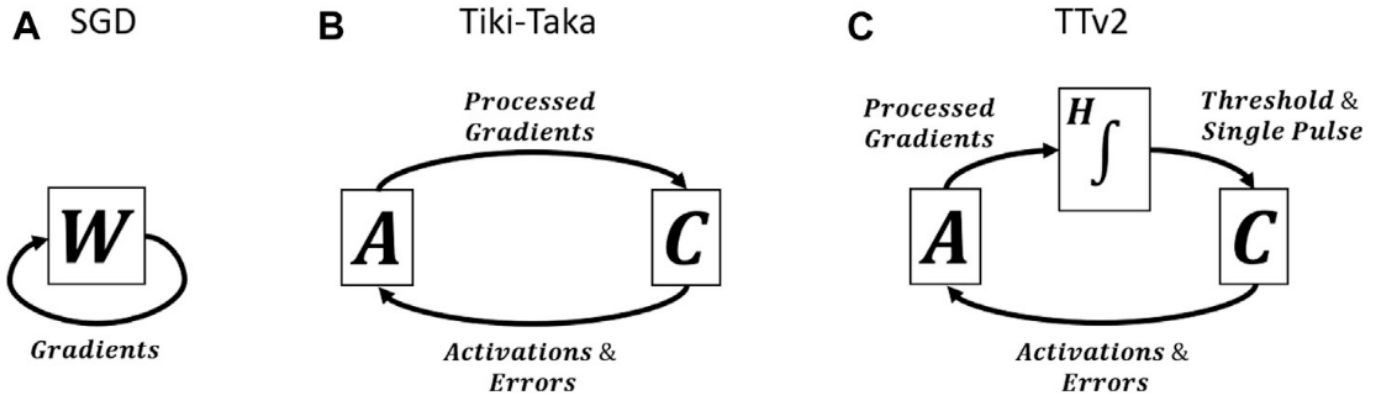
H. Kim *et al.*, "Zero-shifting Technique for Deep Neural Network Training on Resistive Cross-point Arrays," 2019, http://arxiv.org/abs/1907.10228

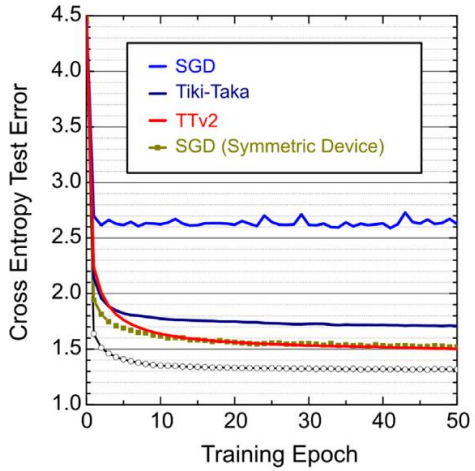




Tiki-Taka Algorithm

T. Gokmen, "Enabling Training of Neural Networks on Noisy Hardware," 2021 doi: 10.3389/frai.2021.699148.





- These training algorithms can compensate for the insufficient linearity and symmetry of synaptic devices
  - For more details, refer to references on this page.

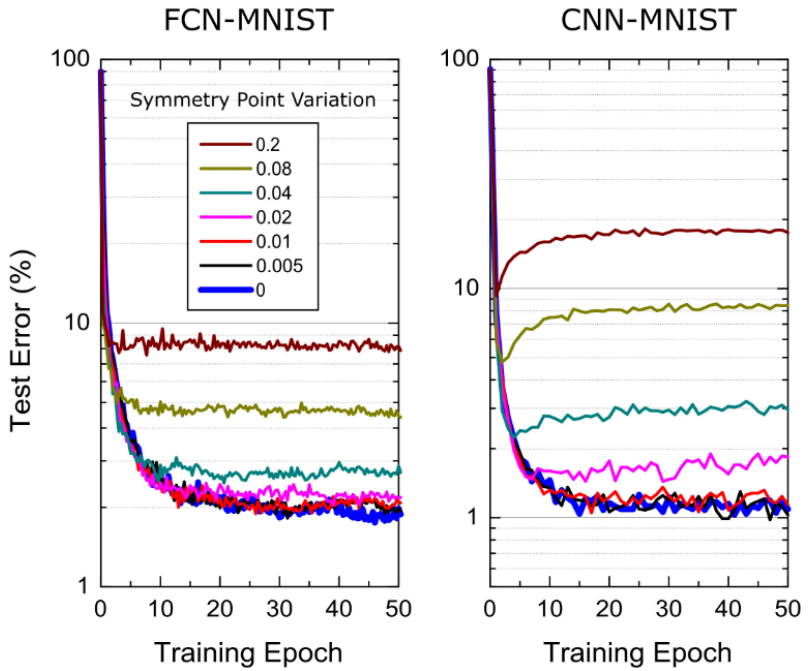
Modified Tiki-Taka Algorithm

Algorithm 1: Tiki-Taka

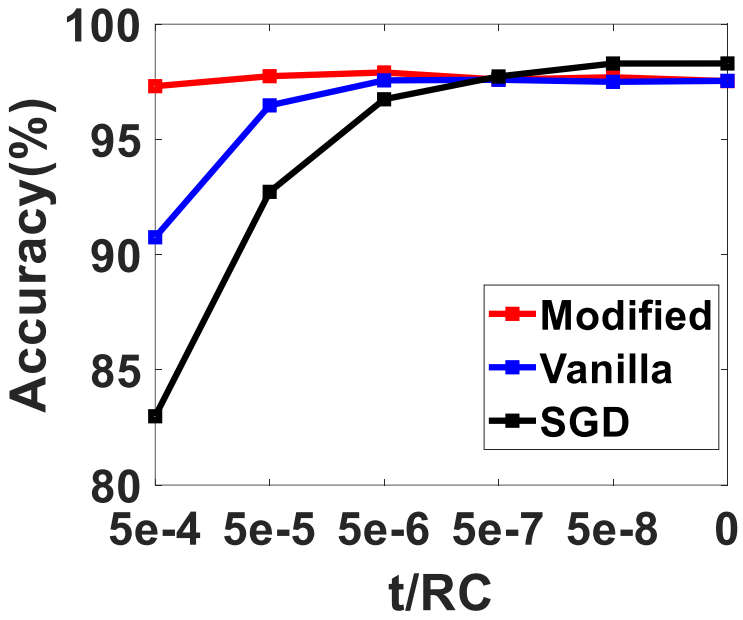
```
initialize A = 0 after symmetry point shifting
initialize C to random values
k = 0
t = 0
nr = Number of rows in A/C
ns = Hyperparameter – Number of update cycles

For each data in the training dataset
    y = (γA + C)x
    z = (γA + C)^T δ
    a_ij ← a_ij + η_a[δ_i × x_j]F_ij(a_ij) − η_a|[δ_i × x_j]|G_ij(a_ij)
    k = mod(k+1, ns)
    If (k = 0)
        u = prepare_vector(t)
        v = Au
        c_ij ← c_ij + η_c[f(v_i) × u_j]F_ij(c_ij) − η_c|[f(v_i) × u_j]|G_ij(c_ij)
        t = mod(t+1, nr)
    end
end
```

T. Gokmen, Frontiers in Artificial Intelligence (2021)



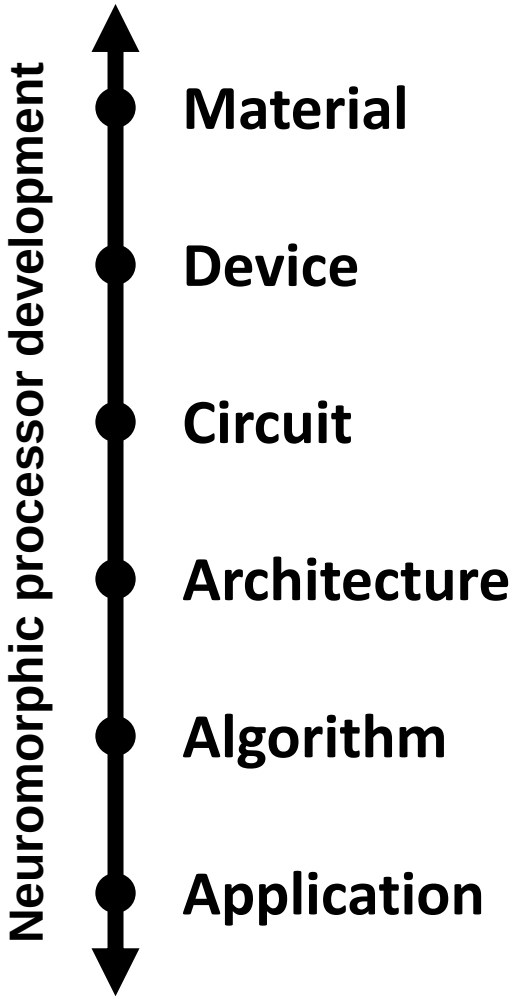
T. Gokmen, Frontiers in Neuroscience (2021)



Capacitor-Based Synaptic Devices with IGZO TFT (Unpublished)

- Tiki-Taka algorithm (v1 and v2)
  - Effectively, mitigates the non-ideal properties of synaptic devices
  - Yet, symmetry point needs to be carefully calibrated and stored in a separate array.
- Modified Tiki-Taka algorithm (optimized for capacitor-based synaptic device)
  - By replacing the symmetry point with resting point, (1) improves the resilience against the retention failures (2) The resting point is stable and can be easily read

Summary

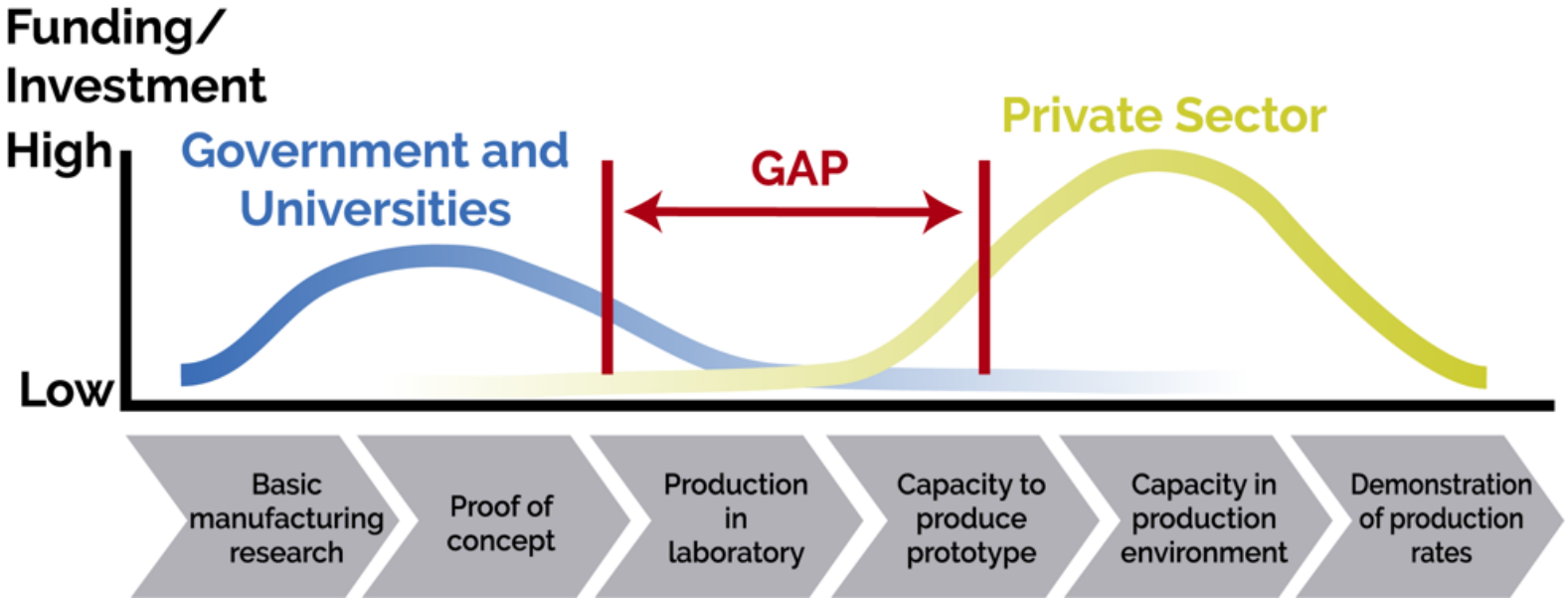


- Stochastic algorithms such as Restricted Boltzmann Machine can not only mitigate the read noise in synaptic devices but also utilize the read noise to efficiently implement stochastic behavior.
- O(1) operation ANN based on 3D PRAM/RRAM can be achieved with equilibrium propagation algorithm which enables end-to-end analog NN.
- Modified Tiki-Taka algorithm along with co-optimized synaptic device can efficiently mitigate the linearity and symmetry of synaptic devices.

ROK-USA Collaboration on Analog Deep Learning Accelerator

Neuromorphic processor development

- **Material**
- **Device**
- **Circuit**
- **Architecture**
- **Algorithm**
- **Application**



https://www.nist.gov/image/niimbl-valley-death

- Lab-to-Fab problem prevents researchers from tacking practical issues.
- Facilities for prototyping analog deep learning accelerator
  - is expensive to build and maintain
  - because diverse technology development is needed (PRAM, RRAM, Ferroelectric memory, IGZO, 2D material, advanced packaging and etc.)
- Can ROK and USA share the burden of building and maintaining the prototyping facilities?

Acknowledgement

SAMSUNG ELECTRONICS
SAMSUNG ADVANCED INSTITUTE OF TECHNOLOGY
SK hynix
LG Display
NRF National Research Foundation of Korea
산업통상자원부 Ministry of Trade, Industry and Energy
차세대지능형 반도체사업단 Next-generation Intelligence Semiconductor Foundation NIS2030
IBM