

# Memory-Based Neuromorphic Hardware for Advanced Neural Network Models

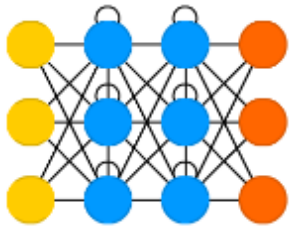
Dmitri Strukov  
UC Santa Barbara  
**SEMICON, Korea**  
February 2021

# Talk Outline

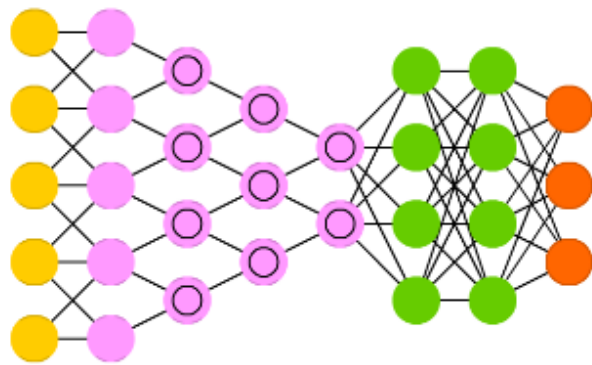
- Motivation and background
- Probabilistic neural networks
  - Stochastic dot product
  - Restricted Boltzmann machine
  - Neurooptimization
- Spiking neural networks
- Summary

# Artificial Neural Network Zoo

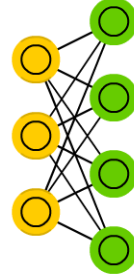
Recurrent Network (RNN)



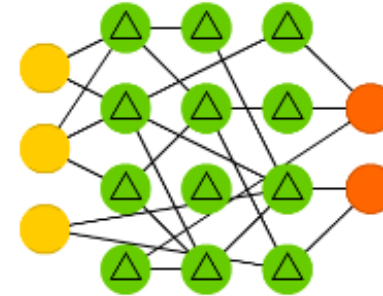
Deep Convolutional Network (CNN)



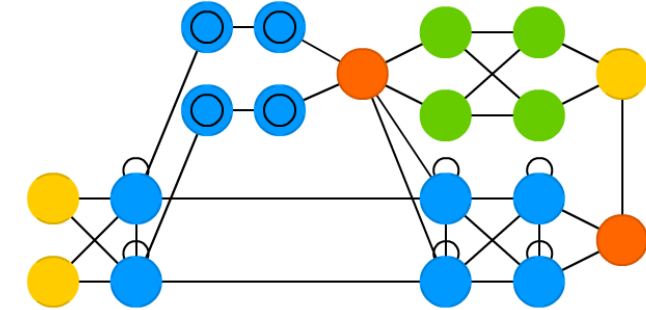
Restricted BM (RBM)



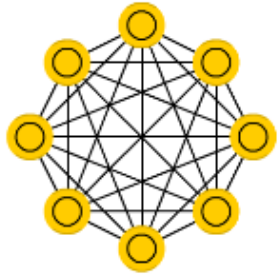
Liquid State Machine (LSM)



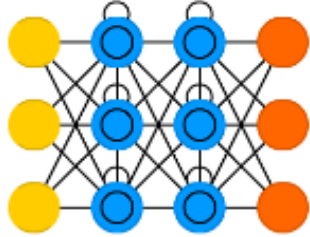
Attention Network (AN)



Hopfield Network (HN)



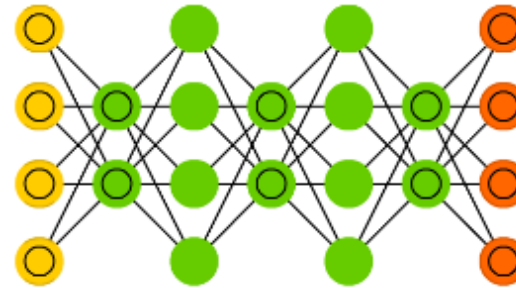
Long / Short Term Memory (LSTM)



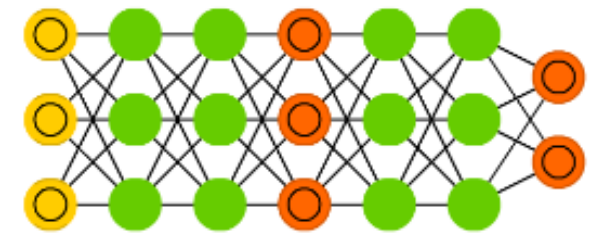
Boltzmann Machine (BM)



Deep Belief Network (DBN)



Generative Adversarial Network (GAN)



Multilayer Perceptron (MLP)

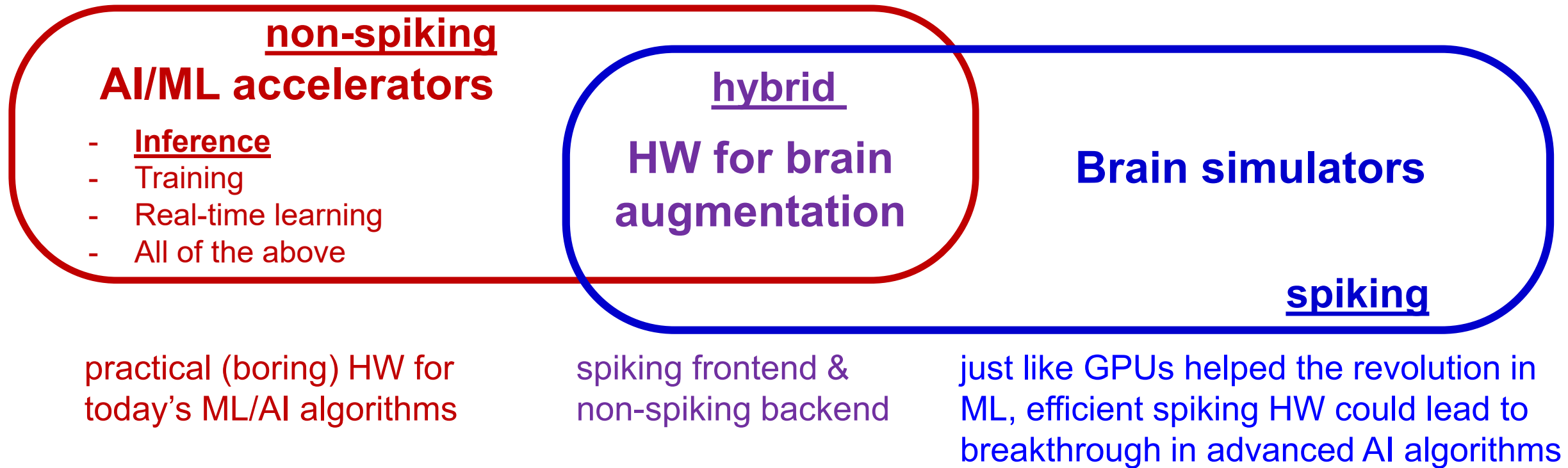


Backfed Input Cell	Noisy Input Cell	Probabilistic Hidden Cell	Output Cell	Recurrent Cell	Different Memory Cell	Convolution or Pool
Input Cell	Hidden Cell	Spiking Hidden Cell	Match Input Output Cell	Memory Cell	Kernel	

**Vector-by-matrix multiplication** (dot-products with the same input vector) is the **most common operation**

$$x_i = f\left(\sum_{j=1}^N w_{ij} y_j\right)$$

# Simplified Big Picture of Neuromorphic Hardware

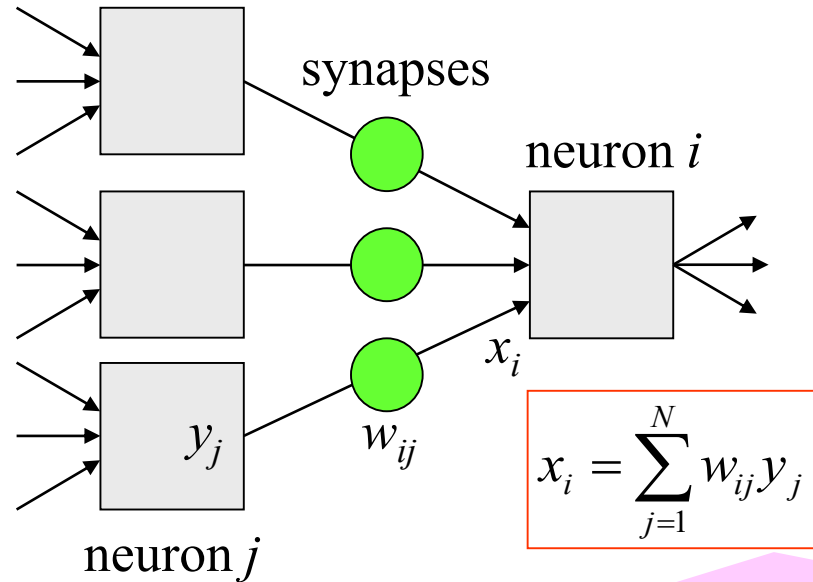


- Mimicking all brain features → ultimate AI hardware?... but biology is limited to biomaterials
- Vector-by-matrix multiplication is the most common operation for virtually all approaches

# Radical Improvement with Analog Computing

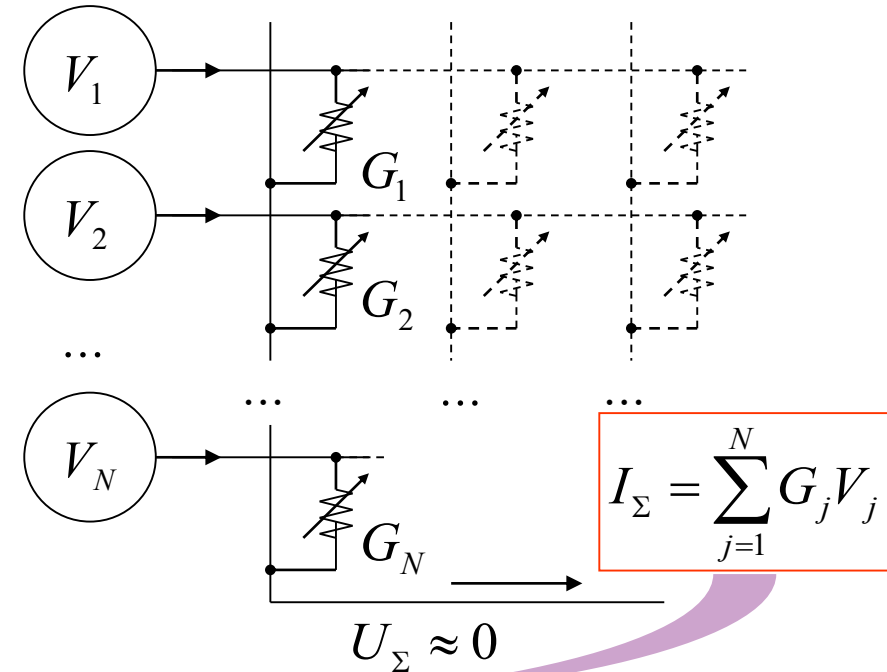
## Vector-by-Matrix-Multiplication (VMM):

basic neuromorphic operation...



## Analog VMM:

...using the Ohm & Kirchhoff laws



## Features:

- physical-level (very compact) and in-memory computation → fast and very energy-efficient
- proposed by Widrow in 1960s, popularized by Mead and his students (CalTech) in the 1980s
- no dense adjustable-conductance crosspoint devices - until recently

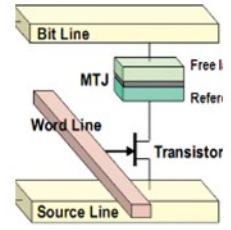
# Tunable Non-Volatile Memory Device Options

would allow to fit extra-large models on chip!

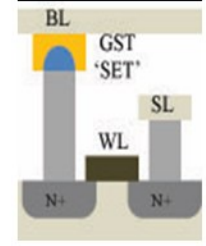


Maturity ↑

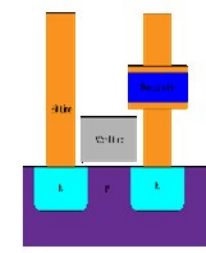
**STTRAM**



**PCRAM**



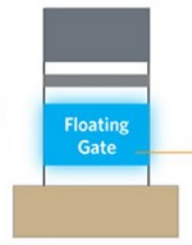
**2D FeRAM**



Active "1T1R"

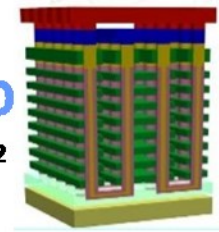
40 F<sup>2</sup> now, <20 F<sup>2</sup> with FinFET

**2D NOR**



**3DNAND**

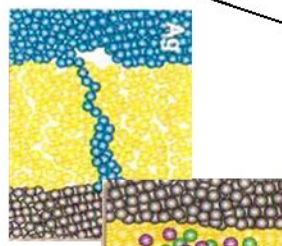
1-10 TB/in<sup>2</sup>



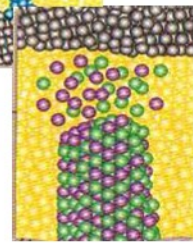
Active "1T"

few 100's F<sup>2</sup> for current, potentially down to 25 F<sup>2</sup>

**CBRAM**



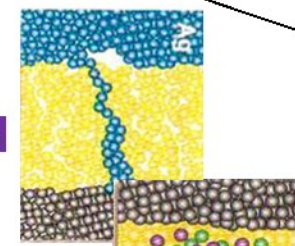
**ReRAM**



Passive "0T1R"

**CBRAM**

~4F<sup>2</sup>



**2D RRAM**

<< 4F<sup>2</sup>

**3D FeRAM?**



**3D RRAM**

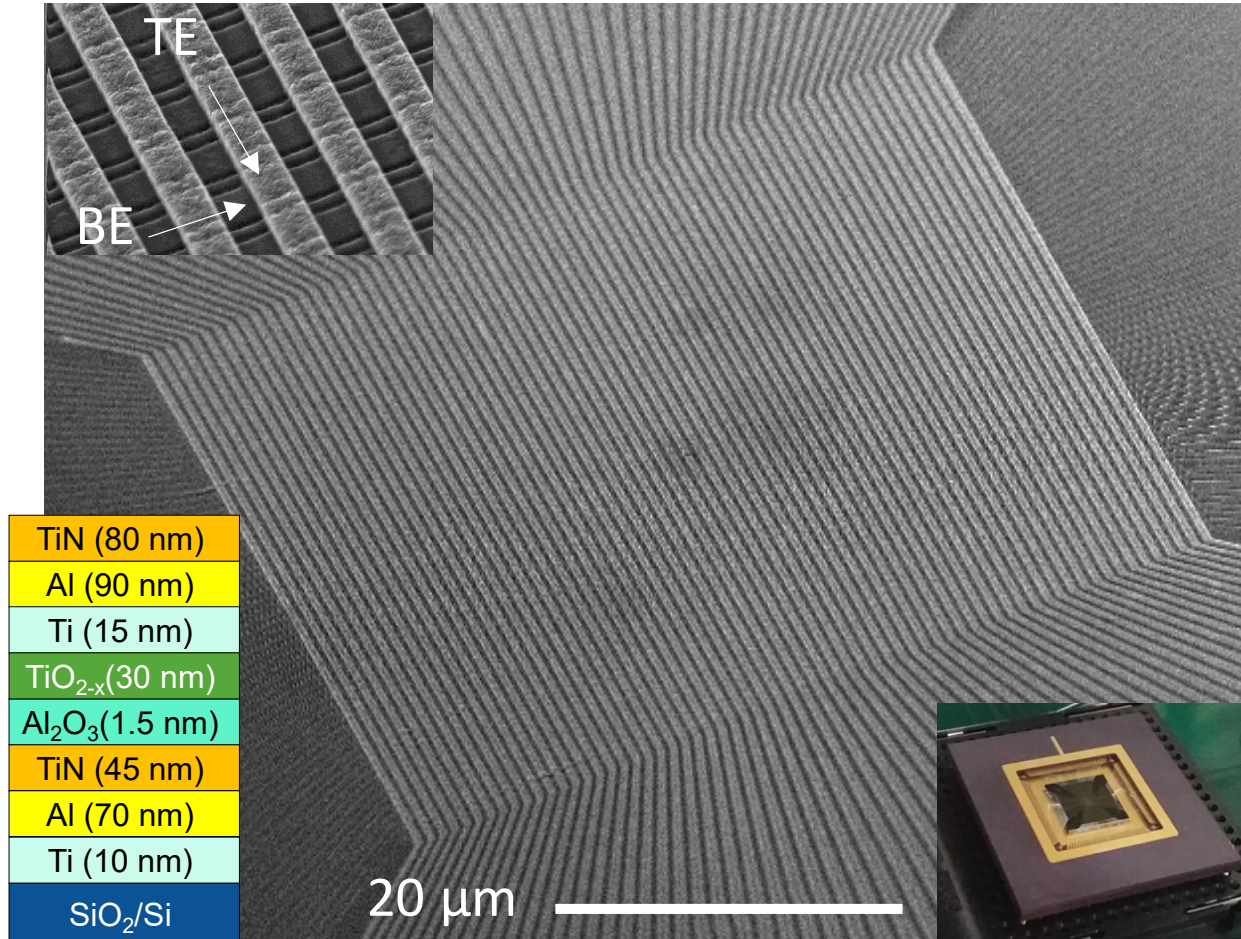
F = feature size

Most important specs: Density, retention, analog switching!

Cell density →

# Long-Term Option: (3D) Passive Metal-Oxide Memristors

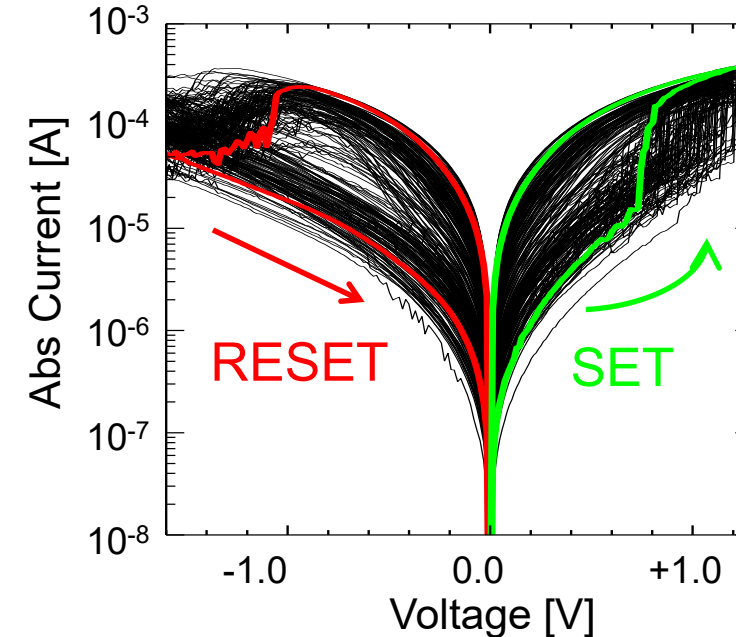
- 64 × 64 passive crossbar circuit



H. Kim et al. arXiv 2019

Background work: M. Prezioso et al., Nature 521, 61 2015, M. Prezioso et al. IEDM'15 p. 17.4.1, 2015, F. Merrih Bayat et al. Nature Comm., 2018

- Typical I-V characteristics

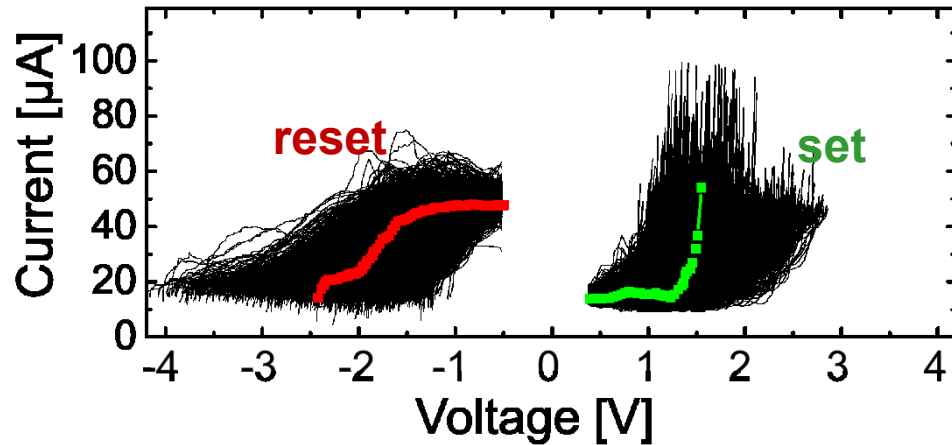


## Details:

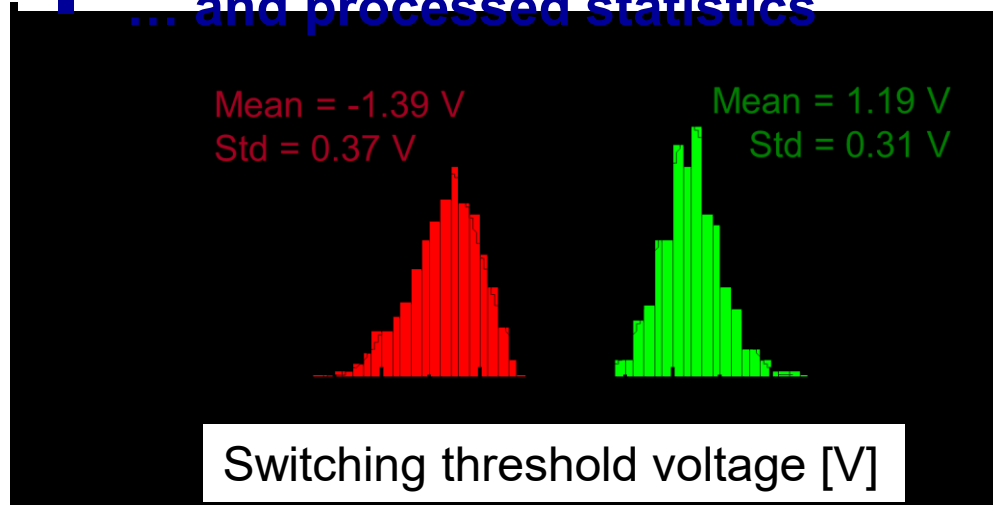
- Al<sub>2</sub>O<sub>3</sub>/TiO<sub>2-x</sub> active bilayer by reactive sputtering
- CMOS-compatible CMP/dry etching process and TiN/Al electrodes for higher conductance
- ~250 nm wide lines, passive (0T1R) integration (e.g. >250x/10,000x better memristor / memory cell density compared to 1T1R work at comparable complexity and yield)
- The largest functional analog-grade passive memristor crossbar circuit supported by proper statistics

# 64×64 Crossbar Yield and I-V Variations

## Raw data (voltage ramp) ...



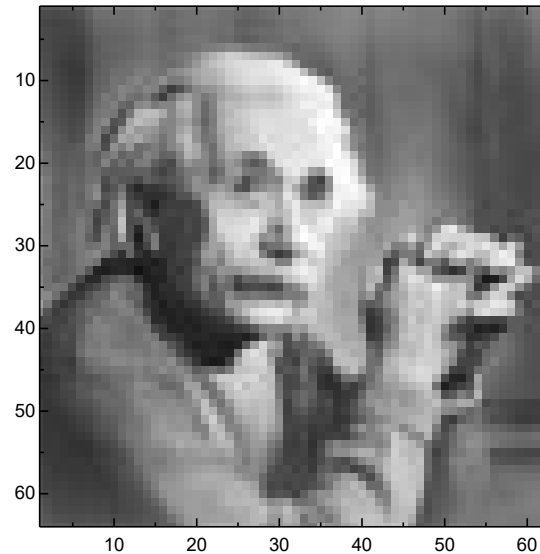
## ... and processed statistics



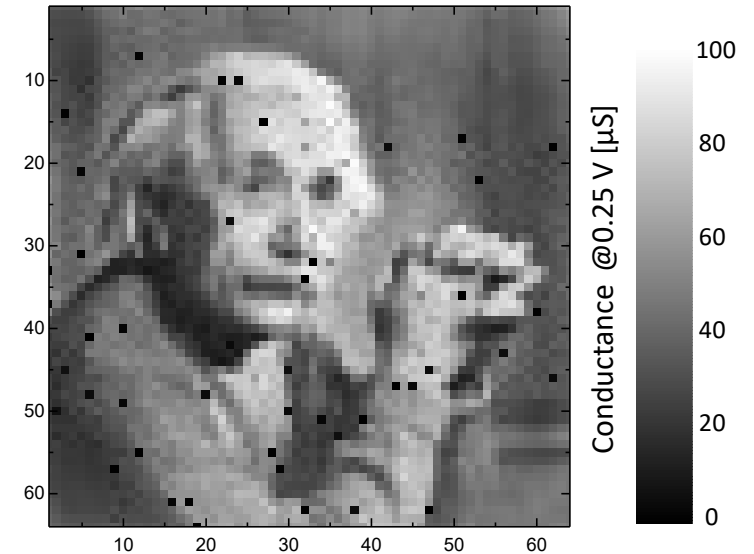
(defined as voltage at which current changes by  $> 10\%$  when applying voltage ramp)

## Conductance tuning in 4k memristor crossbar

desired pattern and ...



...actual one after tuning



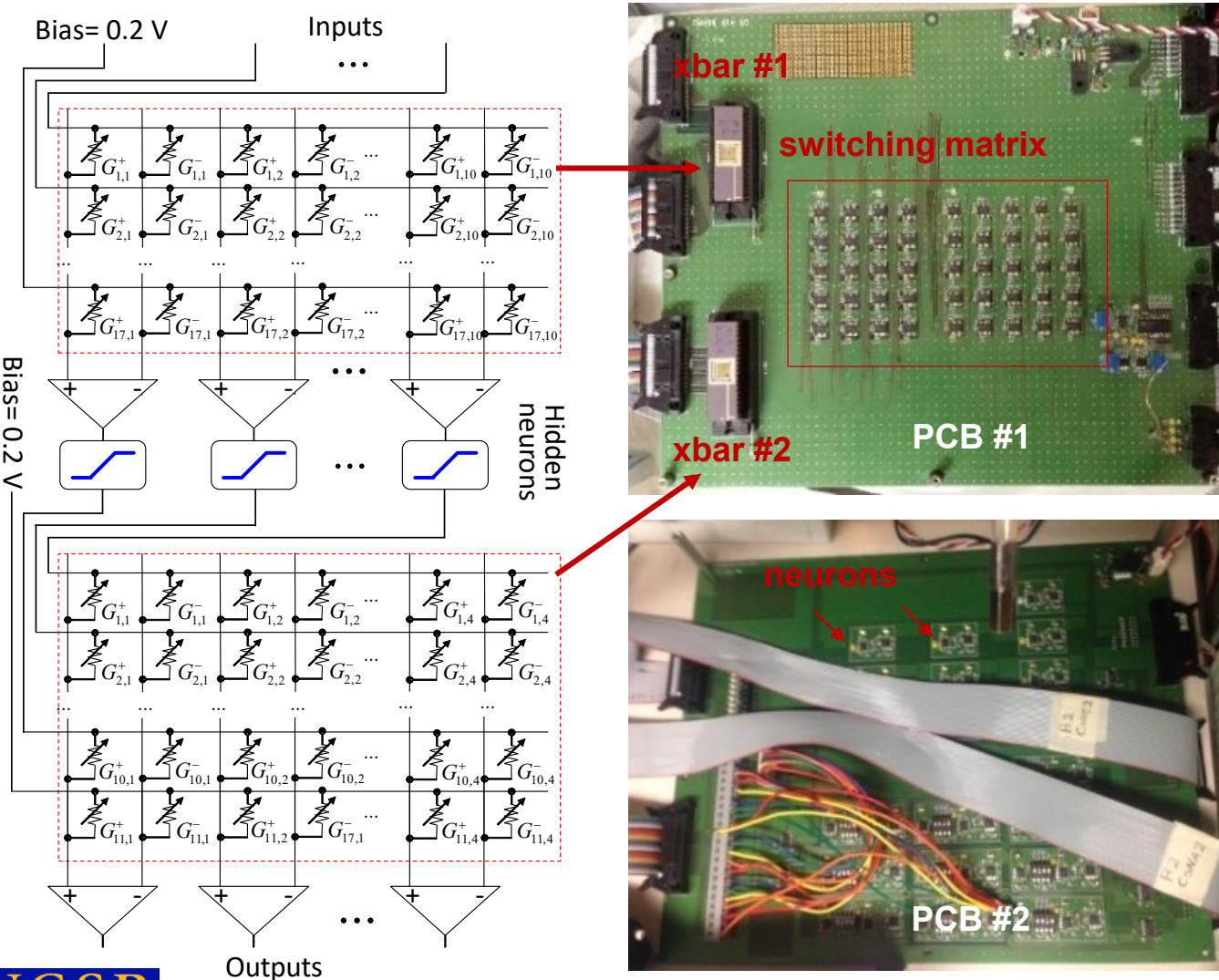
*H. Kim et al. 2019 arXiv*

- Dark dots:  $\sim 1\%$  devices that cannot be switched
- Color encoding: 256 levels from white (10  $\mu\text{S}$ ) to black (100  $\mu\text{S}$ ) @ 0.2V
- $< 5\%$  /  $< 3\%$  absolute / relative tuning error using automated algorithm, with reserves for improvement
- MNIST classification demo within 2% of the software accuracy

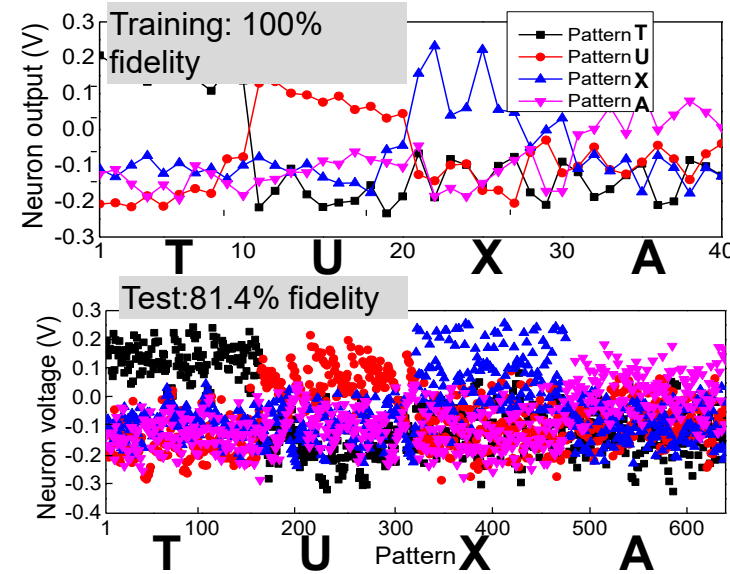


# ReRAM Based Two-Layer Perceptron Demonstration

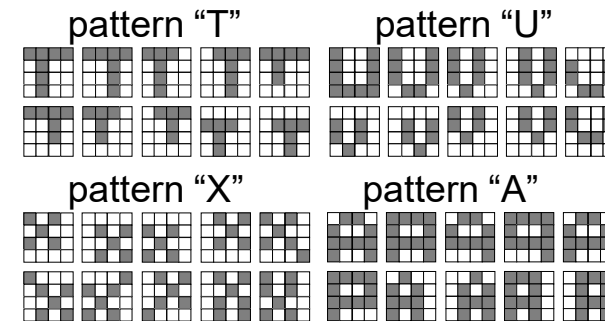
- Equivalent circuit & PCB implementation using 2x20x20 xbars



- Example of 4-pattern classification (ex-situ training)



- Training patterns (4 classes)



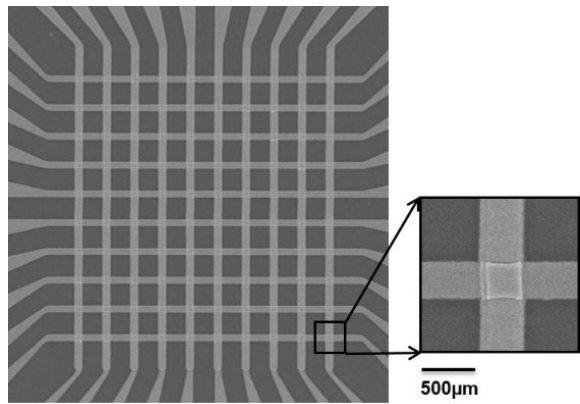
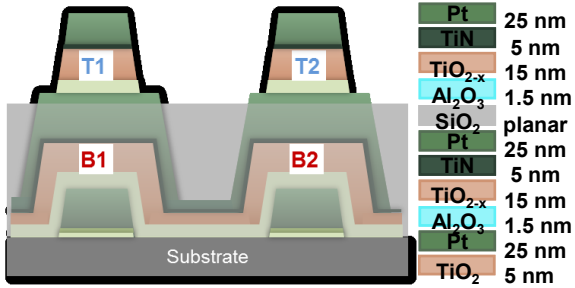
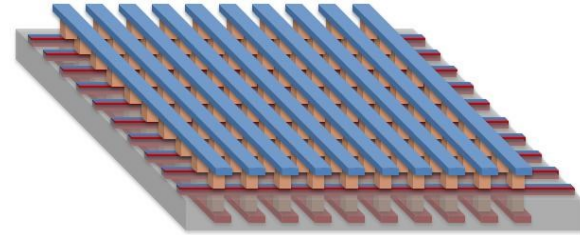
## Summary:

- Complete neural network functionality in hardware, with 428 memristor-based synapses and 14 discrete IC CMOS based neurons
- Both ex-situ and in-situ training

F. Merrikh-Bayat *et al.*,  
*Nature Comm.*, 2018

# Monolithic Analog-Grade 3D ReRAM Circuits

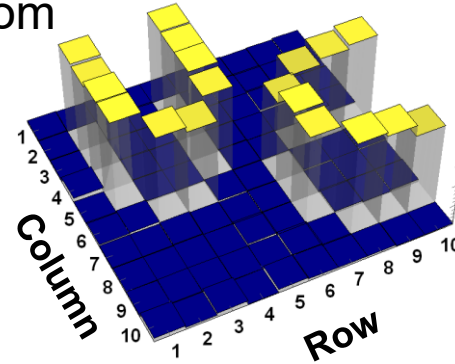
## Device structure



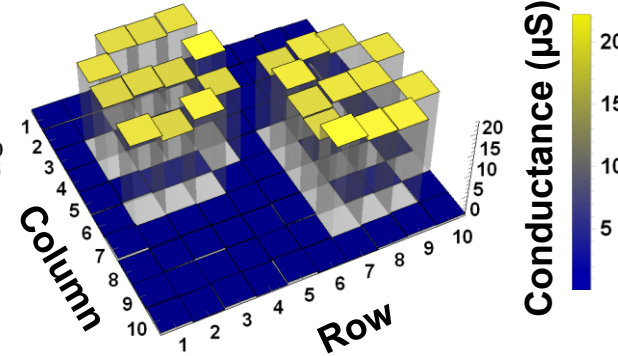
G. Adam et al. (ESSDERC'16, TED'17)

## Digital operation

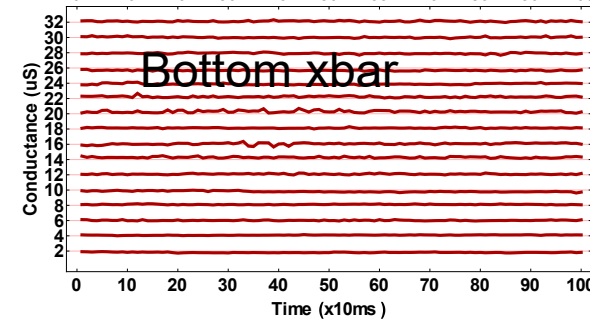
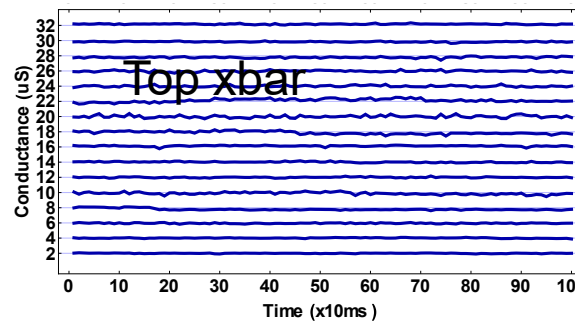
Bottom xbar



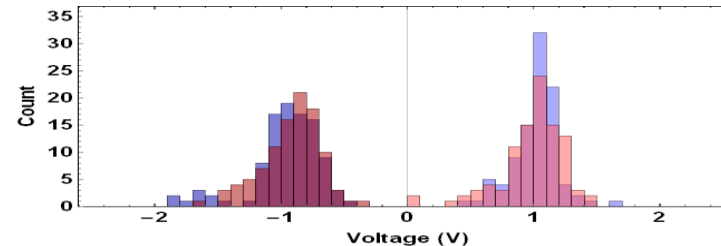
Top xbar



## Analog memory operation



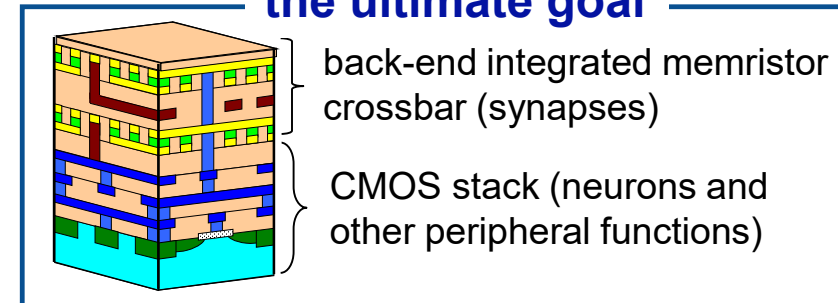
## Switching threshold statistics



### Details:

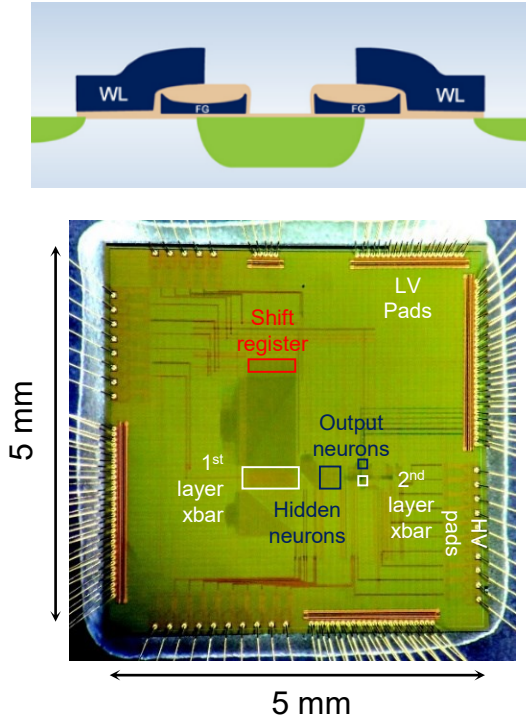
- Functional multilayer analog-grade xbar with passive Al<sub>2</sub>O<sub>3</sub>/TiO<sub>2-x</sub> devices
- In-situ material stack fabrication by ion milling
- High uniformity and device yield

## the ultimate goal

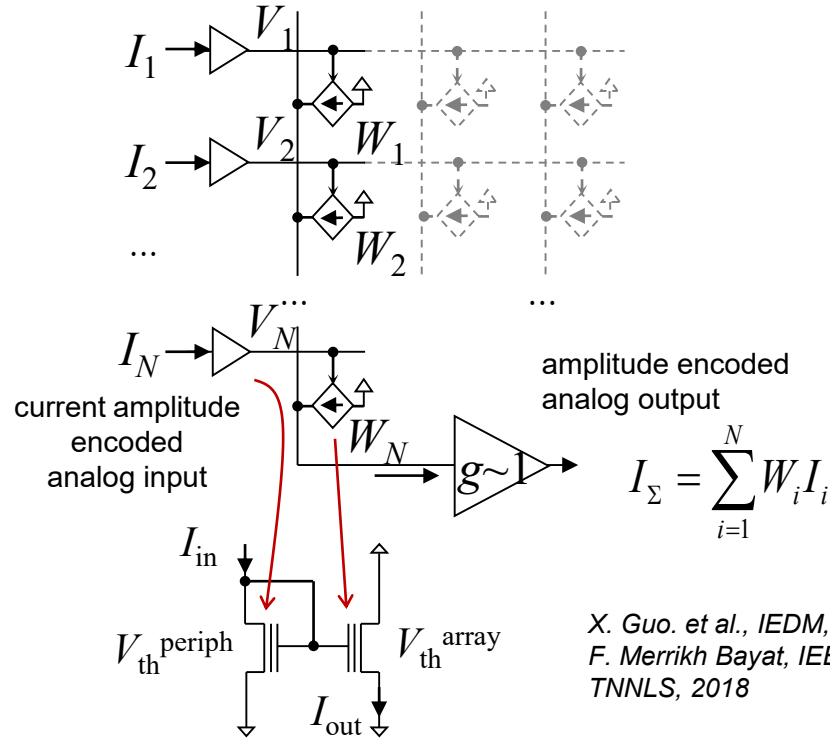


# Near-Term Option: Floating-Gate Devices

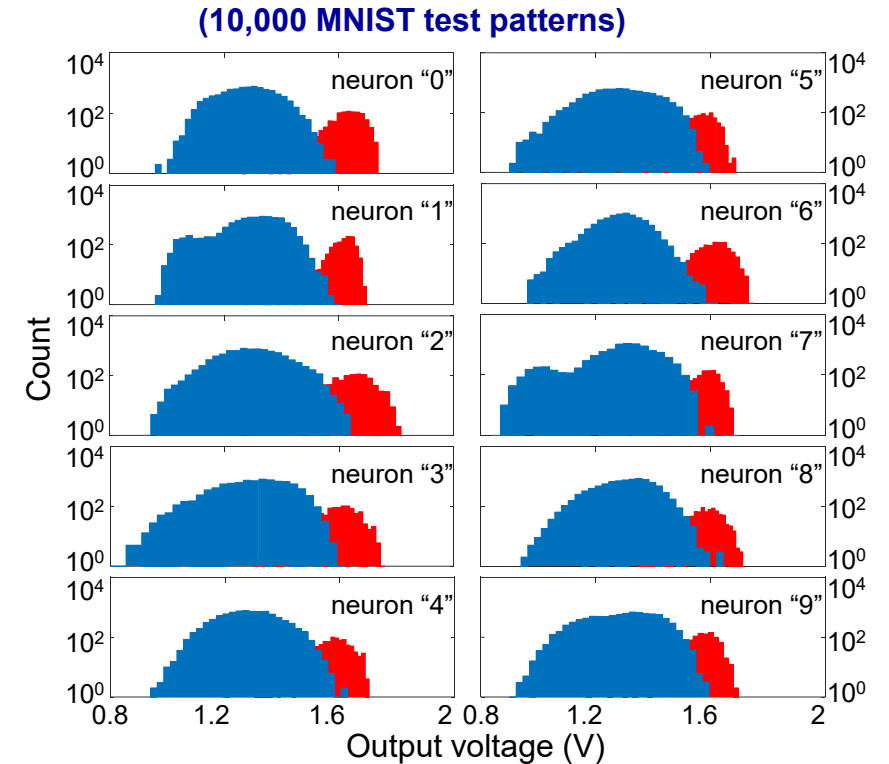
## NOR eFlash device & chip



## Vector-by-Matrix Multiplier Circuit



## 2-layer MLP classification results

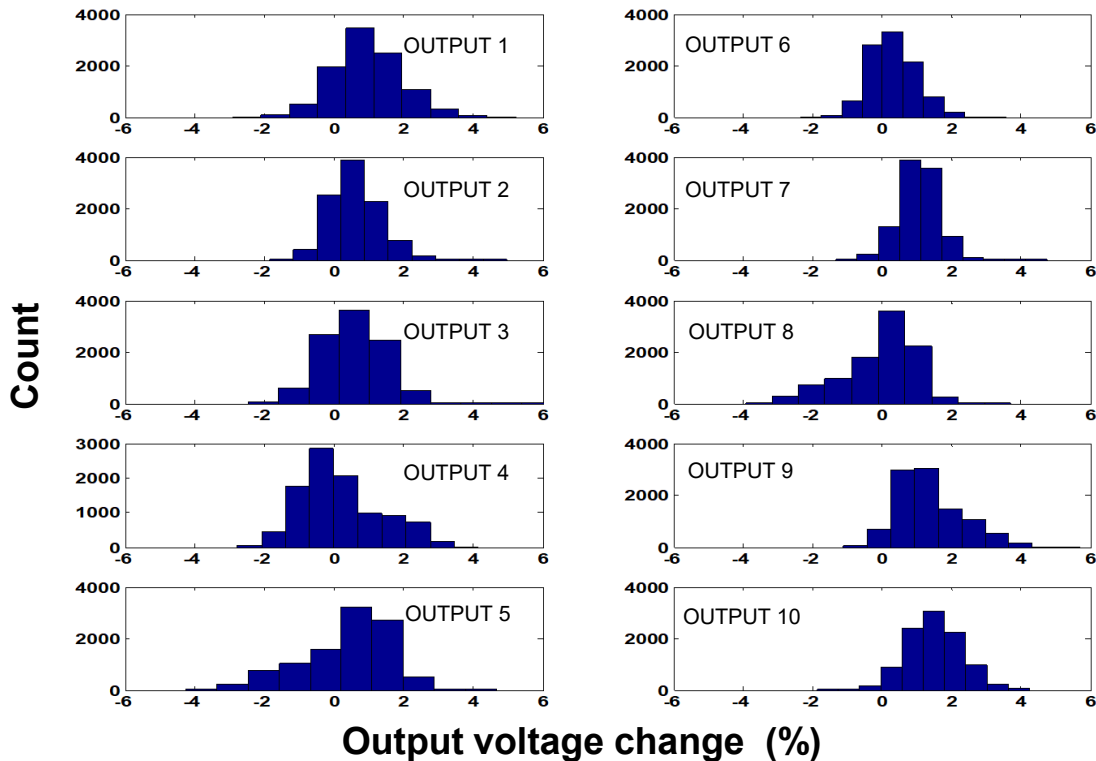


### Summary:

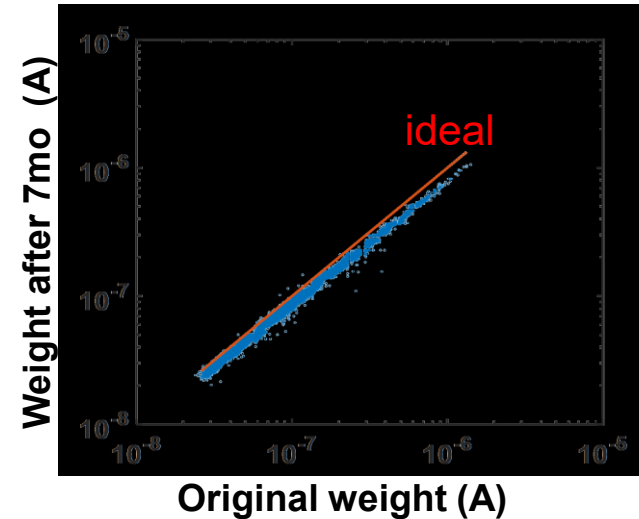
- 28x28 B/W input, 10-class output, >100,000 NOR flash synapses, 64 hidden layer CMOS neurons, 180-nm process with eFlash
- 94.65% experimental fidelity (96.5% theoretical)
- < 1- $\mu$ s latency, < 20 nJ energy per pattern (reserves for improvement for both with better neuron design)
- Much better in speed and energy efficiency over digital circuits at comparable MNIST fidelity ( $10^6$  better energy-delay than IBM TrueNorth)
- Reproducible, temperature insensitive, no change in performance after 7 months shelf-time, without any cell retuning
- More recent work using 55-nm ESF3 NOR-flash technology (CICC'17, IEDM'18'19), scalable to 28 nm

# Experimental Results: Retention, Temperature Sensitivity, and Chip-to-Chip Statistics

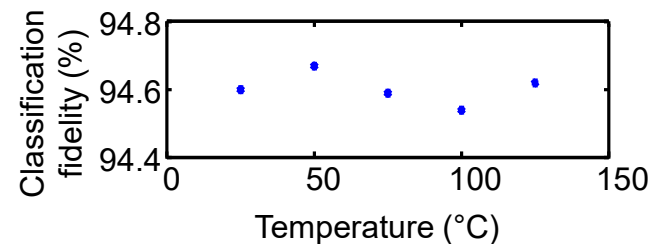
## Change in output after 7 months (10,000 MNIST test patterns)



## Weight drift in 7 months



## Temperature sensitivity



## Summary:

- Reproducible results: 94.6%, 94.1%, 94.2% fidelity for three different chips (with ~ 5% tuning accuracy for all)
- Temperature insensitive:  $\pm 0.1\%$  change in fidelity over 100°C range
- Slight drift in cells' conductance in 7 month but no impact on classification fidelity!
- Cost function optimized for to maximize output margins

F. Merrikh Bayat et al., *IEDM'17*

# Part II.

# Probabilistic Neural Networks

# Noise in Biological and Artificial Neural Networks

Molecular-level operations in the brain, e.g. neurotransmitter release in synaptic clefts and voltage gating of ion channels, are stochastic

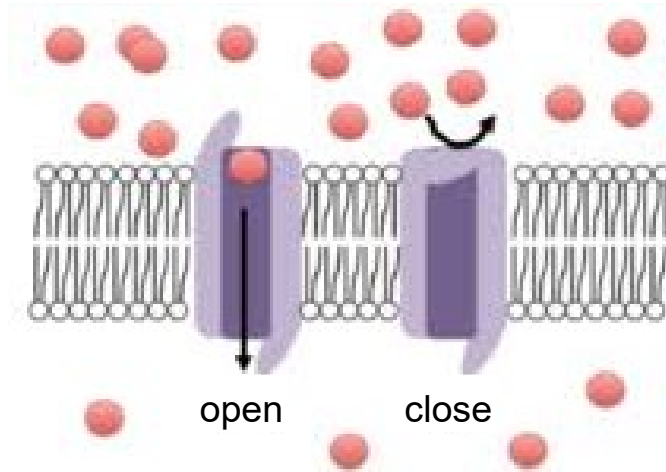
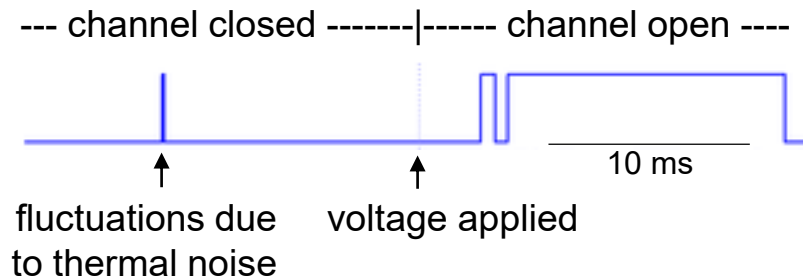
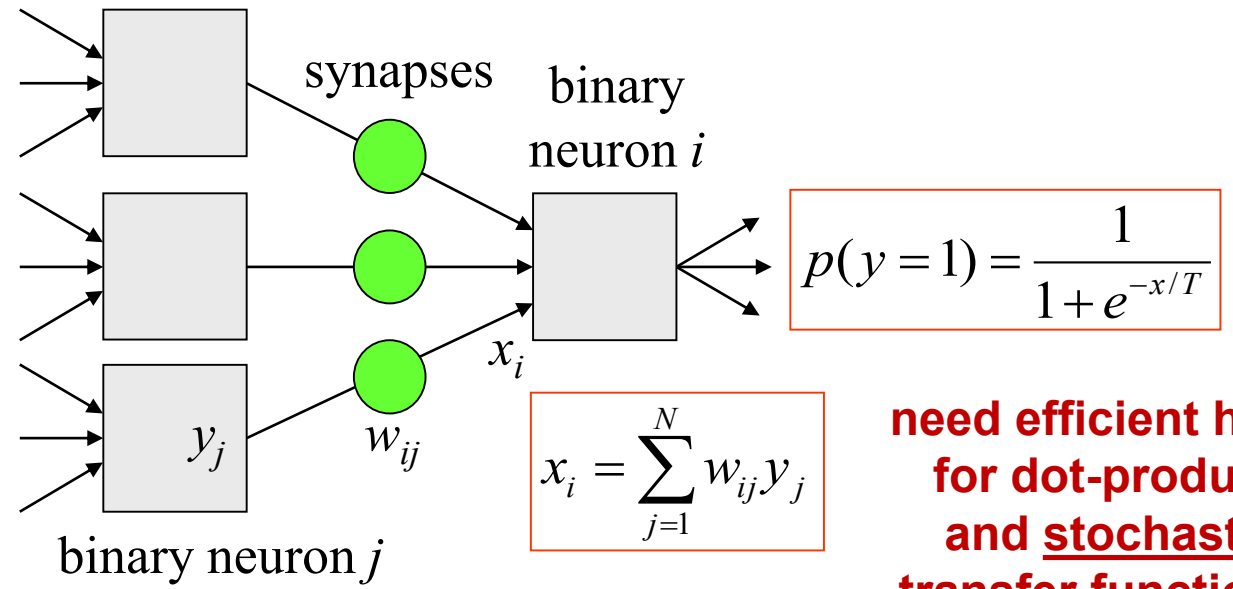


Image sources:  
Scholarpedia

Example: fluctuations in K channel



## Stochastic (binary) neuron



**need efficient hw  
for dot-product  
and stochastic  
transfer function**

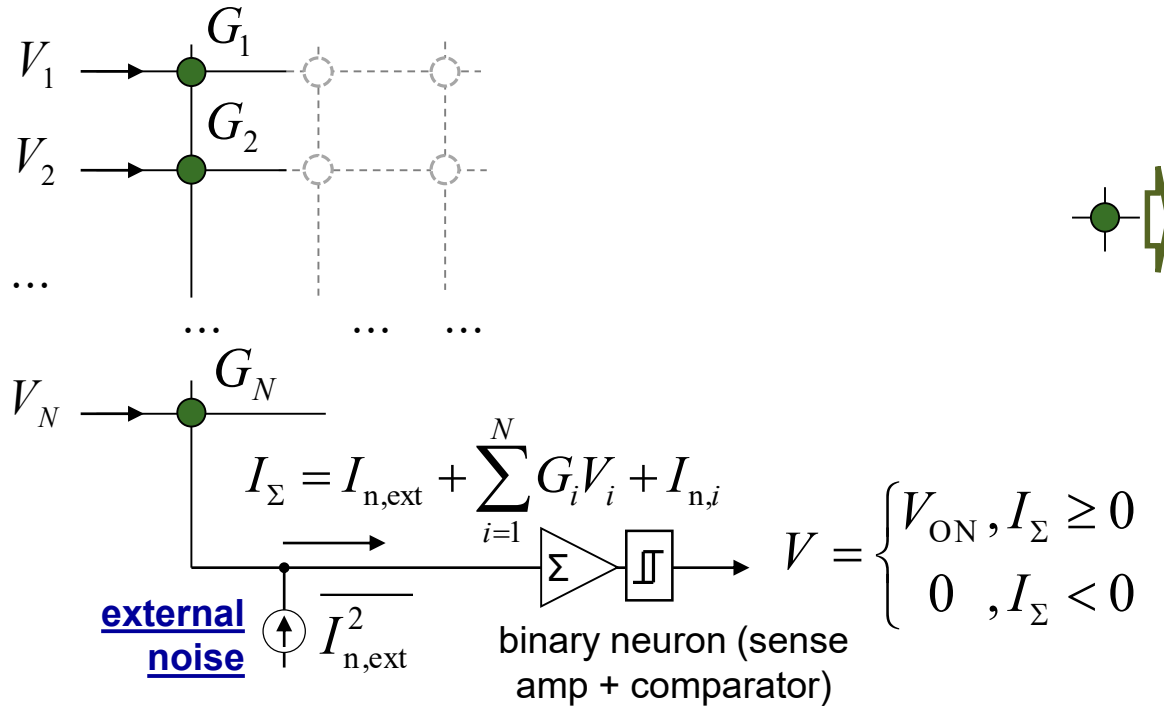
Stochastic neural networks:

- (Restricted) Boltzmann machines
- Stochastic Hopfield networks
- Deep believe networks
- Bayesian networks
- ...

# Stochastic Analog Vector-by-Matrix Multiplier

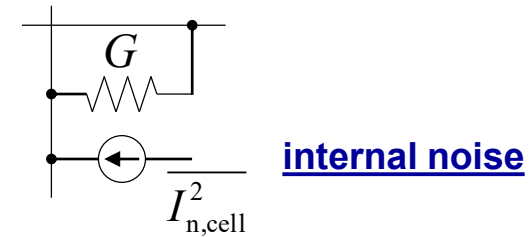
## Basic Idea:

add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator

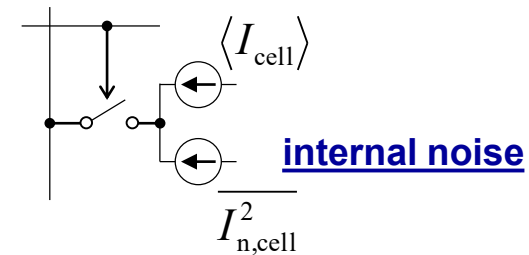


## Two Implementation Options:

0T1R memristor cell (works for 1T1R as well)



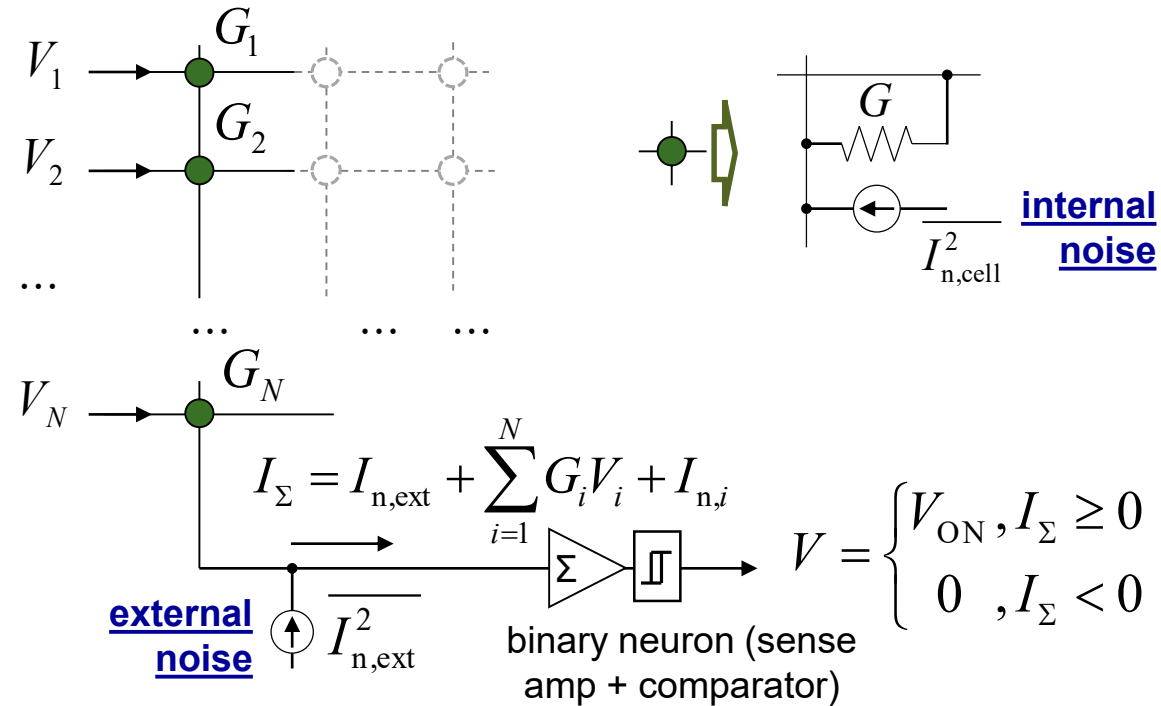
Floating gate transistor



# Stochastic Analog Vector-by-Matrix Multiplier

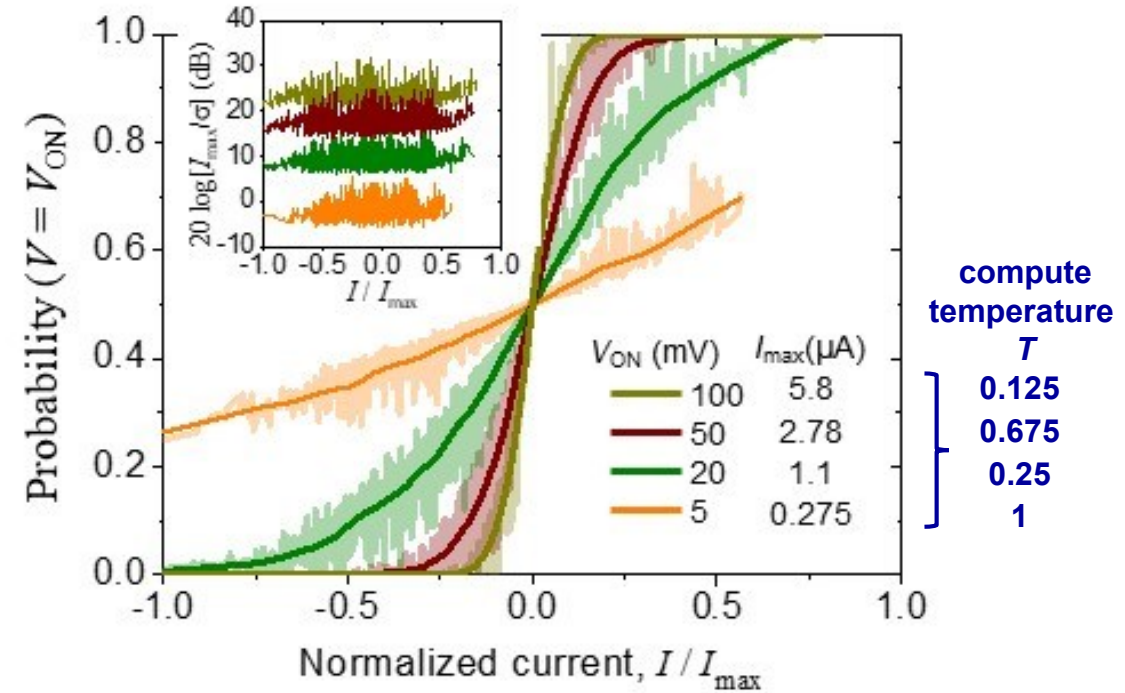
## Basic Idea:

add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator



## Experimental Demo:

using  $20 \times 20$  passive array with externally-injected noise from readout circuitry



M.R. Mahmoodi et al. Nature Communications, 2019

## Features:

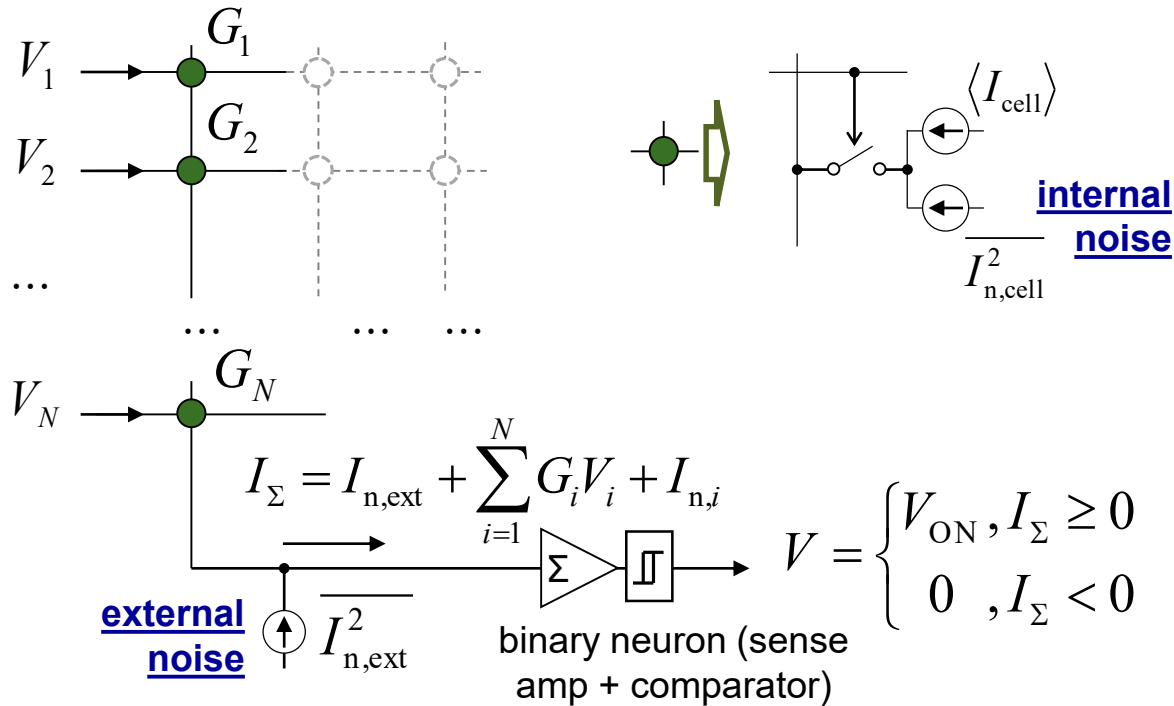
- Sigmoid slope (i.e. SNR or compute temperature  $T$ ) controlled dynamically by the applied voltage  $V_{\text{ON}}$
- Some smearing of output probabilities due to input-dependent noise and device imperfections



# Stochastic Analog Vector-by-Matrix Multiplier

## Basic Idea:

add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator

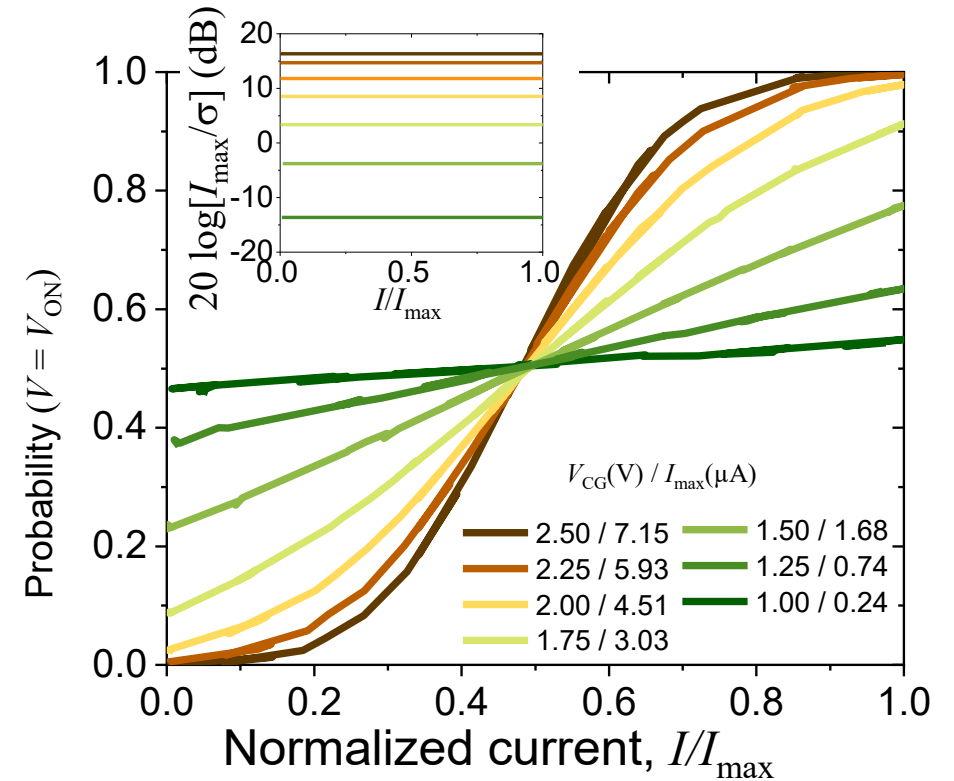


## Features:

- Sigmoid slope (i.e. SNR or compute temperature  $T$ ) controlled dynamically by the applied gate voltage

## Experimental Demo:

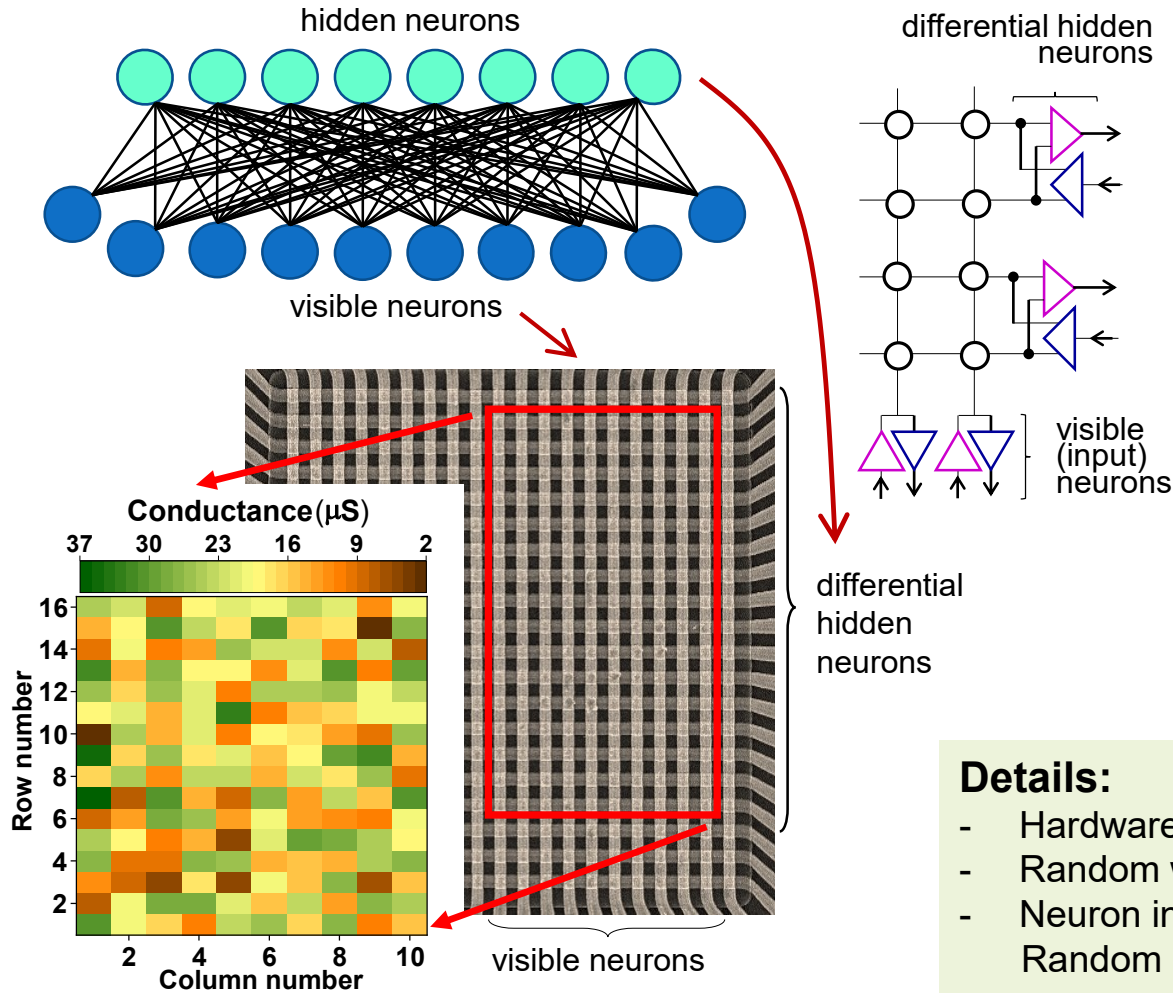
using 180nm embedded ESF1 NOR-flash memory technology



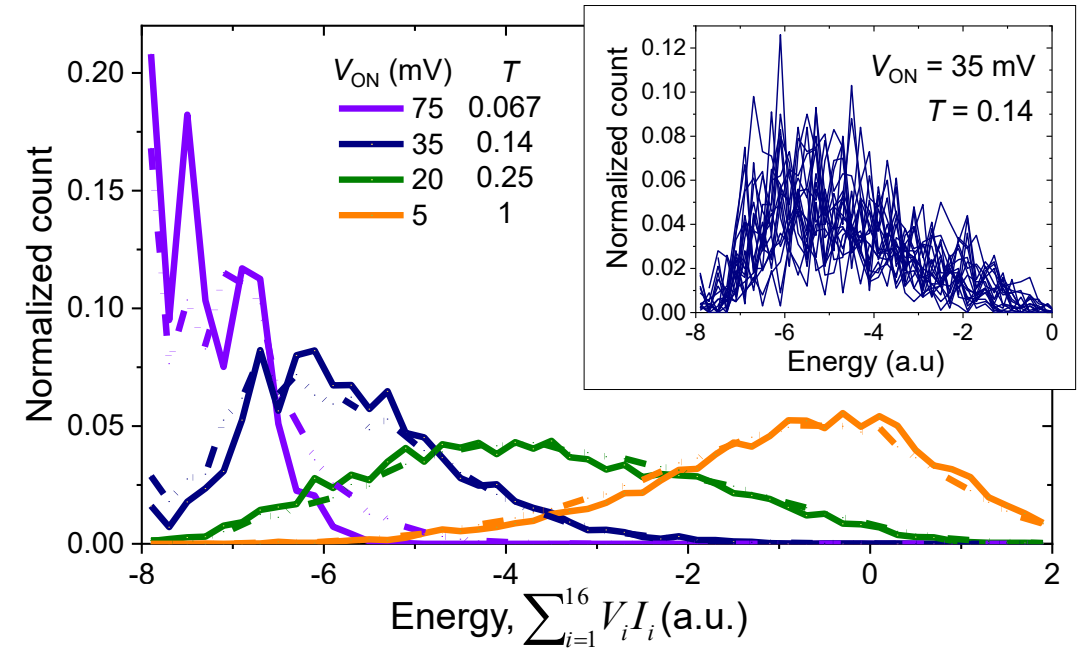
M.R. Mahmoodi et al. Nature Communications, 2019

# Restricted Boltzmann Machine Demo

- 10-input 8-hidden neuron RBM network



- Experiment (solid) vs. simulation (dash-dot)



### Details:

- Hardware injected noise with software-emulated neuron functionality
- Random weights (from  $-32 \mu\text{S}$  to  $+32 \mu\text{S}$ ) mapped to  $10 \times 16$  portion of memristor xbar
- Neuron input currents sampled at 1 MHz bandwidth after applying Random Input  $\rightarrow$  Visible  $\rightarrow$  Hidden  $\rightarrow$  Visible  $\rightarrow$  Hidden  $\rightarrow$  ...

## Part IIb

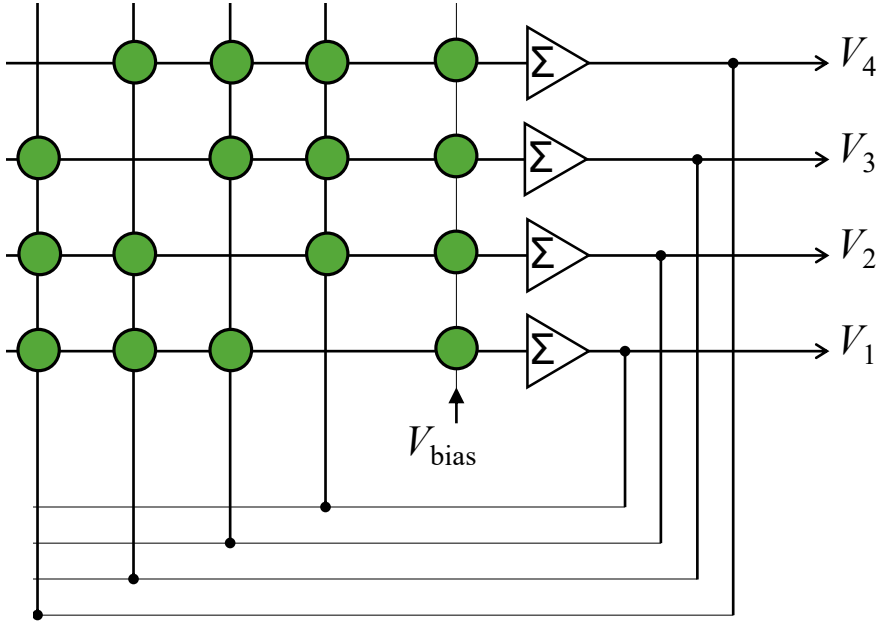
# Probabilistic Neural Networks for Neurooptimization

# Solving Optimization Problems with Hopfield Neural Network

- Combinatorial optimization problems

Application	Problem
Logistics / package delivery	Traveling salesman
Power grid	Maximum flow
Design automation	Vertex cover
Molecular dynamic simulations	Graph partitioning

- Example of continuous time / binary neuron Hopfield network



Σ = sum amp & comparator

- Solving TSP with Hopfield neural network

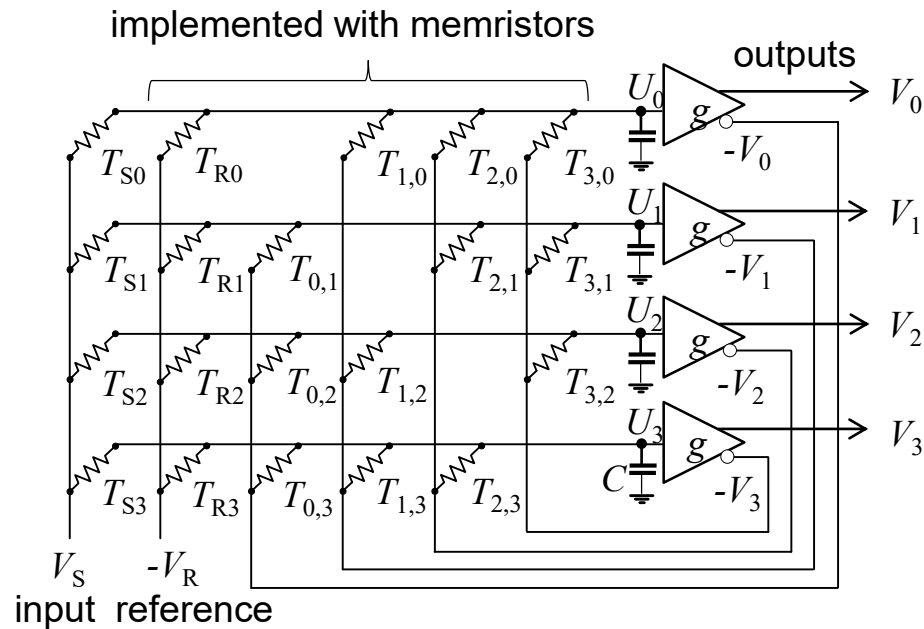
Traveling Salesman Problem: NP hard → use heuristics, e.g.

- single route = specific neuron outputs
- finding optimal solution = minimizing “energy” function of neuron outputs
- dynamics of the recurrent network with *proper weights* minimizes energy function over time



# Earlier Work: (Deterministic) Hopfield Network Experimental Demonstration with Discrete Memristors

## ▪ Hopfield network for A-to-D conversion

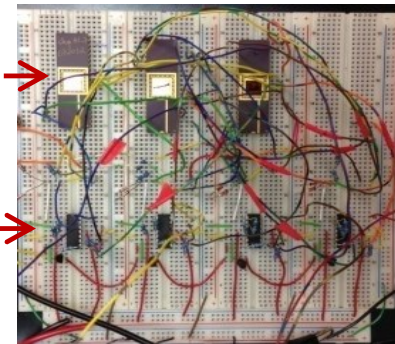


### Major features:

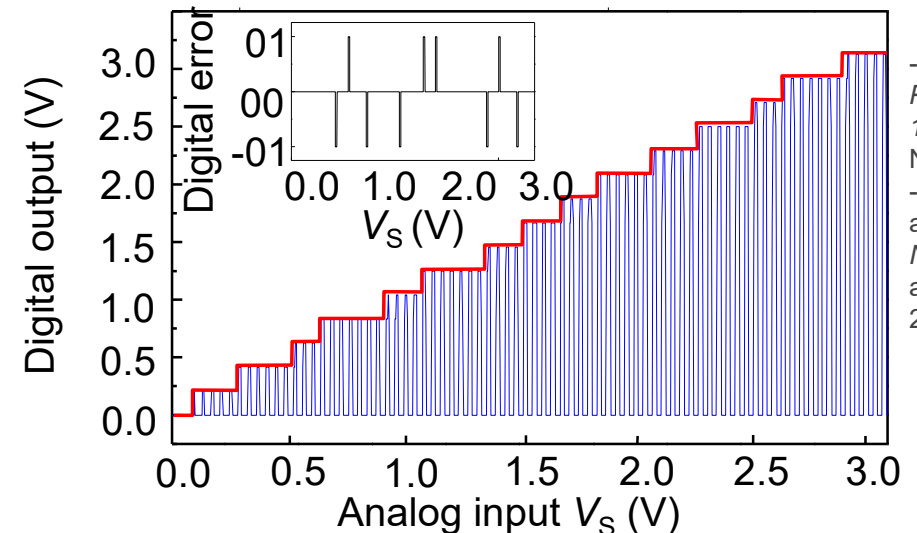
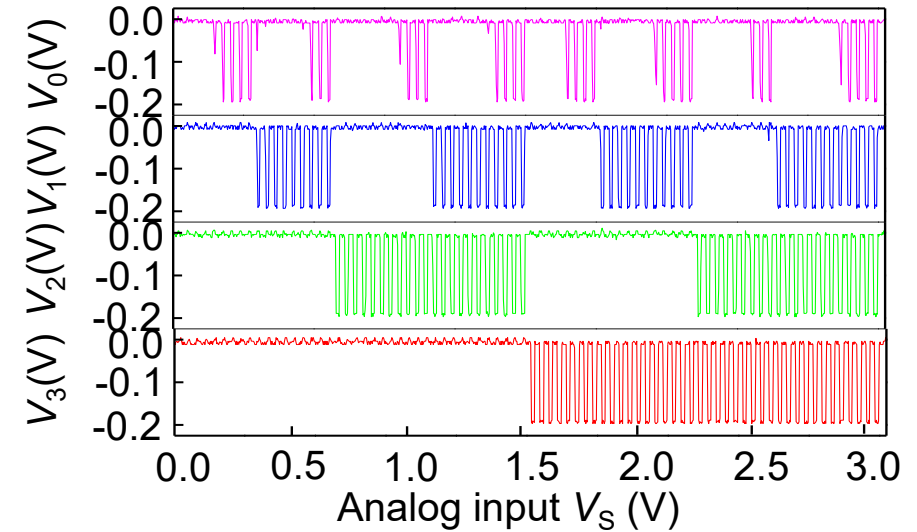
- 4-bit ADC implemented as a Hopfield network
- The first demo for the memristor-based Hopfield neural network
- CMOS discrete IC neurons
- Discrete packaged memristors
- Fine-tuning to cope with offsets and variations

chips with single-device memristors

TL074CN opamp

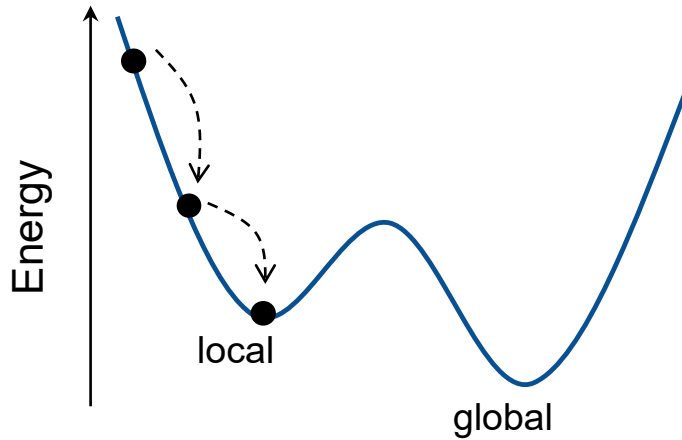


## ▪ Experimental results

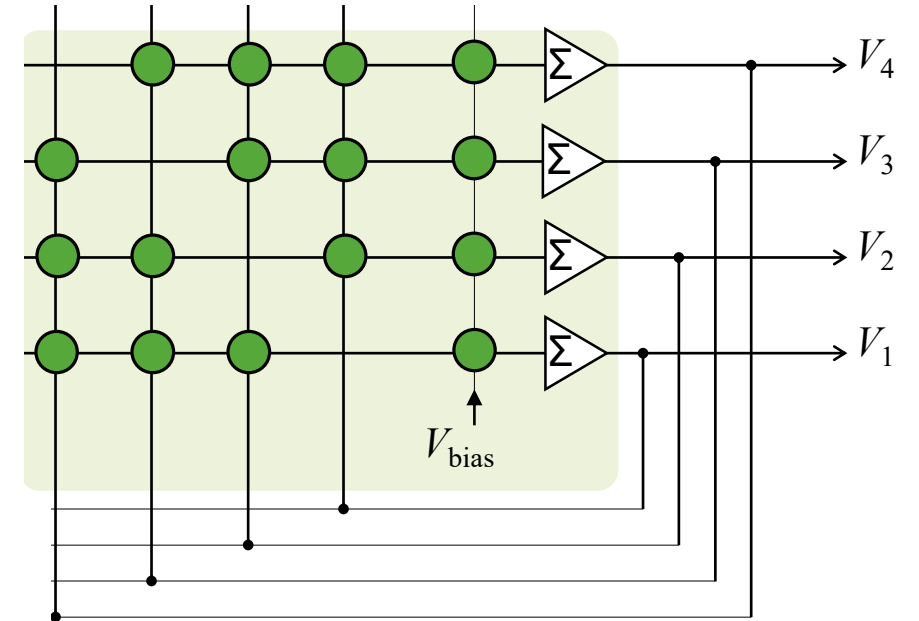


- L. Gao et al, in: *Proc. NanoArch'13*, Ney York, NY July 2013;  
 - X. Guo et al., *Frontiers in Neuroscience* **9**, art. 488, Dec. 2015

# Local Minima in Hopfield Network



Local minima present problems!

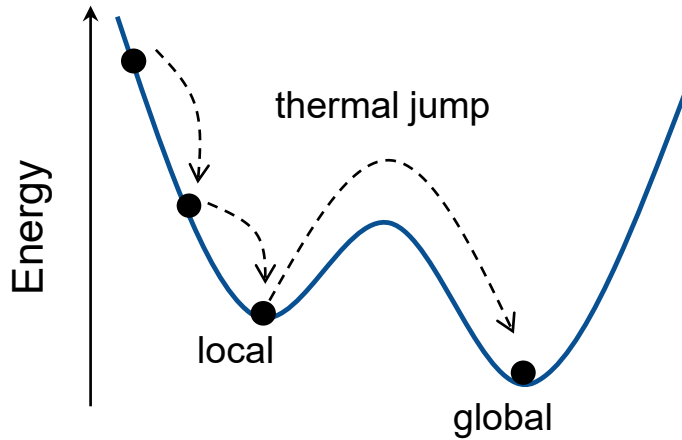


Color background:

 Baseline Hopfield neural network

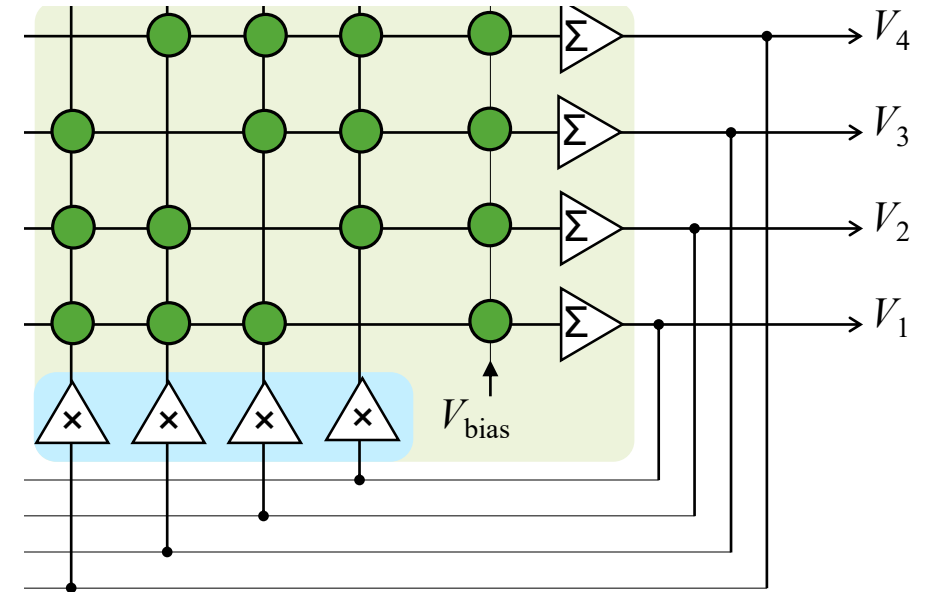
$\Sigma$  = sum amp & comparator

# Simulated Annealing with Generalized Hopfield Network (Boltzmann Machine)



Local minima present problems!

Solution: employ probabilistic neurons (stochastic VMMs) to implement simulated annealing



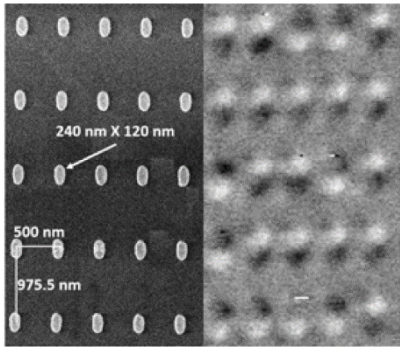
**Color background:**

- Baseline Hopfield neural network
- Stochastic annealing

Σ = sum amp & comparator  
× = scaling

# Emerging (Custom) Hardware for Combinatorial Optimization

## Nanomagnets / P-bits



Experimentally measured ground states for the network consisting of up to 3 coupled magnetic devices with fixed coupling;

- **Limited (near neighbor, fixed) coupling and/or ...**

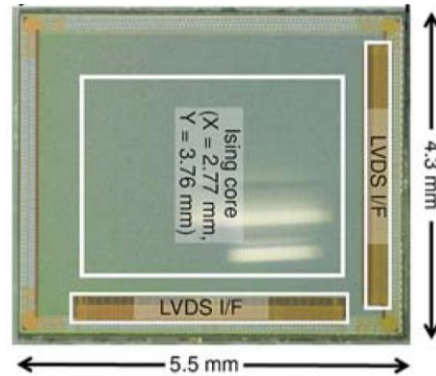
Debashis *et al.* *IEDM* 2016

Integer (up to 945) factorization with 8 p-bits

- **... high CMOS overhead**

Borders *et al.* *Nature* 573 2019

## CMOS

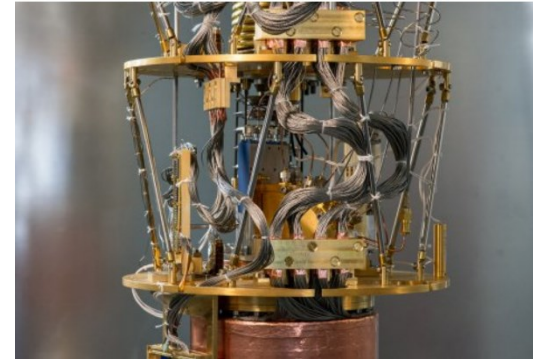


Experimental results for solving maximum-cut problem with 2×30K-spin Ising network 40-nm 23.65-mm<sup>2</sup> SRAM-based chips

- **Not in-memory (bulky, slower, power hungry)**
- **Binary weights**

Takemoto, *et al.* *ISSCC* 2018

## Josephson Junction

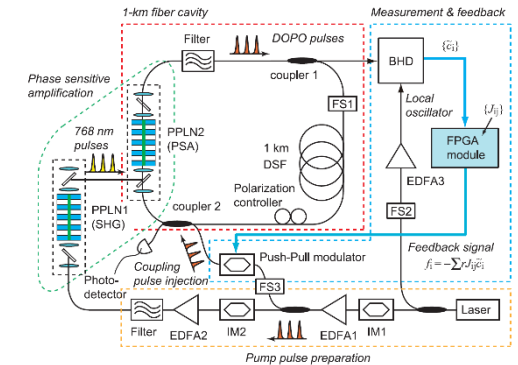


Experimentally measured ground state of random spin glass problems based on 108-qubit D-Wave One system (with evidence of quantum annealing)

- **Low temperature operation**
- **Many issues unsolved**

Boixo, *et al.* *Nature phys.* 2018

## Photonics



Experimental results for solving max-cut problems with up to 2,000 nodes with Ising network based on degenerate optical parametric oscillators

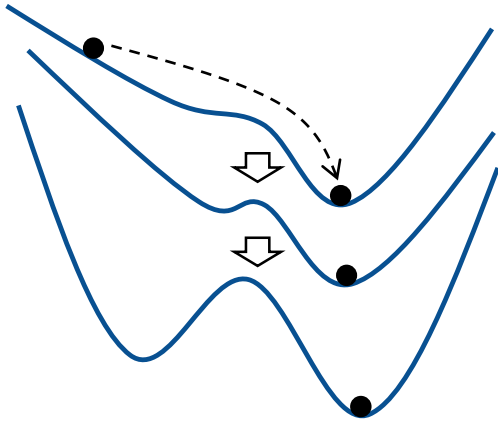
- **Slow due to high overhead of the electronic feedback used for updating spatial light modulator**

Inagaki, *et al.* *Science* 2016

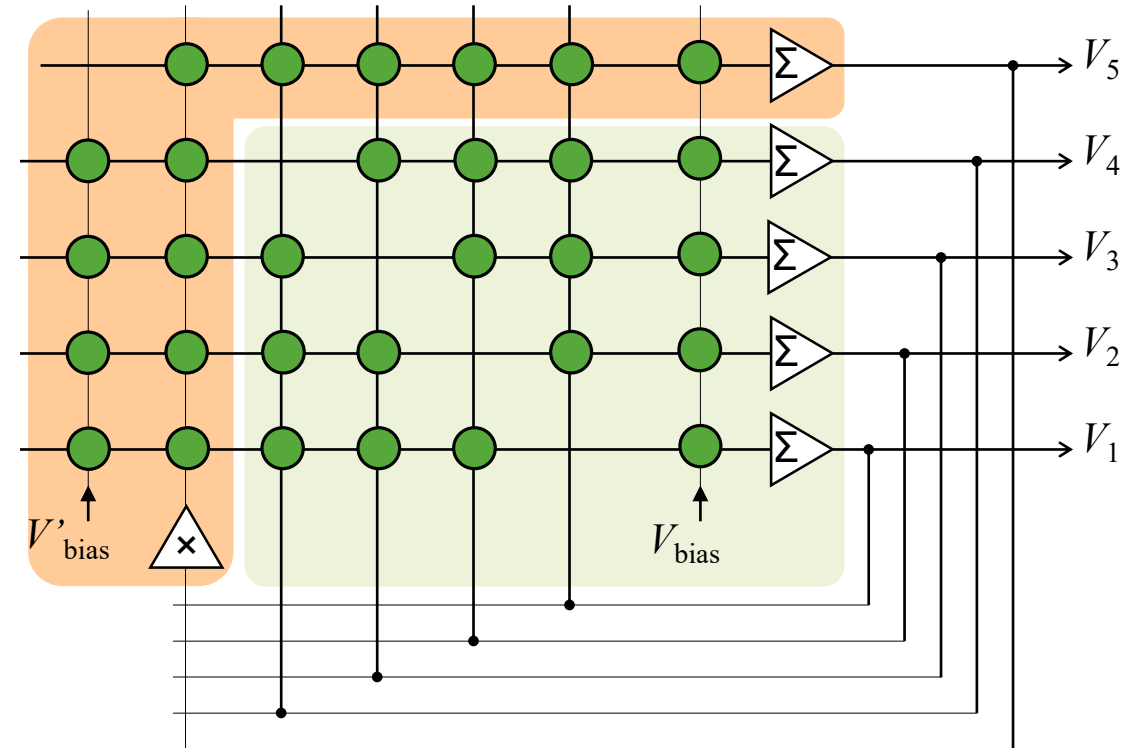


# Adjustable Energy Function Annealing

$$\text{Energy} = E_{\text{original}} + \exp(-\text{time})E_{\text{addon}}$$



Another solution inspired by quantum annealers: Dynamically adjustable energy function



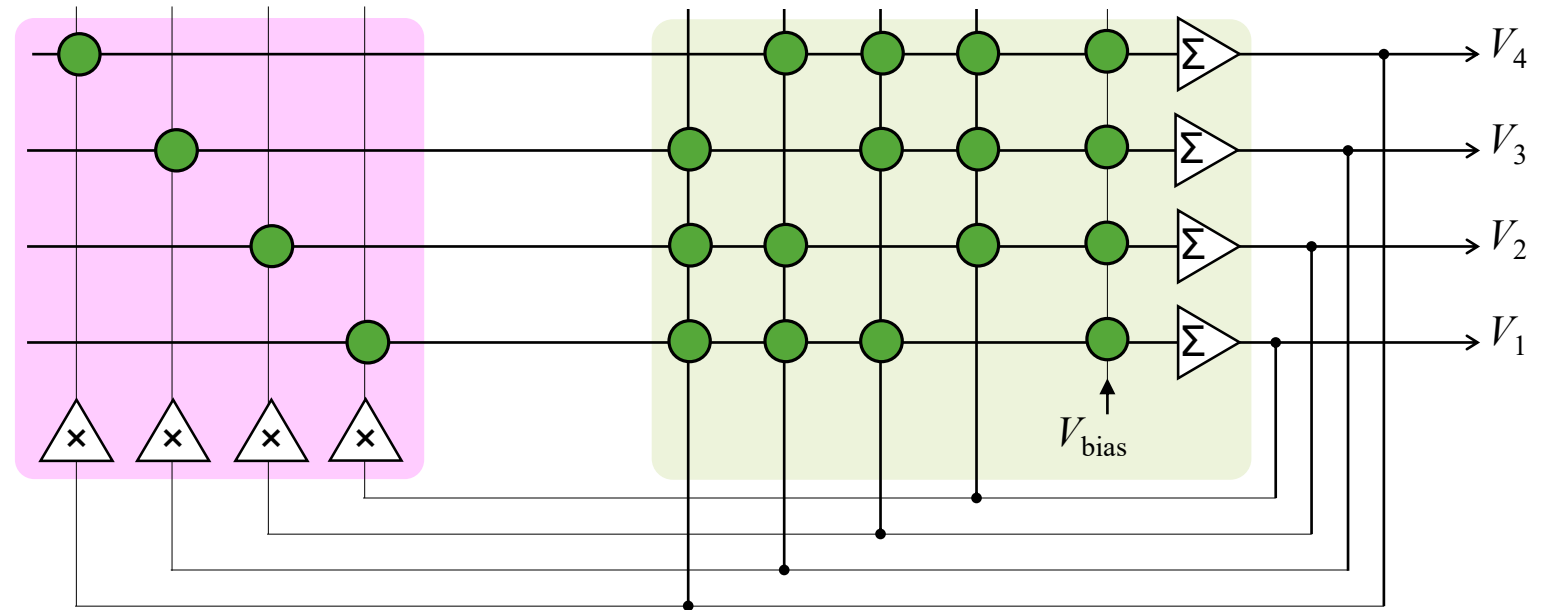
**Color background:**

Baseline Hopfield neural network

Adjustable energy function / weight annealing

$\Sigma$  = sum amp & comparator  
 $\times$  = scaling

# Yet Another Approach: Chaotic Annealing



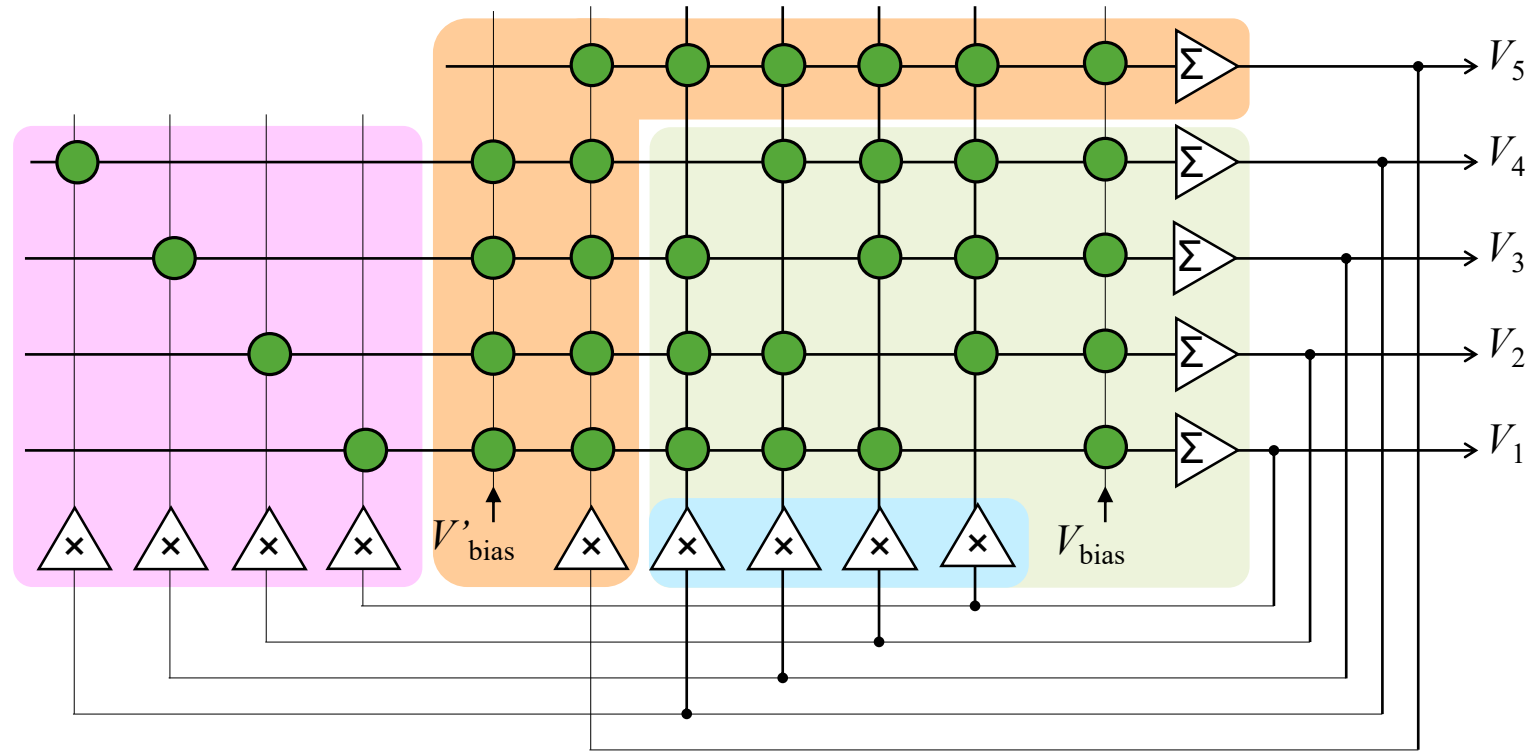
**Color background:**

Baseline Hopfield neural network

Chaotic annealing

$\Sigma$  = sum amp & comparator  
 $\times$  = scaling

# Flexible-Annealing Mixed-Signal Generalized Hopfield Networks for Combinatorial Optimization



**Color background:**

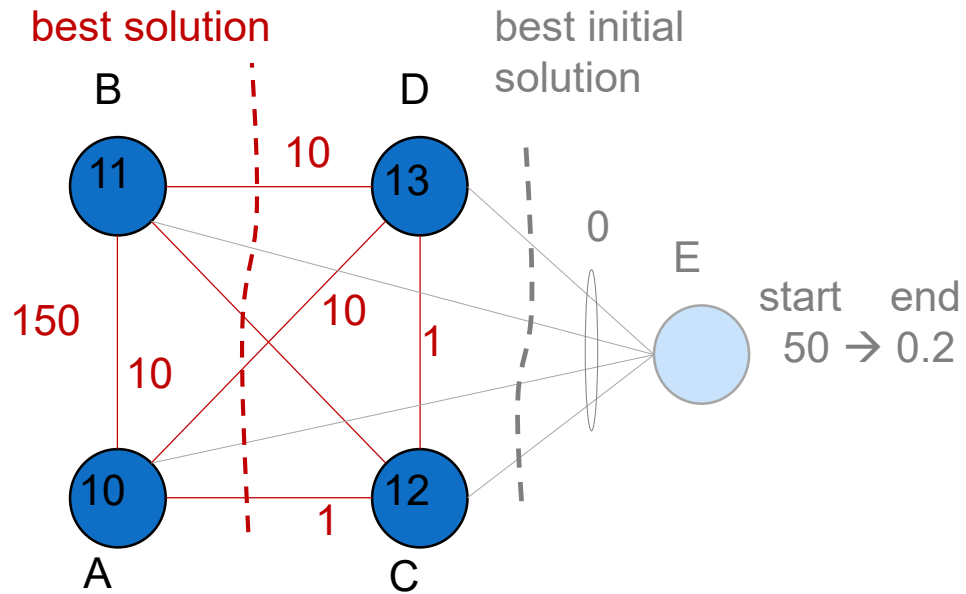
- Baseline Hopfield neural network
- Stochastic annealing
- Adjustable energy function / weight annealing
- Chaotic annealing

$\Sigma$  = sum amp & comparator

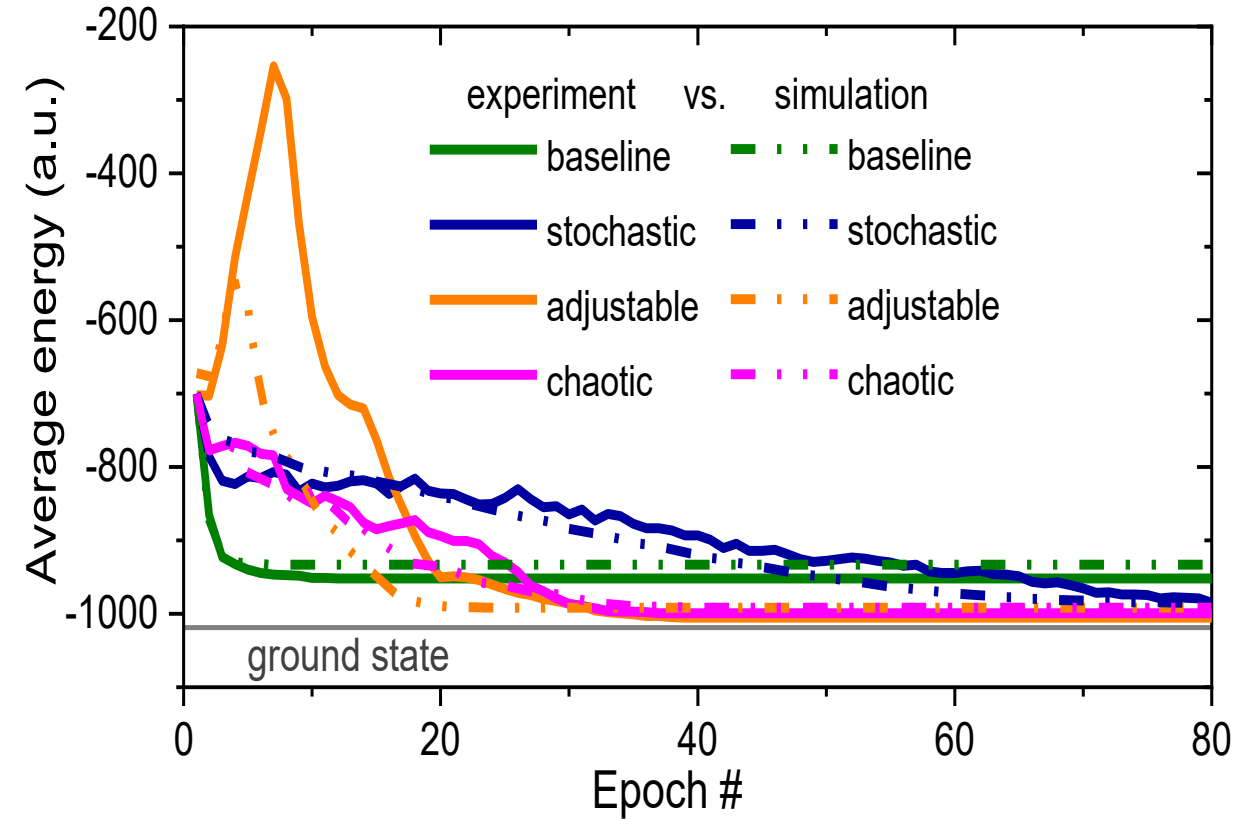
$\times$  = scaling

# Combinatorial Optimization Demo with FG

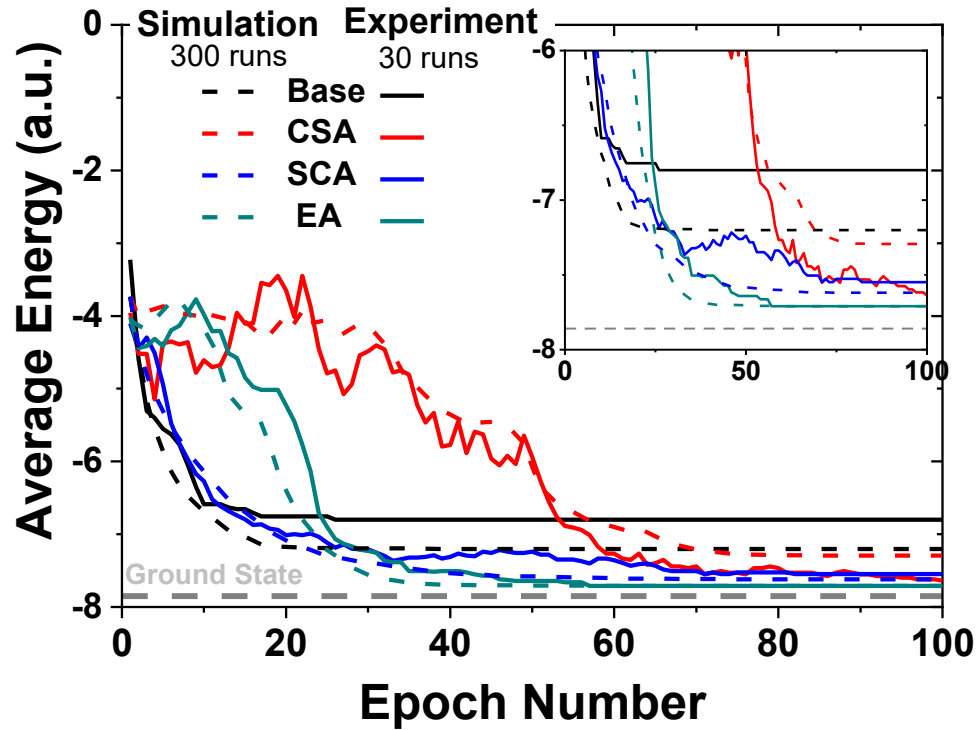
- **Weighted graph partitioning problem...**  
(finding two mutually exclusive, set of nodes with maximally balanced node weights and minimized edge weights between two sets)



... and experimental results using  
**10×20 180-nm NOR flash memory array**



# Combinatorial Optimization Demo with Passive 64×64 Metal-Oxide Memristive Crossbar Circuits



Example of solving 5-node maximum-weighted clique problem

Additional problems demonstrated:

- 12-node maximum-weighted vertex cover
- 10-node maximum-weight independent set
- 6-node maximum-weight graph partitioning

M.R. Mahmoodi et al., *Proc. IEDM'19*

## Performance estimates & comparison to competition\*

	Conventional		Emerging technology		This work	
	CPU	GPU	D-Wave	Fiber optics	Memristor	NOR flash
Time to solution ( $\mu\text{s}$ )	220	10	$10^{10}$	600	3	10
Energy to solution ( $\mu\text{J}$ )	4000	2500	$250 \times 10^{12}$	?	0.2	0.6

\* benchmarked on noisy mean-field algorithm, adapted from *ArXiv:1903.11194*, which presented similar idea though with 1T1R devices, simulated (and less controllable) noise, and binary weights.

# Part III.

# Spiking Neural Networks

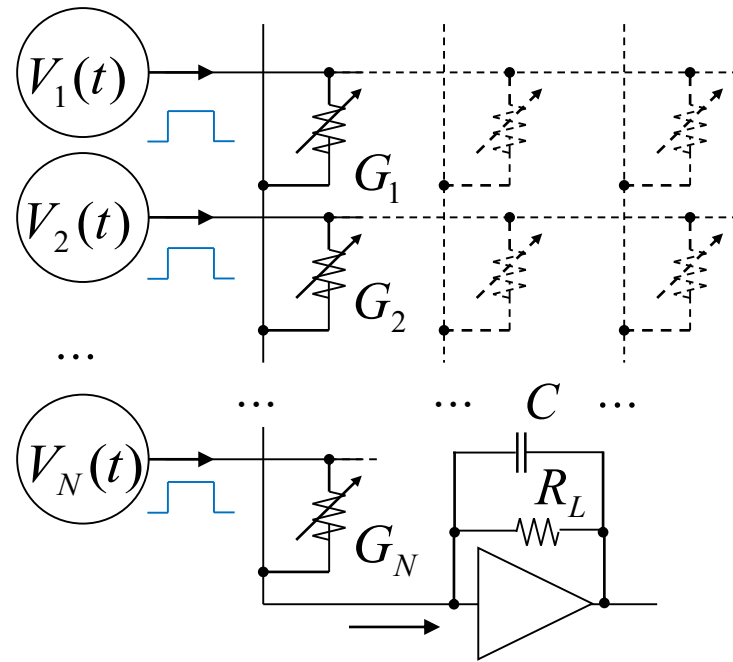
# Spiking Neural Networks

(the most advanced and biologically plausible artificial neural networks)

- Information encoded in timing of spikes → natural for spatial-temporal information processing, event-driven encoding
- Local learning rules for synaptic weight update → suitable for online HW-friendly training but ...
- ...outperformed (so far) by simpler backprop-trained models on common image/speech benchmarks

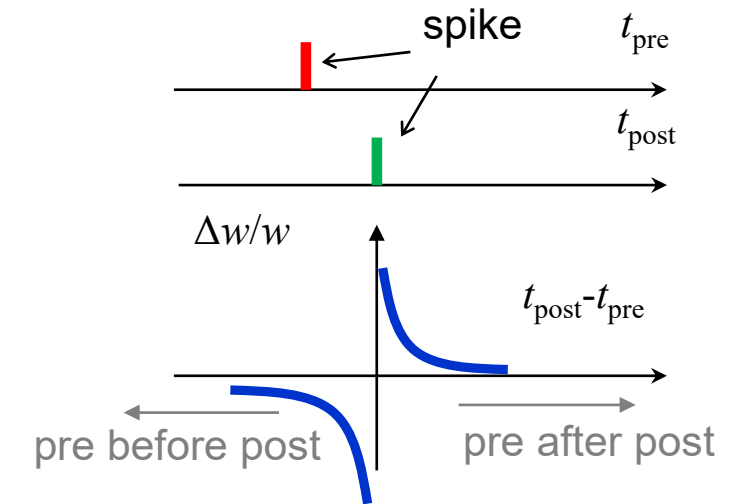
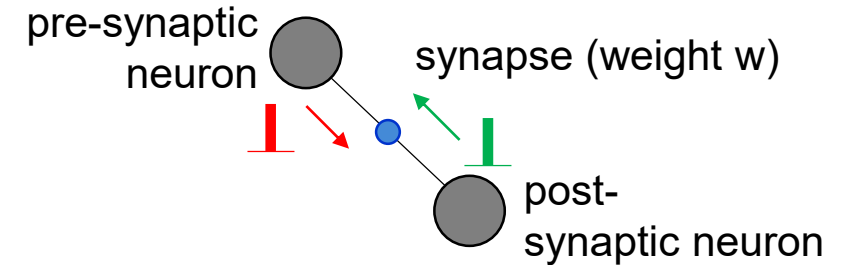
**Efficient spiking hardware to enable large-scale brain simulations**

## VMM + leaky-integrate-and-fire neuron (+ STDP)

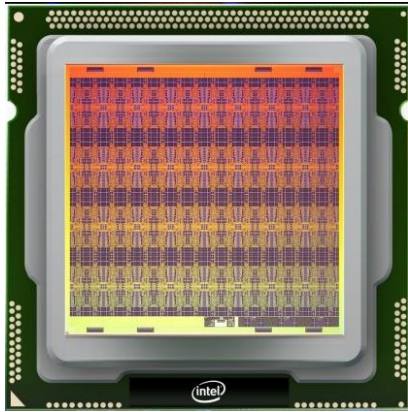


$$C \frac{dU(t)}{dt} = \sum_{j=1}^N G_j V_j(t) - \frac{1}{R_L} U(t)$$

## Spike-timing-dependent plasticity essential feature for online learning



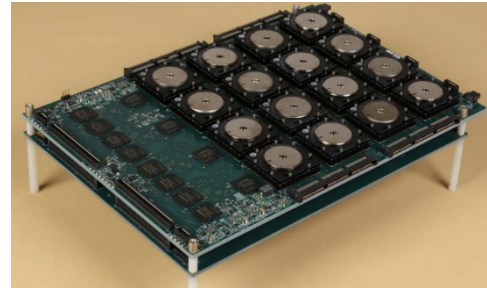
# Spiking Digital / Mixed-Signal Neuromorphic Hardware



## Loihi (Intel, 2018)

14 nm, 2.07 B transistors  
130 M synapses (up to 9b)  
131 K neurons  
on-chip learning

## TrueNorth (IBM, 2014)



28 nm, 5.4 B transistors  
256 M synapses (1b)  
1 M neurons  
inference only



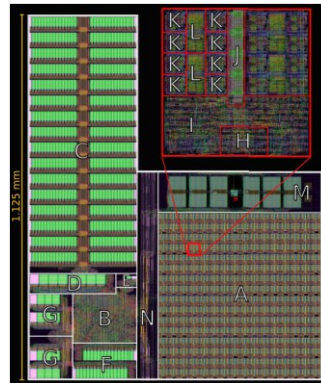
## SpiNNaker (U Manchester, 2019)

130 nm  
ARM-uP based  
1 K neurons / core  
1 M cores, 7 TB RAM

## BrainScaleS (Heidelberg U, 2015)

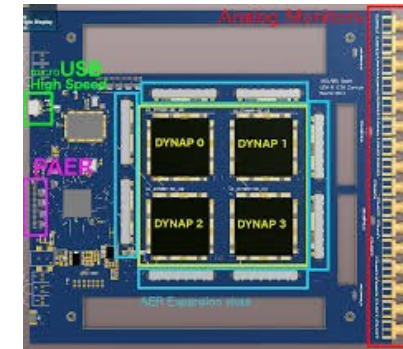


180 nm  
wafer-scale  
40 M synapses  
180 K neurons



## BrainDrop (Stanford, 2018)

28 nm  
64 K synapses  
4096 neurons  
all-to-all



## DYNAP-SE (INI Zurich, 2018)

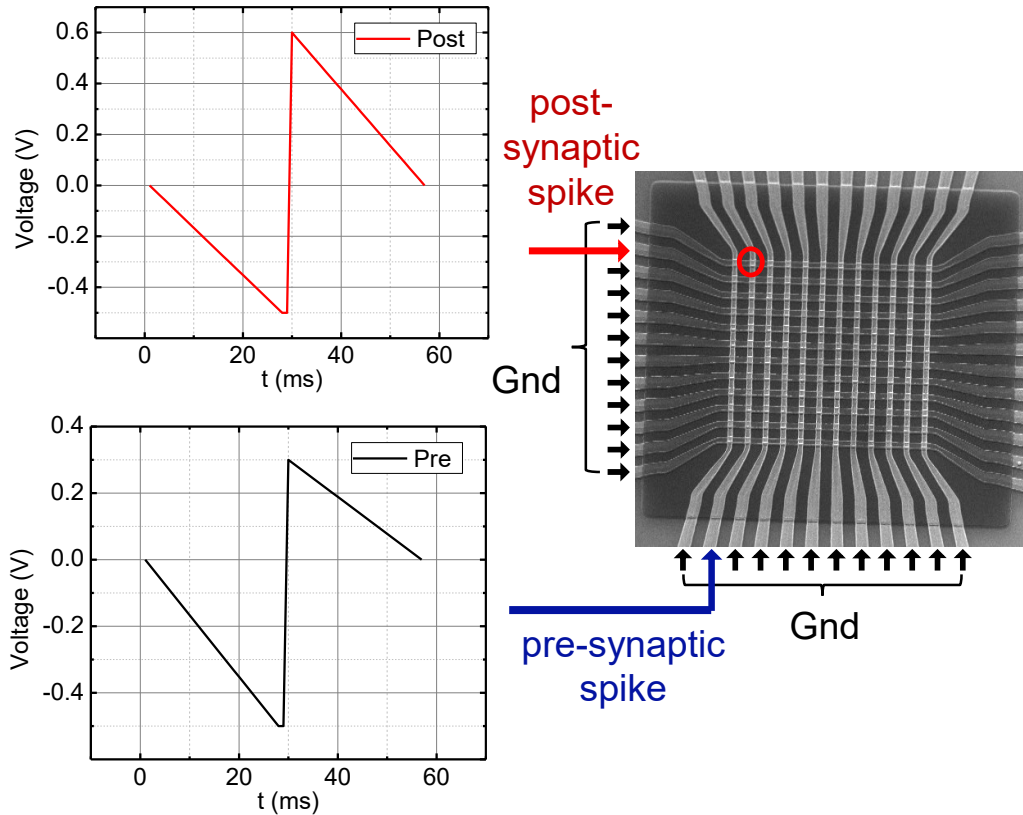
180 nm  
64 K synapses  
1024 neurons (12b)

- Experimental systems (e.g., to explore new algorithms or perform brain simulations)
- Loihi: digital synapses, not in-memory computing,  $\sim 25$  pJ/Sop



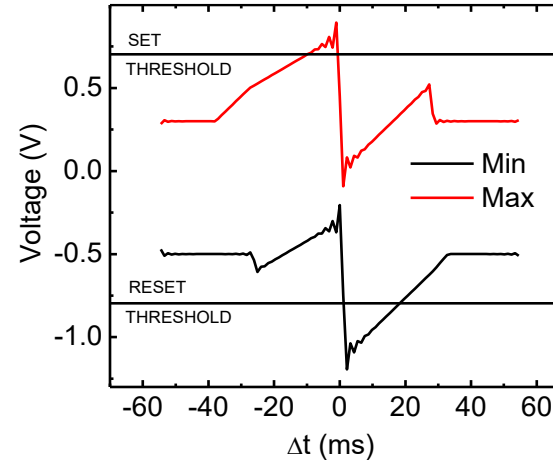
# Spike-Timing-Dependent Plasticity with Memristors

## Applied pulses

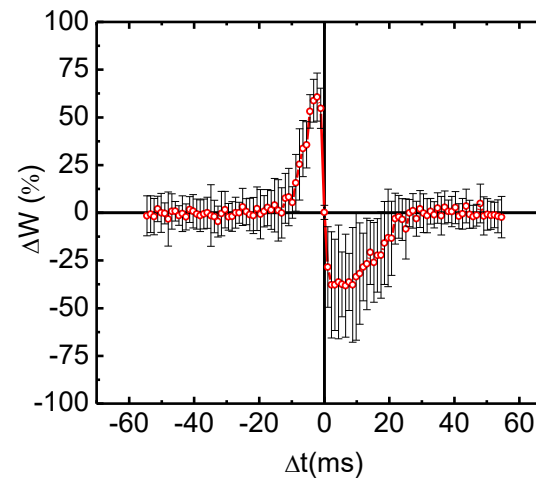


M. Prezioso et al., Nature Scientific Reports 2016

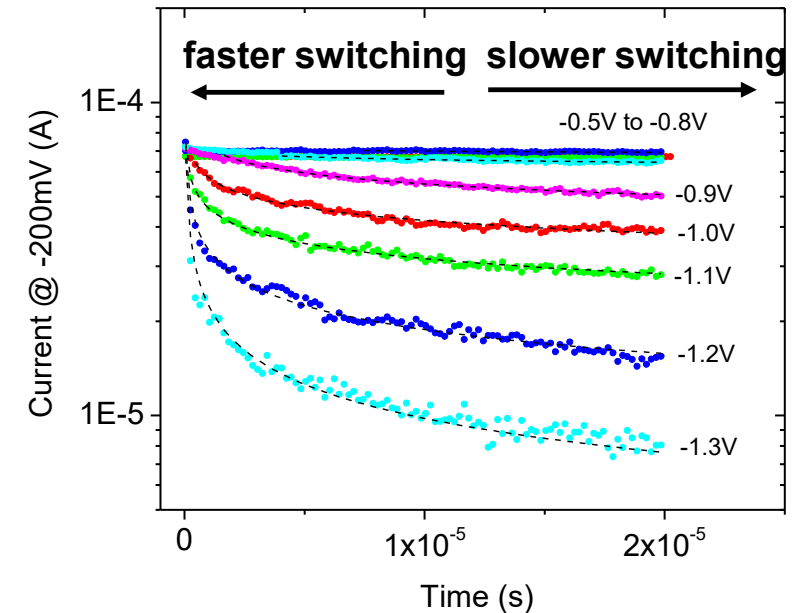
## Voltage on memristor



## Measured STDP



## Typical memristor switching dynamics



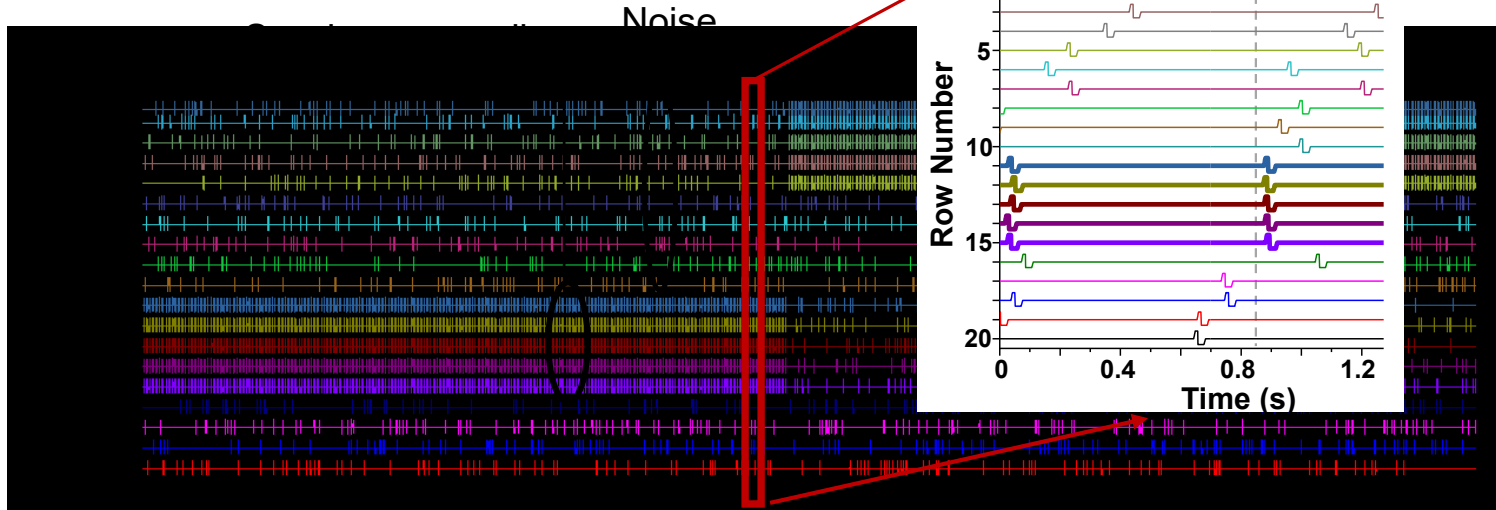
F. Alibart et al., Nanotechnology 2012

## Summary:

- Reproducible, clean STDP due to the utilized quasi-saturated switching regime (pulse duration  $\sim 60$  ms)
- Impact of initial memory state on STDP and fitted model

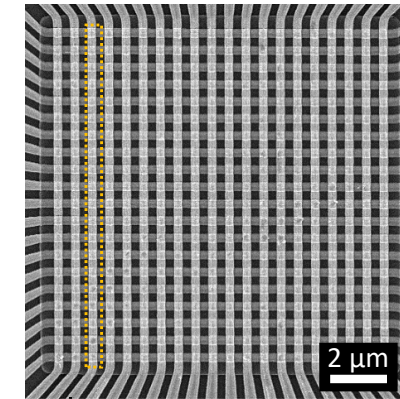
# Coincidence Detection Learning by STDP

## Applied spiking input

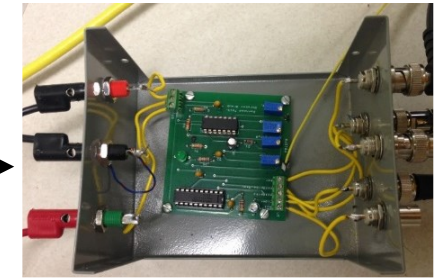


M. Prezioso et al. Nature Comm (2018)

20x20 crossbar circuit

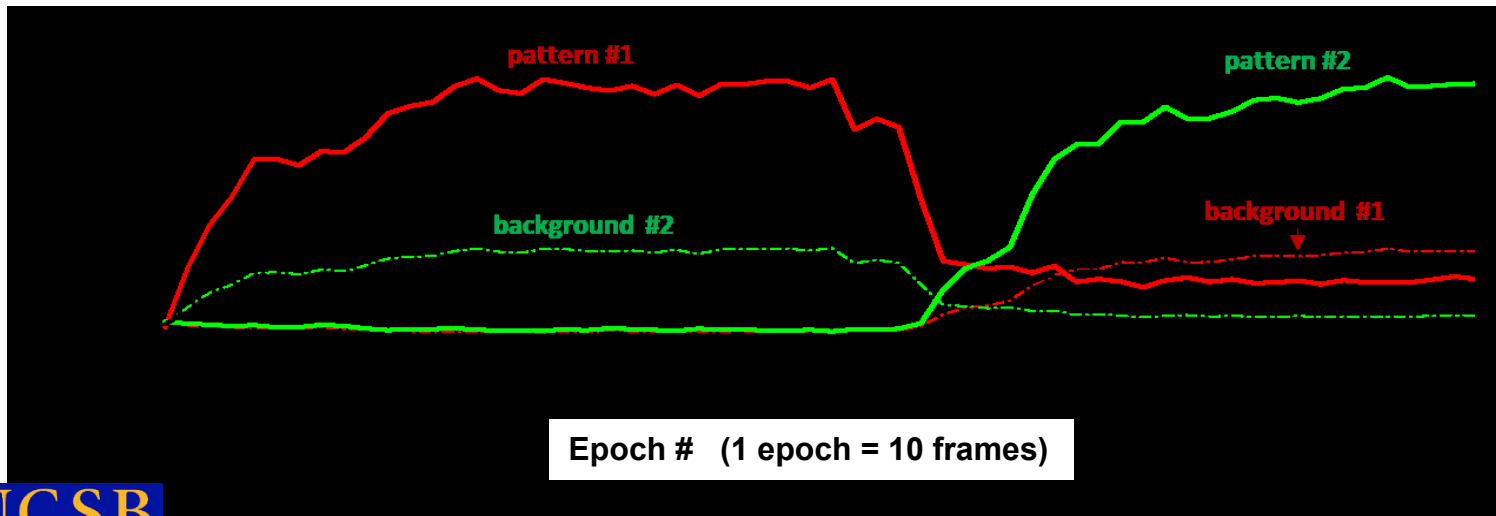


CMOS IC leaky-integrate-and-fire neuron



backward spike

## Weight evolution (learning via STDP)

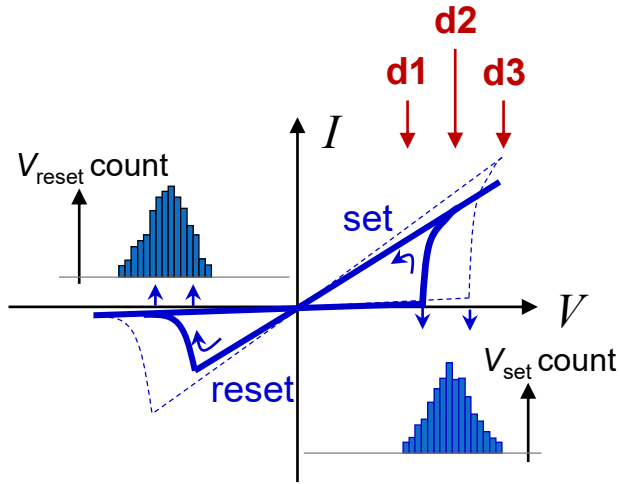


## Summary:

- Unsupervised learning for spike coincidence (i.e. of spiking on synchronous input), which is a fundamental low-level brain operation
- Weights of synchronous inputs are gradually potentiated, while all other depressed via STDP mechanism
- Higher noise density compared to prior work

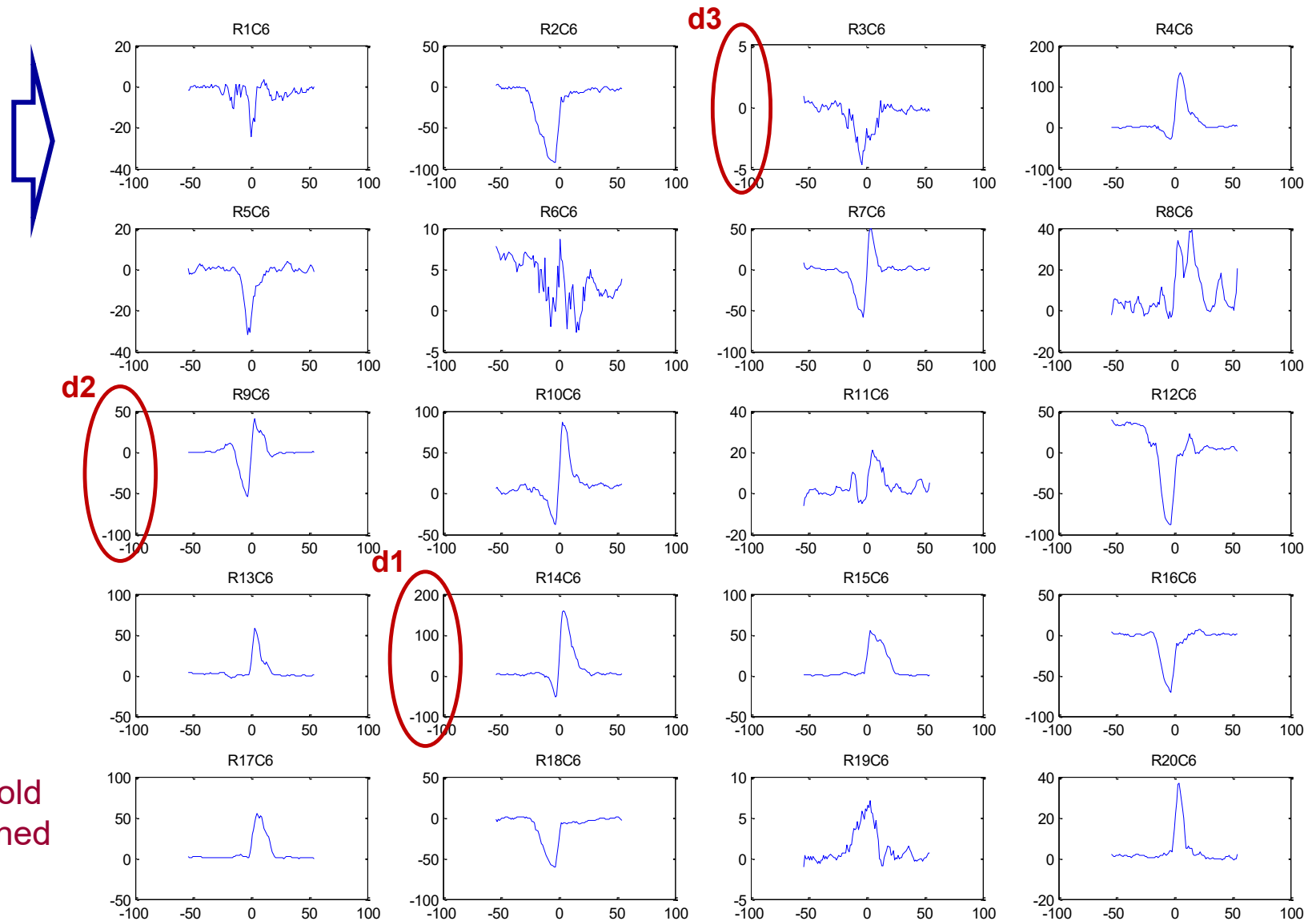
# Challenges for Memristor-Based Spiking Hardware

- STDP for 20 different memristors (sharing column 6<sup>th</sup>), collected using fixed-amplitude pre- and post-synaptic pulses



Stricter device requirements (passive or active cells):

- More severe impact of threshold variations than for ex-situ-trained systems
- Higher switching endurance



# Summary

- Neuromorphic inference with ex-situ training as natural (checkpoint) entry-level application of emerging nonvolatile memory analog neural networks
  - Near term: Embedded NOR floating gate memories (available at foundries now)
  - Long term: Metal oxide memristors (the most dense though least mature)
  - >10,000x in energy-delay over digital system based on experimental results for small-scale system, and system-level projections to bigger systems (e.g. NOR-flash aCortex) for ex-situ trained inference accelerators
- Straightforward extension to any inference (e.g. stochastic neural networks, neurooptimization, inference for spiking neural networks)
- Experimental demonstration of Boltzmann machines based on small-scale stochastic VMMs circuits with applications in deep believe networks and combinatorial optimization
  - Intrinsic noise of memory devices to implement stochastic transfer function or stochastic vector-by-matrix multiplication
- Experimental demonstration of coincidence pattern detection with STDP learning in spiking neural networks
- Major memristor challenges: poor yield, device uniformity, high cell currents
- Much more demanding device uniformity requirements for training accelerators or real-time learning (e.g. SNN with STDP learning)

# Acknowledgments

Current members at my research group at UC Santa Barbara:



Tinish Bhattacharya



Nikita Buzov



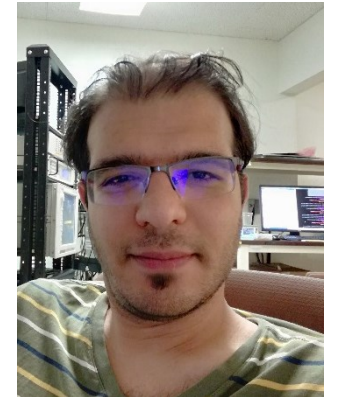
Zahra Fahimi



Hae Jin Kim



Shabnam Larimian



M. Reza Mahmoodi

Former group members and collaborators: H. Kim, H. Nili, V. Polishchuk, V. Sedov, M. Prezioso, F. Merrikh Bayat, G. Adam, B. Hoskins, X. Guo, N. Do, B. Charkabarti, M. Klachko, K. Likharev

Sponsors (past and present):

