High-Performance Stochastic Memristive Networks for Neurocomputing and Neurooptimization

Dmitri Strukov

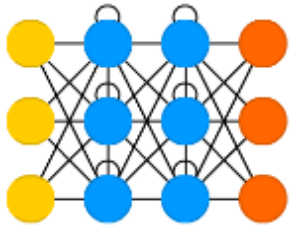
UC Santa Barbara

PRiME

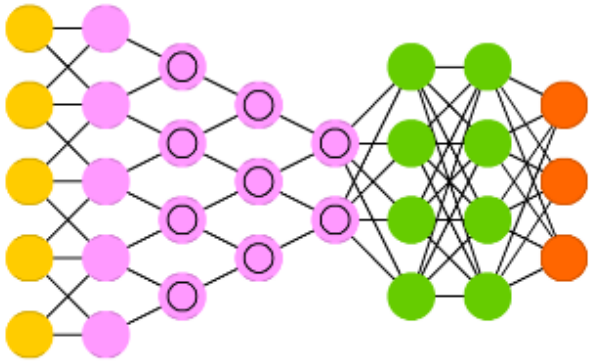
October 2020 (virtual)

Artificial Neural Network Zoo

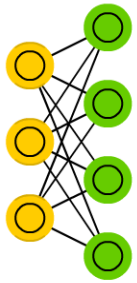
Recurrent Network (RNN)



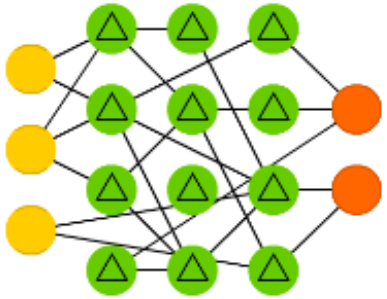
Deep Convolutional Network (CNN)



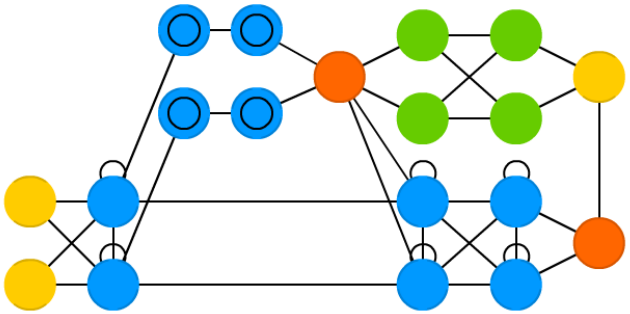
Restricted BM (RBM)



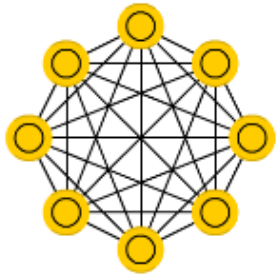
Liquid State Machine (LSM)



Attention Network (AN)

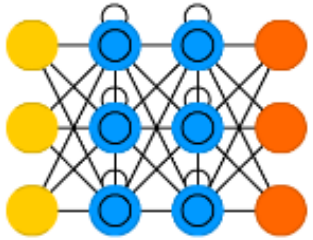


Hopfield Network (HN)

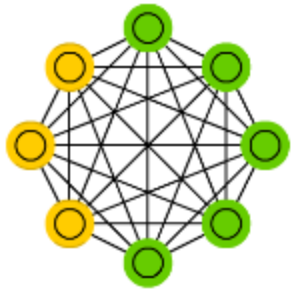


Multilayer Perceptron (MLP)

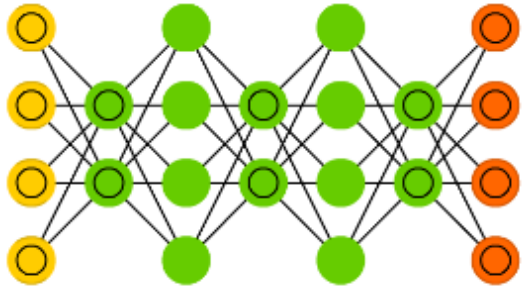
Long / Short Term Memory (LSTM)



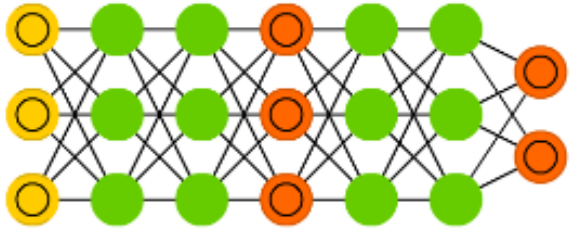
Boltzmann Machine (BM)

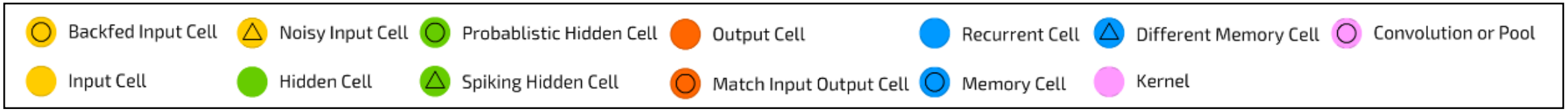


Deep Belief Network (DBN)

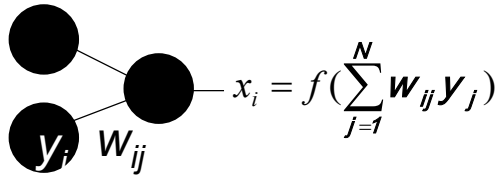


Generative Adversarial Network (GAN)





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



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

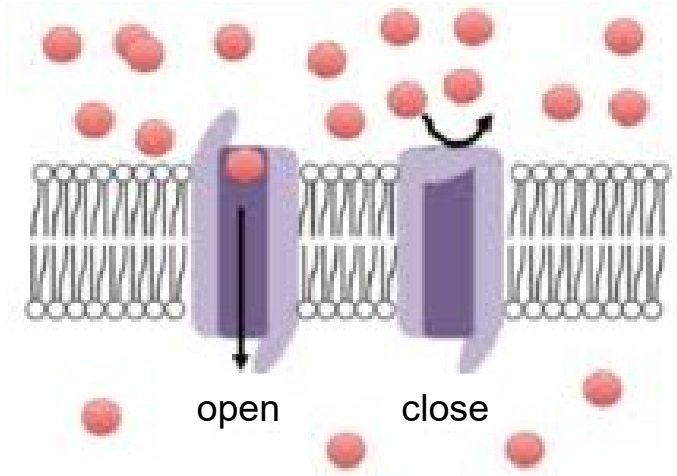
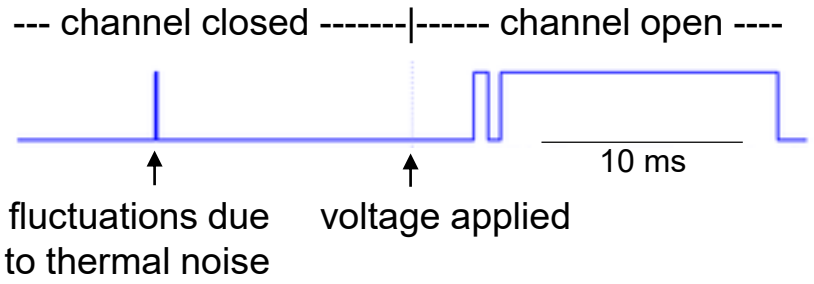
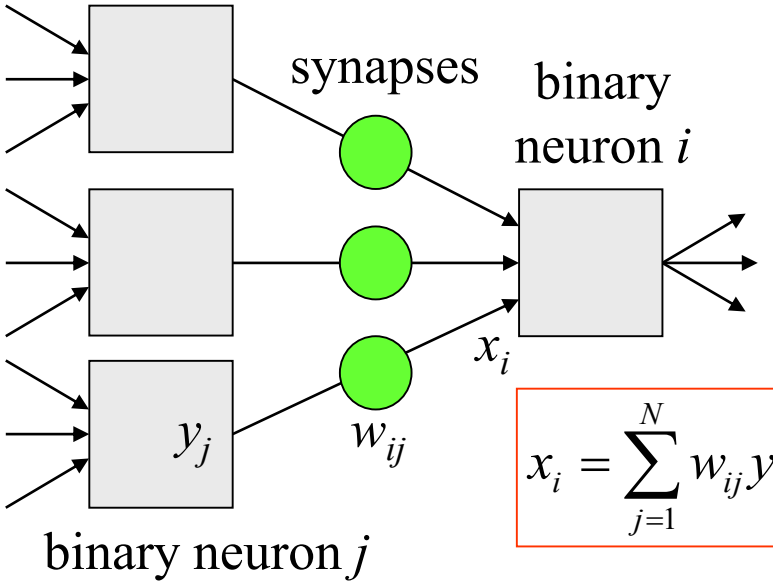


Image sources: Scholarpedia

Example: fluctuations in K channel



Stochastic (binary) neuron



$$p(y=1) = \frac{1}{1+e^{-x/T}}$$

$$x_i = \sum_{j=1}^{N} w_{ij} y_j$$

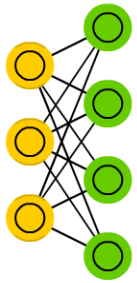
need efficient hw for dot-product and stochastic transfer function

Stochastic neural networks:

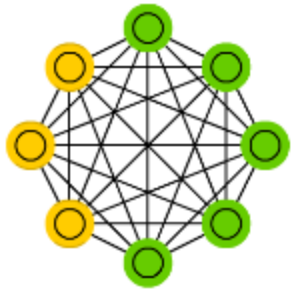
- (Restricted) Boltzmann machines
- Stochastic Hopfield networks
- Deep believe networks
- Bayesian networks
- ...

Focus of This Talk

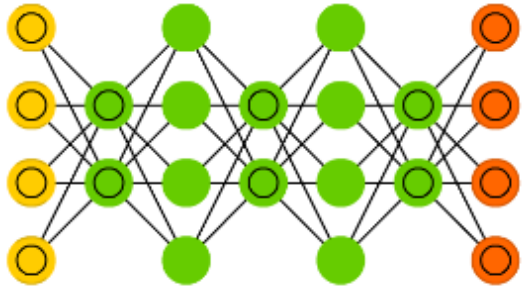
Restricted BM (RBM)

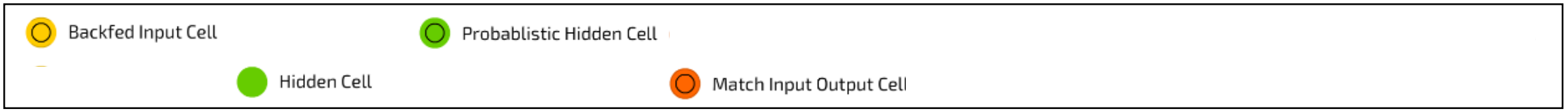


Boltzmann Machine (BM)

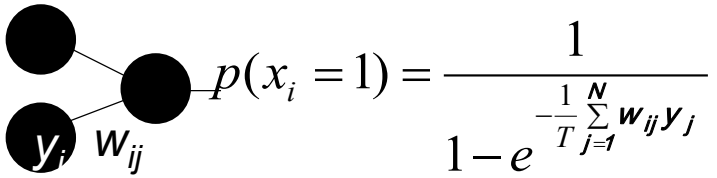


Deep Belief Network (DBN)





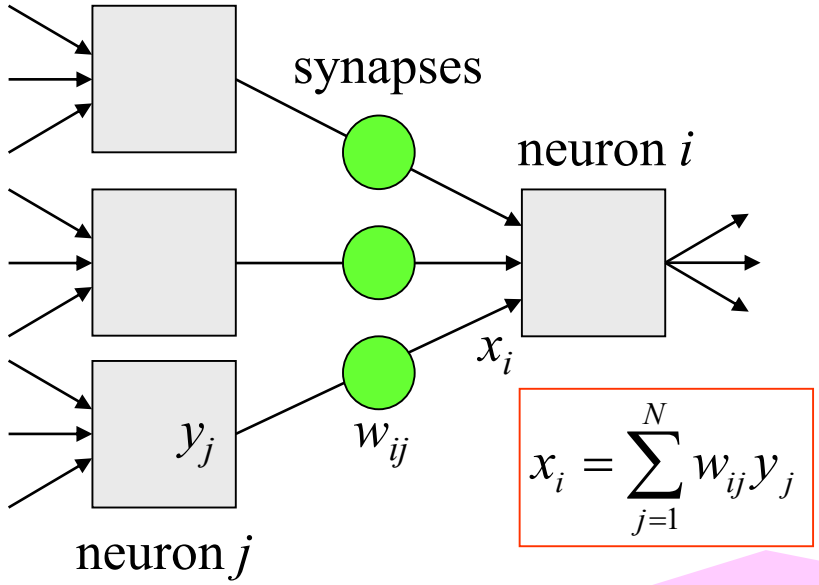
Stochastic vector-by-matrix multiplication is the **most common operation**



Radical Improvement with Analog Computing

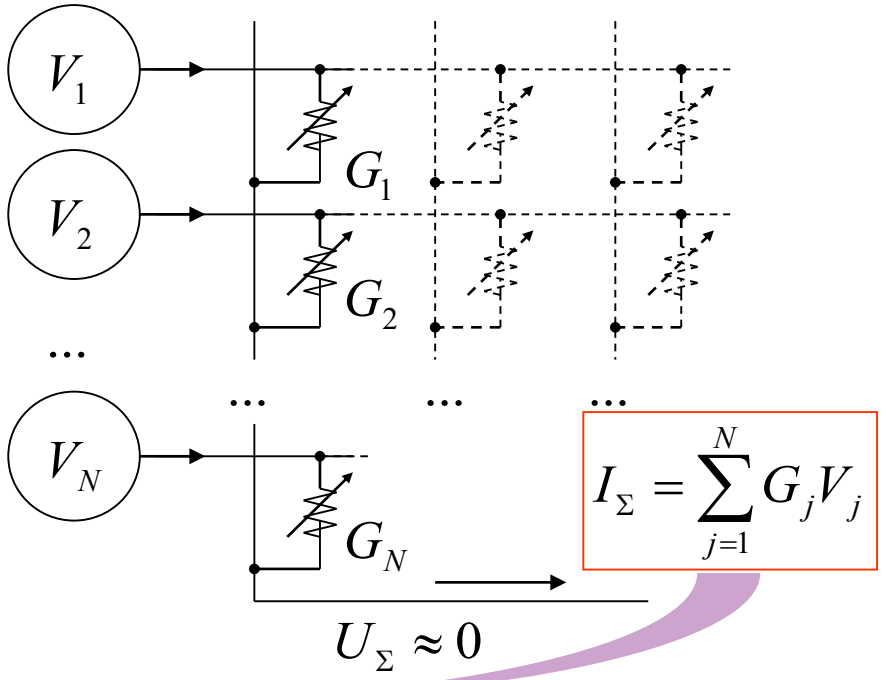
Vector-by-Matrix-Multiplication (VMM):

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

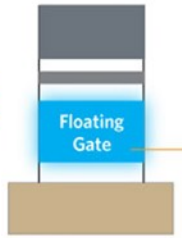
would allow to fit extra-large models on chip!

Maturity

Active "1T"

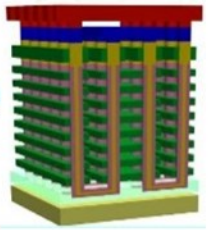
40 F^2 now, <20 F^2 with FinFET

2D NOR



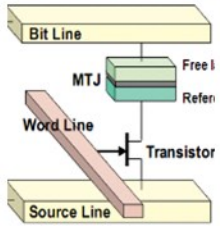
3DNAND

1÷10 TB/in^2

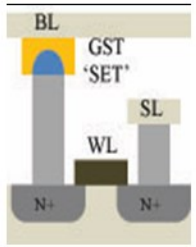


Active "1T1R"

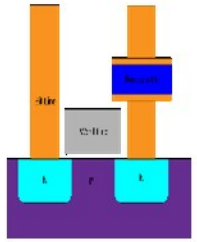
STTRAM



PCRAM

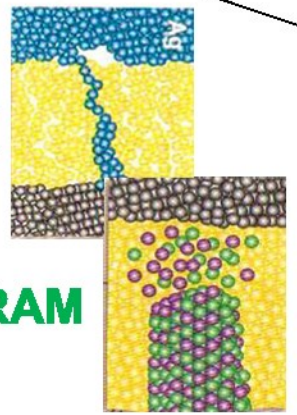


2D FeRAM



CBRAM

ReRAM



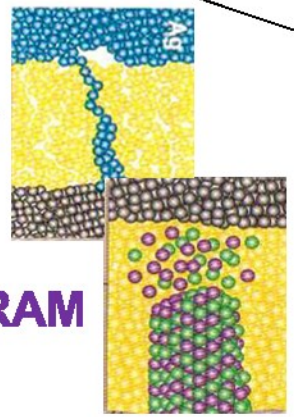
few 100's F^2 for current, potentially down to 25 F^2

Passive "0T1R"

CBRAM

~4F^2

2D RRAM



<< 4F^2

3D FeRAM?

3D RRAM

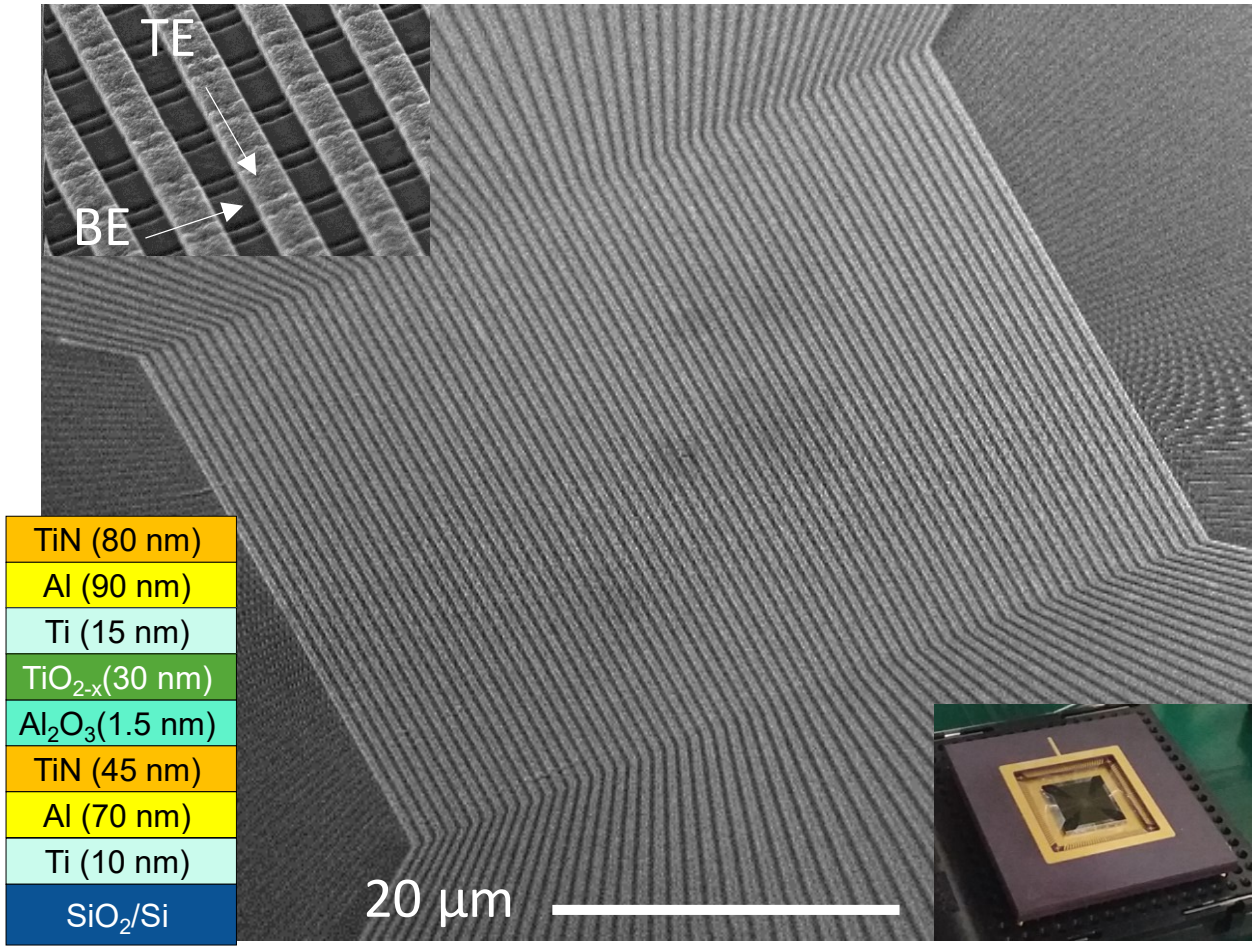
F = feature size

Most important specs: Density, retention, analog switching!

Cell density

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

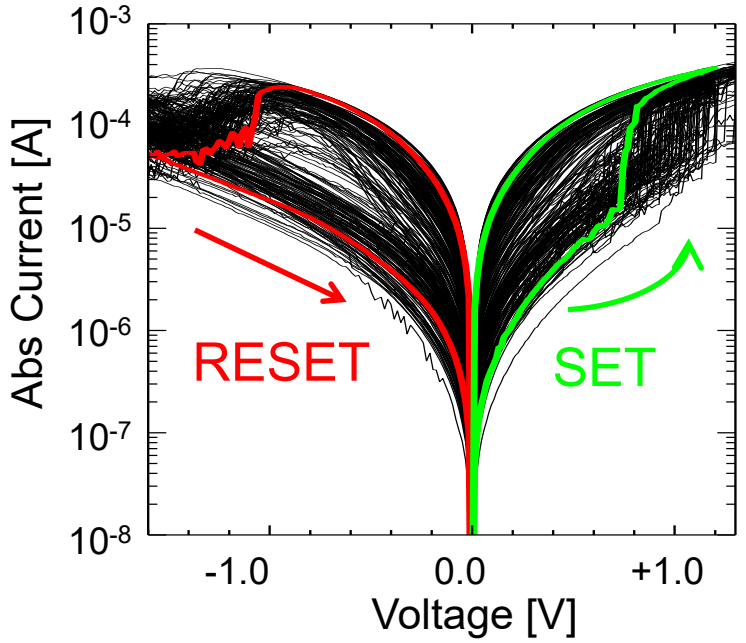
- **64 × 64 passive crossbar circuit**



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

- **Typical I-V characteristics**

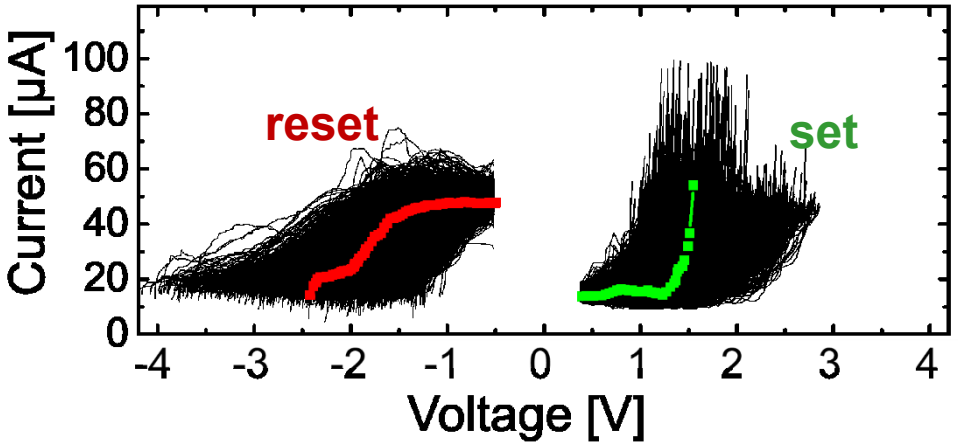


Details:

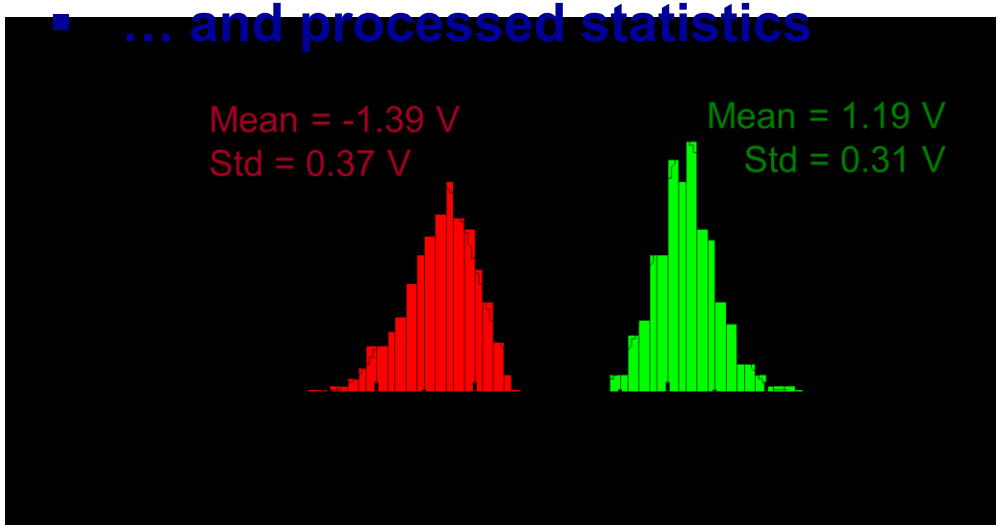
- Al_2O_3/TiO_{2-x} active bilayer by reactive sputtering
- CMOS-compatible CMP/dry etching process and TiN/Al electrodes for higher conductance
- ~250 nm wide lines, passive (0T1R) integration (e.g. >250x/10,000x better memristor / memory cell density compared to 1T1R work at comparable complexity and yield
- The largest functional analog-grade passive memristor crossbar circuit supported by proper statistics

Most Important Metric: Yield and Switching Threshold Variations in 64×64 Xbar

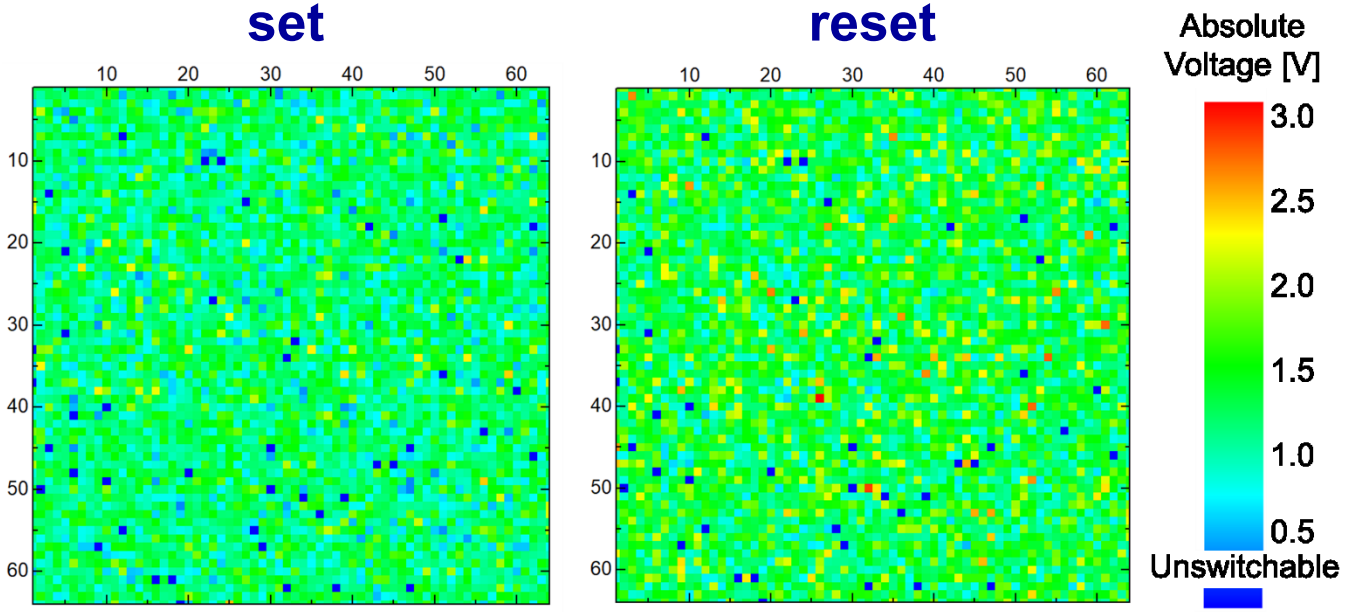
- Raw data (voltage ramp) …



- … and processed statistics



- Switching threshold spatial map:

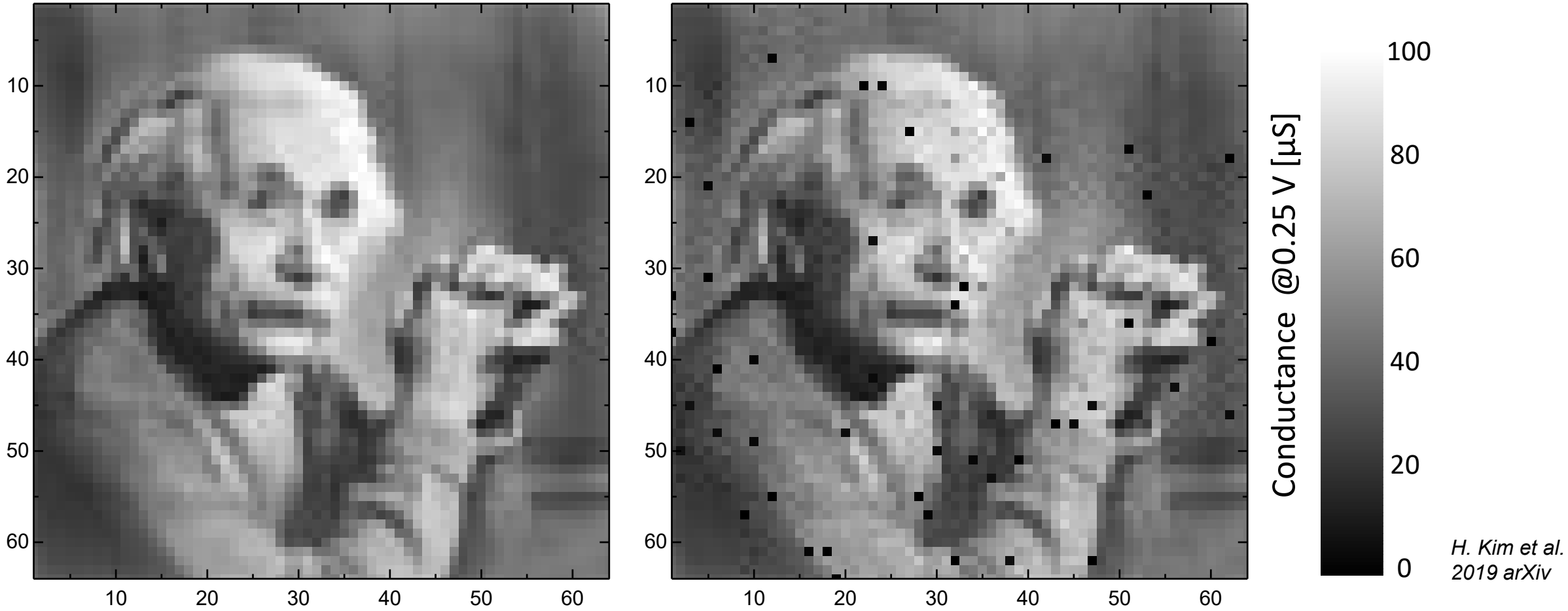


- Switching threshold is defined as voltage at which current changes by > 10% when applying voltage ramp
- Dark blue dots: ~1% devices that cannot be switched

H. Kim et al. 2019 arXiv

Conductance Tuning in 64×64 Memristor Crossbar

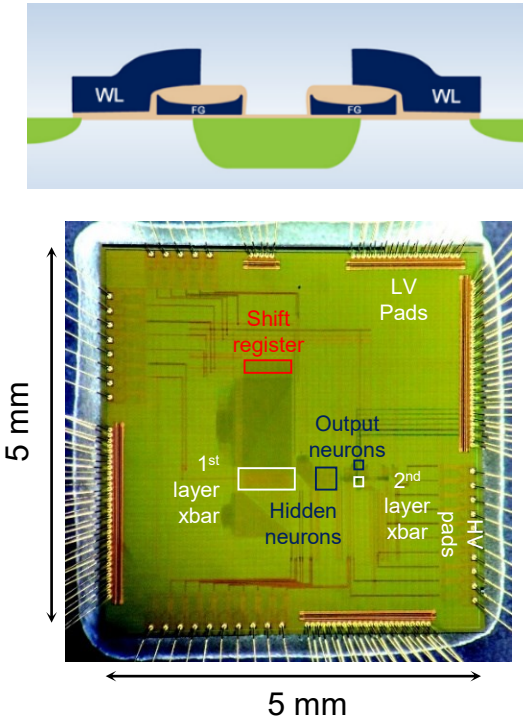
- **Desired pattern**
- **Actual pattern**



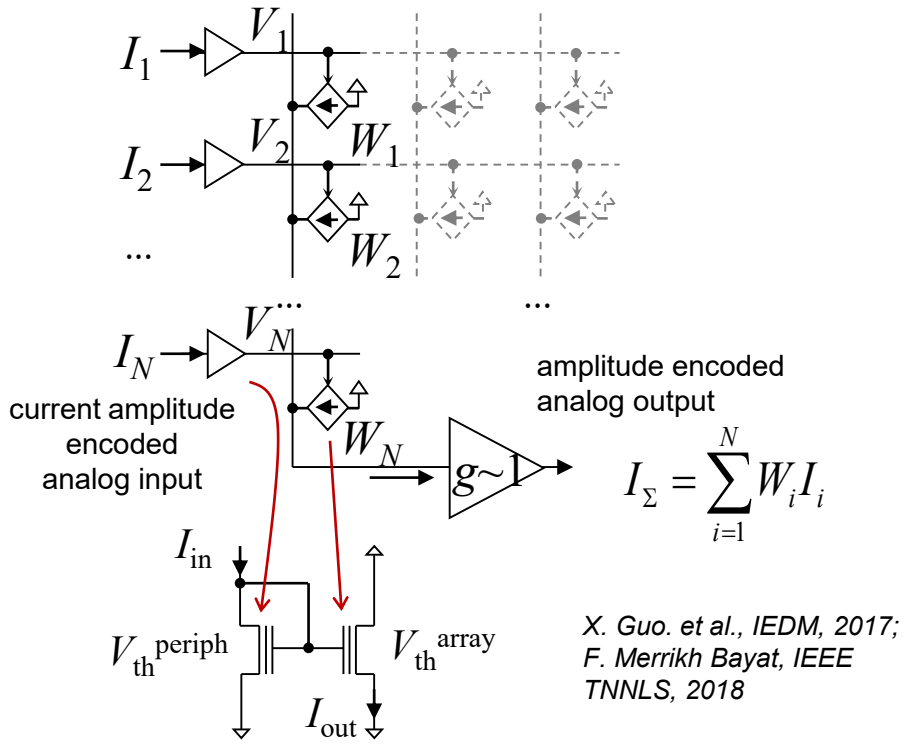
- Color encoding: 256 levels from white (10 μS) to black (100 μS) @ 0.2V
- < 5% / < 3% absolute / relative tuning error using automated algorithm, with reserves for improvement

Near-Term Option: Floating-Gate Devices

- **NOR eFlash device& chip**

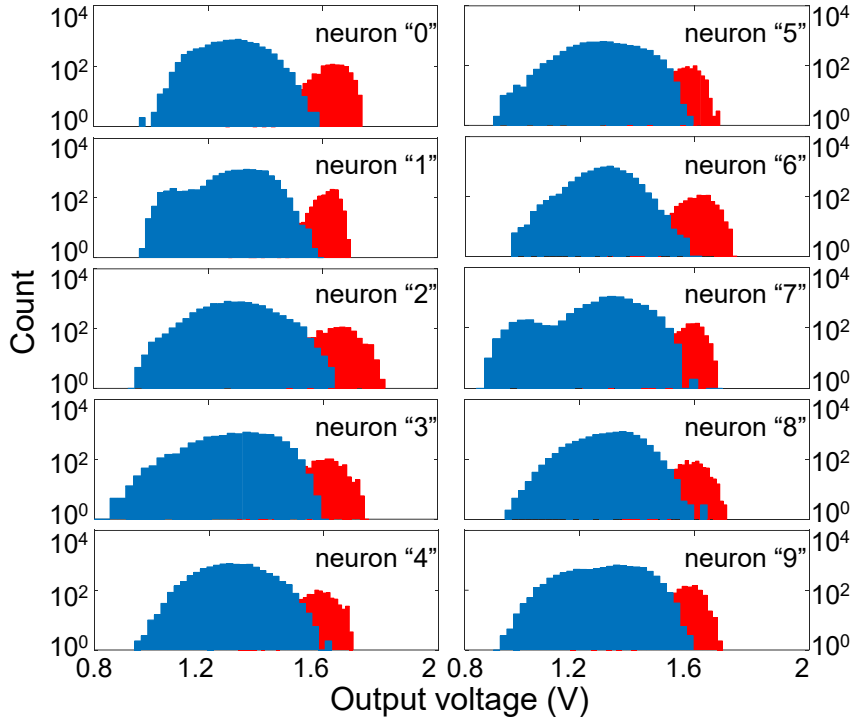


- **Vector-by-Matrix Multiplier Circuit**



X. Guo. et al., IEDM, 2017; F. Merrikh Bayat, IEEE TNNLS, 2018

- **2-layer MLP classification results**
  (10,000 MNIST test patterns)



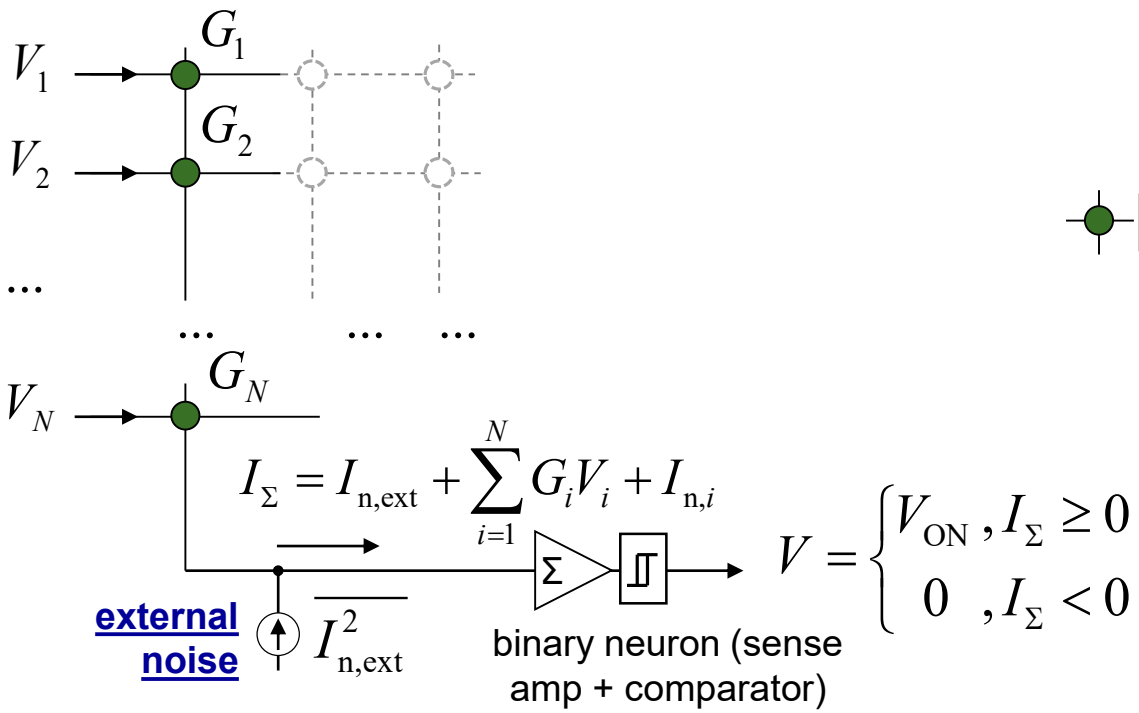
Summary:

- 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-μs 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

New Result #1: Stochastic Analog Vector-by-Matrix Multiplier

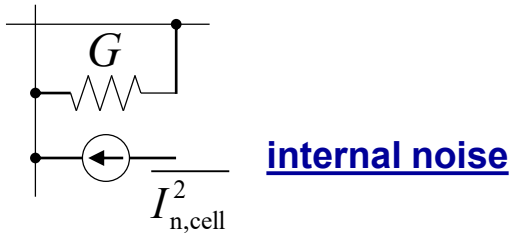
Basic Idea:

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

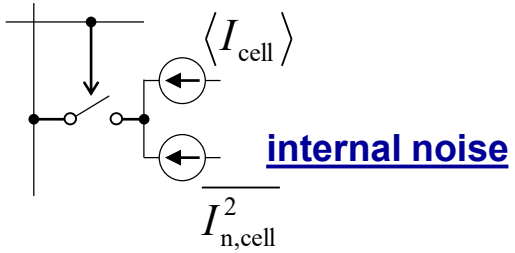


Two Implementation Options:

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

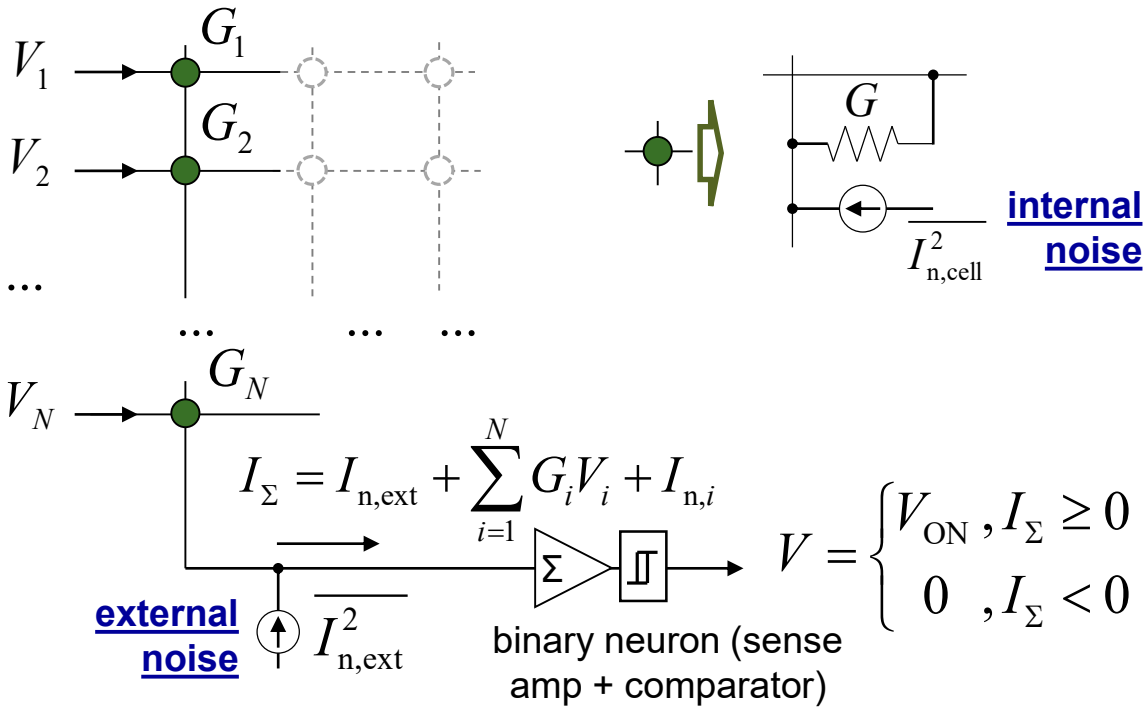


Floating gate transistor

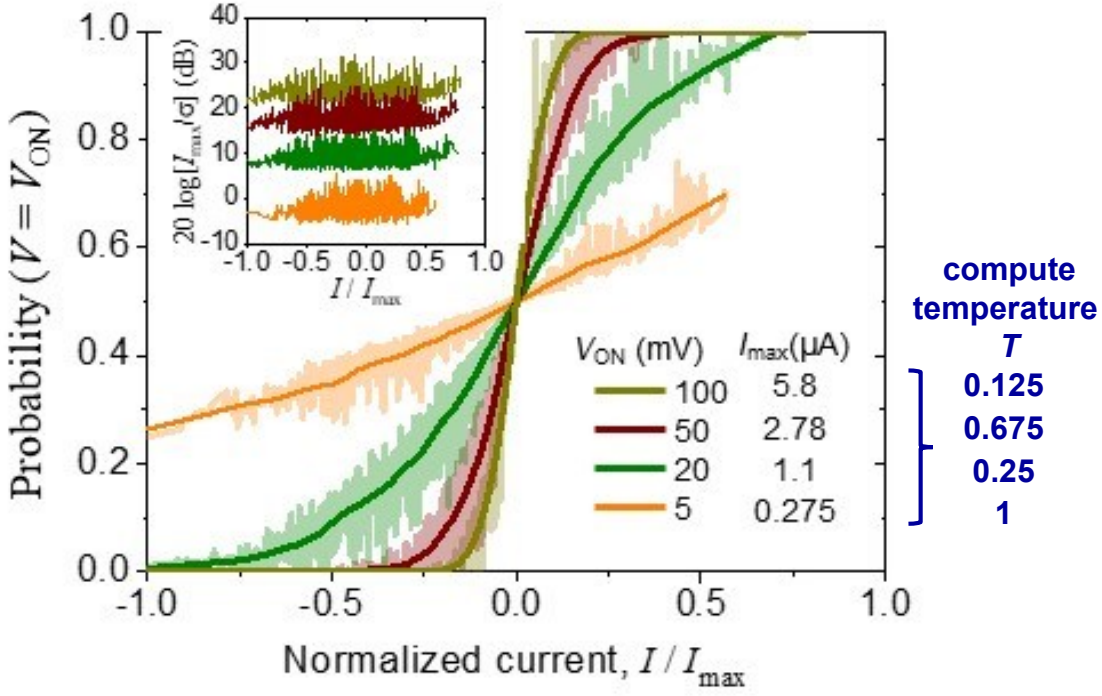


New Result #1: Stochastic Analog Vector-by-Matrix Multiplier

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



Experimental Demo:
using 20×20 passive array with externally-injected noise from readout circuitry



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

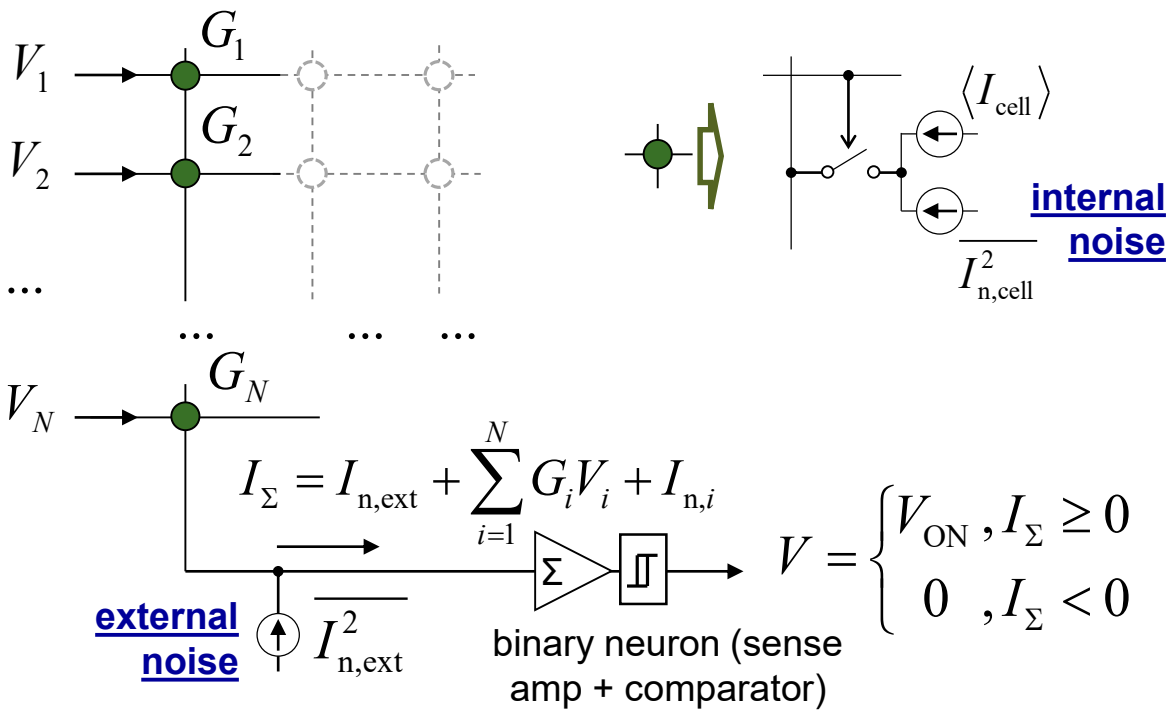
Features:

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

New Result #1: Stochastic Analog Vector-by-Matrix Multiplier

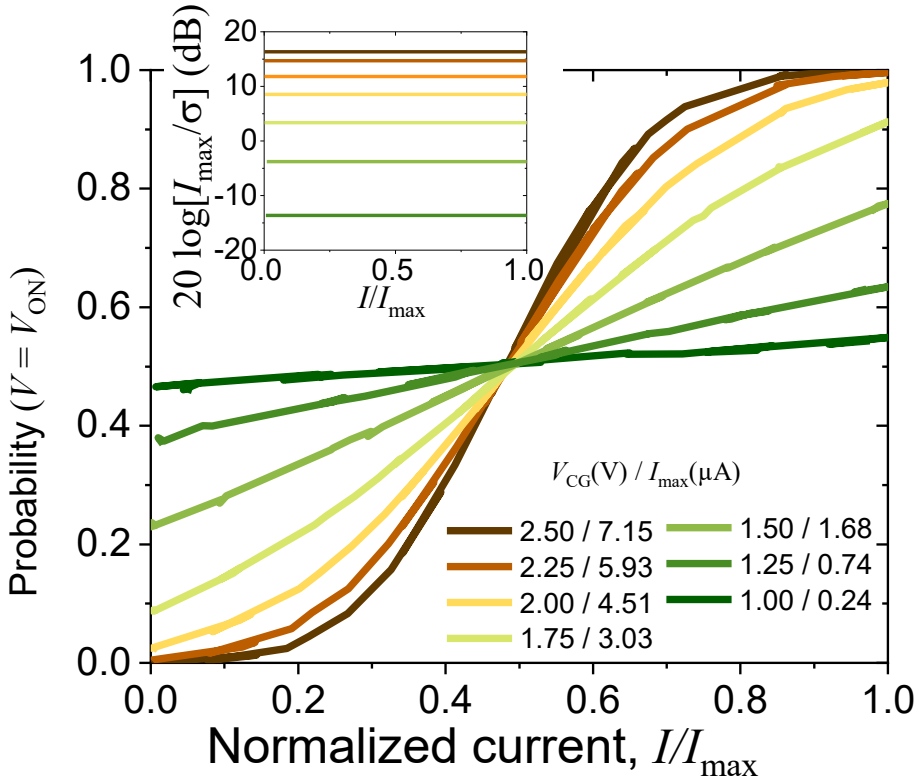
Basic Idea:

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



Experimental Demo:

using 180nm embedded ESF1 NOR-flash memory technology



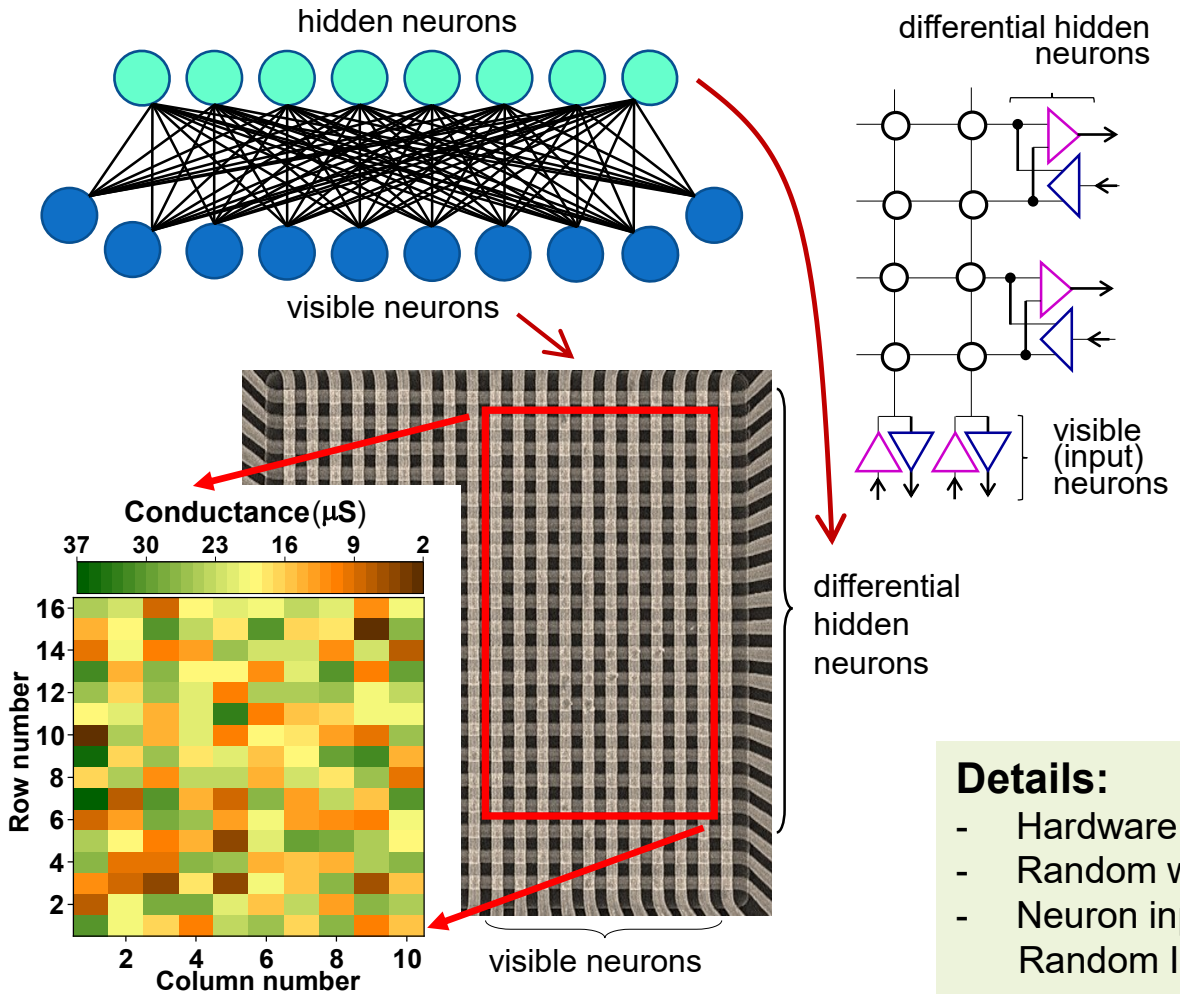
M.R. Mahmoodi et al. Nature Communications, 2019

Features:

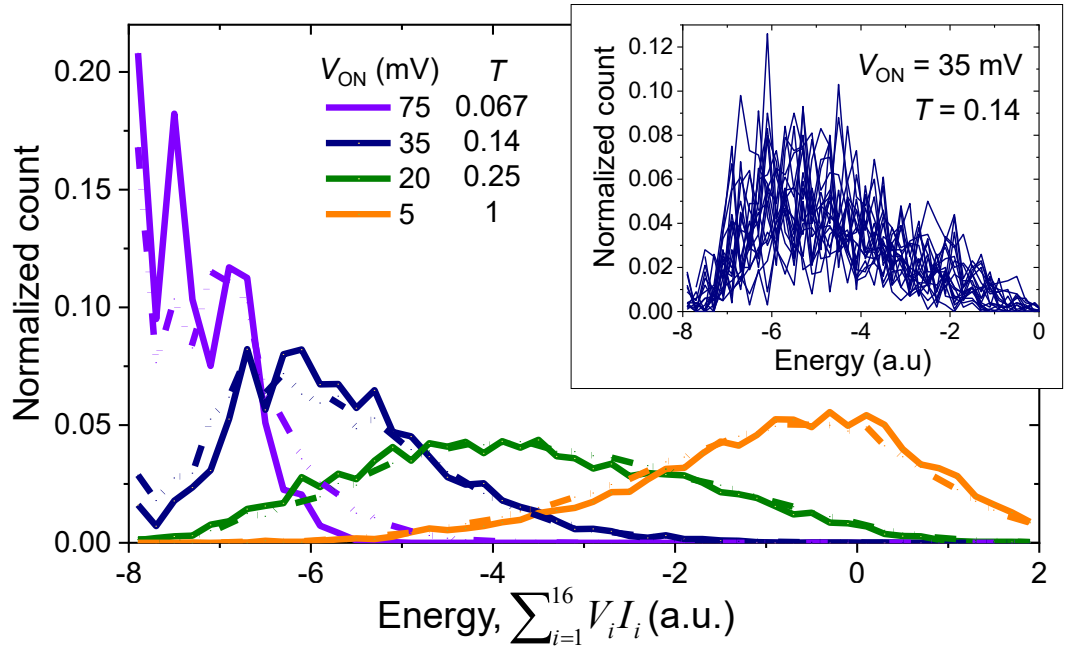
- Sigmoid slope (i.e. SNR or compute temperature *T*) controlled dynamically by the applied gate voltage

New Result #2: 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 μS to + 32 μS) mapped to 10×16 portion of memristor xbar
- Neuron input currents sampled at 1 MHz bandwidth after applying
  Random Input→ Visible → Hidden → Visible → Hidden → …

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

Solving Optimization Problems with Hopfield Neural Network

- **Combinatorial optimization problems**

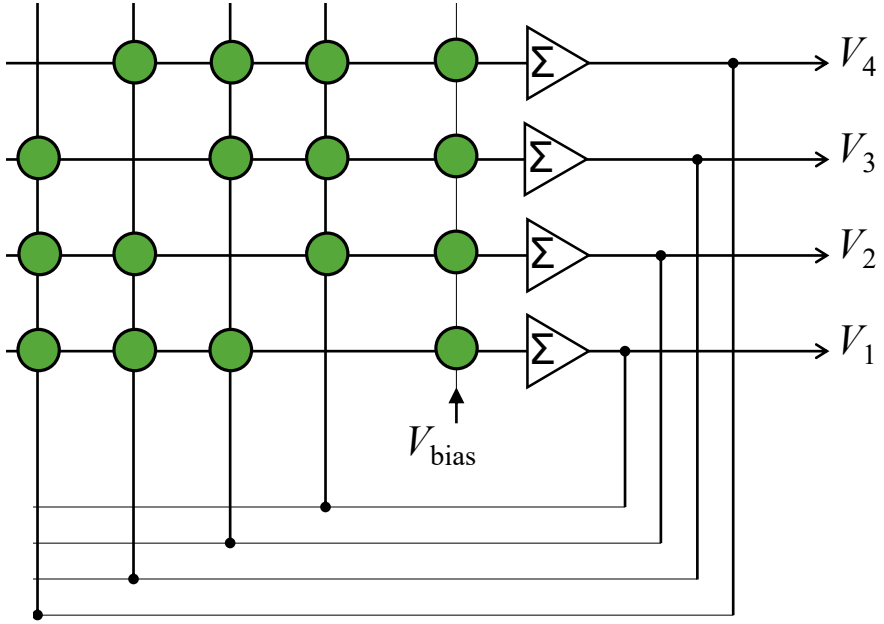
Application	Problem
Logistics / package delivery	Traveling salesman
Power grid	Maximum flow
Design automation	Vertex cover
Molecular dynamic simulations	Graph partitioning

- **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

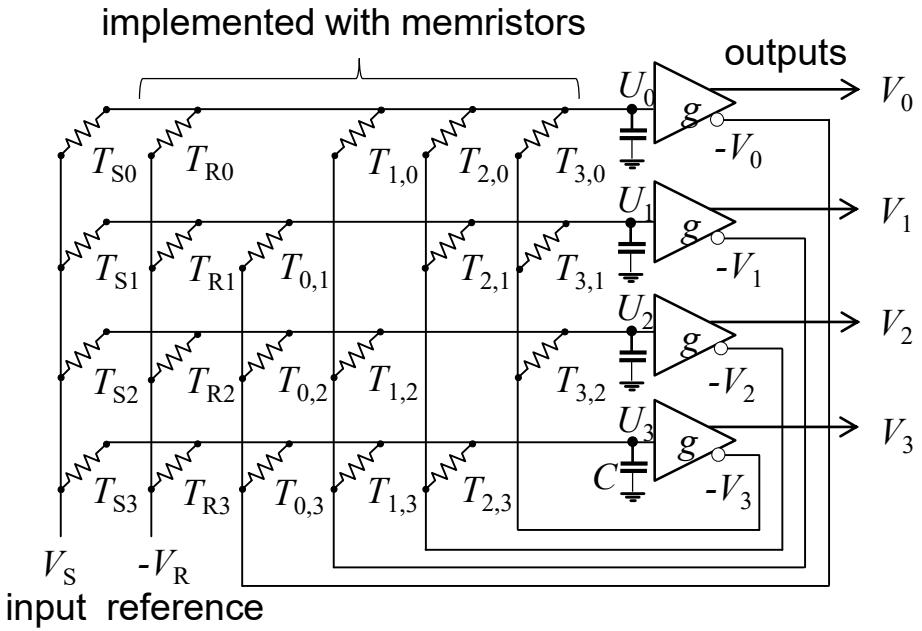
- **Example of continuous time / binary neuron Hopfield network**



∑ = sum amp & comparator

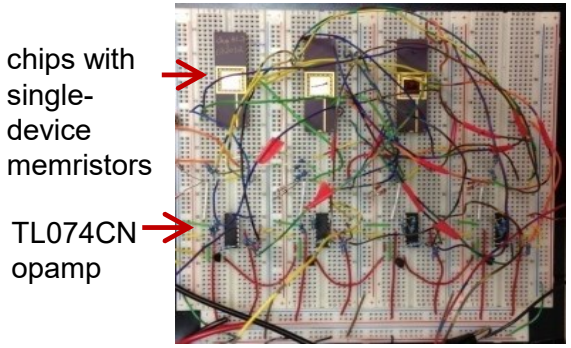
Earlier Work: (Deterministic) Hopfield Network Experimental Demonstration with Discrete Memristors

- **Hopfield network for A-to-D conversion**

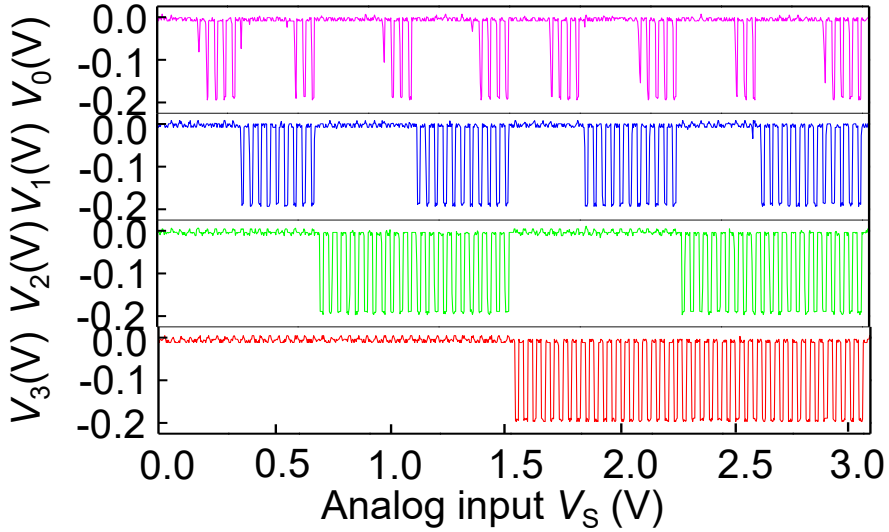


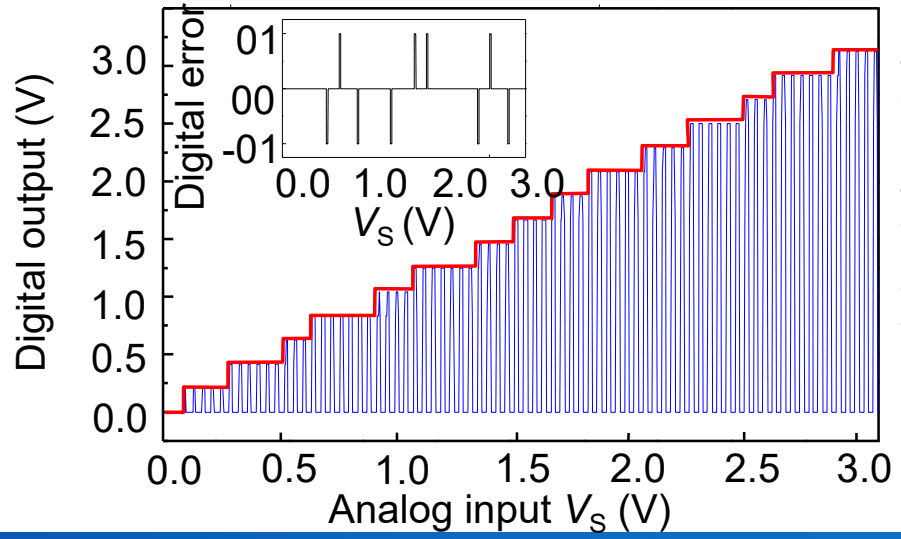
Major features:

- 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



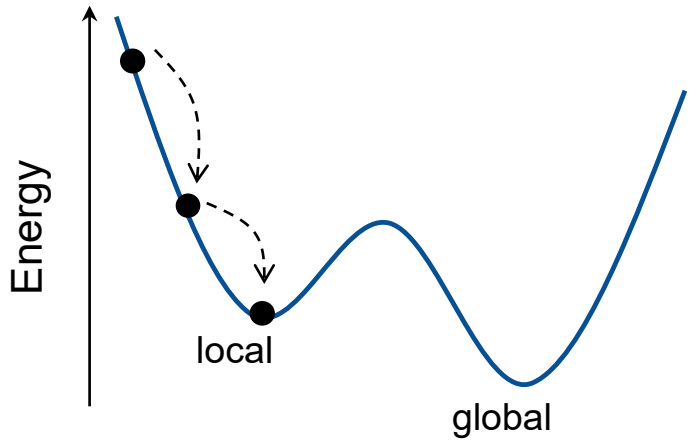
- **Experimental results**



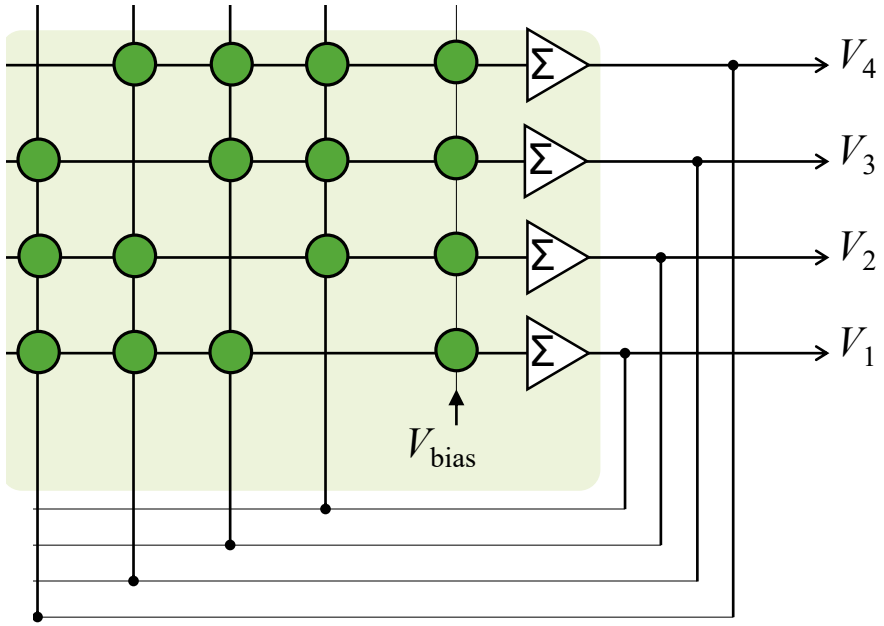


- L. Gao et al, in: *Proc. NanoArch' 13*, Ney York, NY July 2013;
- X. Guo et al., *Frontiers in Neuroscience* **9**, art. 488, Dec. 2015

Local Minima in Hopfield Network



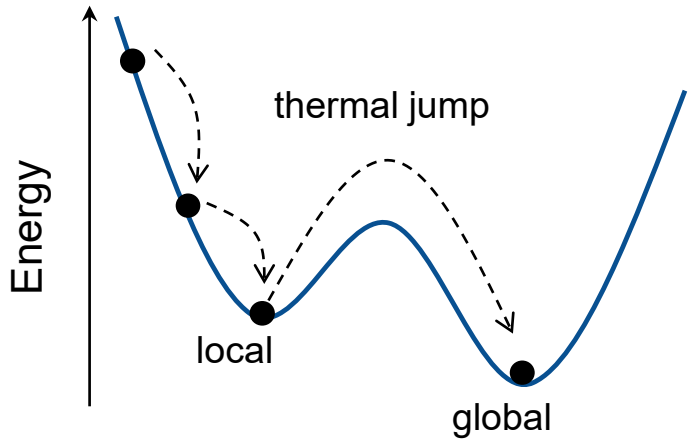
Local minima present problems!



Color background:
Baseline Hopfield neural network

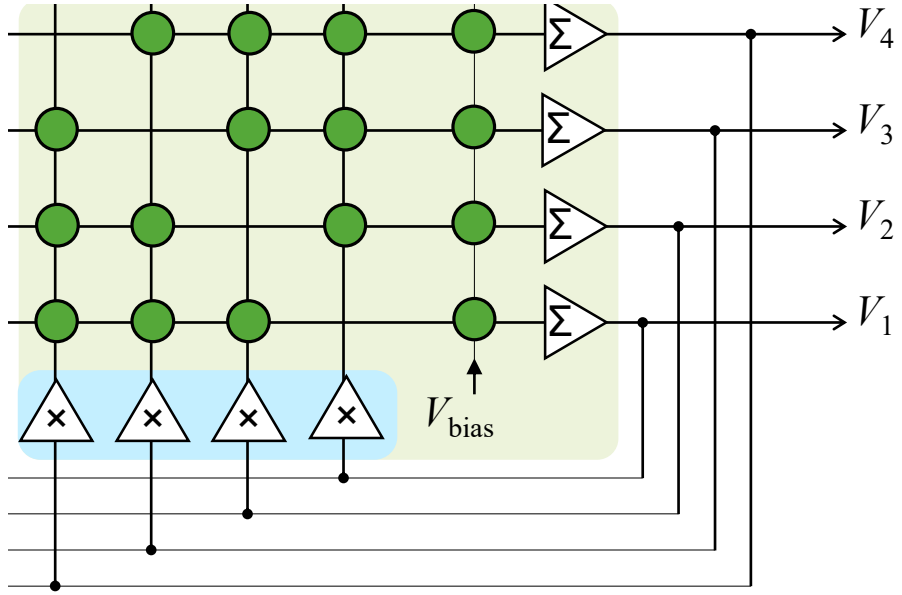
∑ = sum amp & comparator

Simulated Annealing with Generalized Hopfield Network (Boltzmann Machine)



Local minima present problems!

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

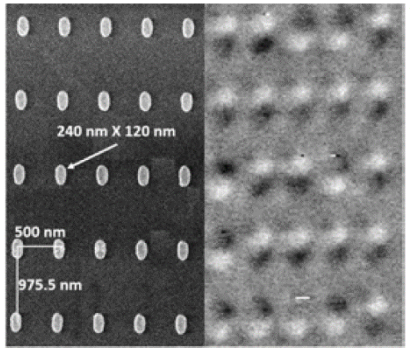


Color background:
Baseline Hopfield neural network
Stochastic annealing

∑ = 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 …**

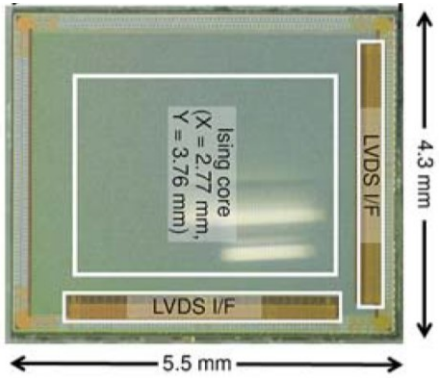
Debashis *et al. IEDM* 2016

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

- **… high CMOS overhead**

Borders *et al. Nature* 573 2019

CMOS

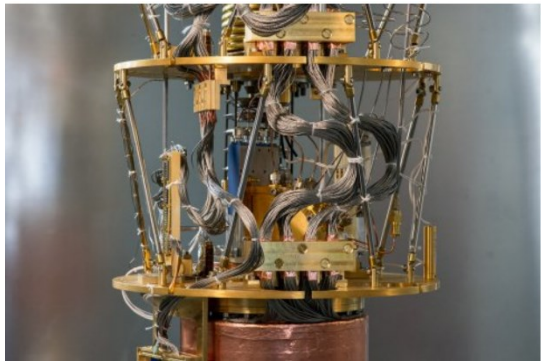


Experimental results for solving maximum-cut problem with 2×30K-spin Ising network 40-nm 23.65-mm^2 SRAM-based chips

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

Takemoto, *et al. ISSCC* 2018

Josephson Junction

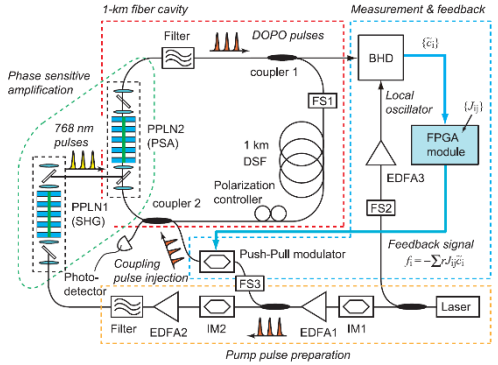


Experimentally measured ground state of random spin glass problems based on 108-qubit D-Wave One system (with evidence of quantum annealing)

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

Boixo, *et al. Nature phys.* 2018

Photonics



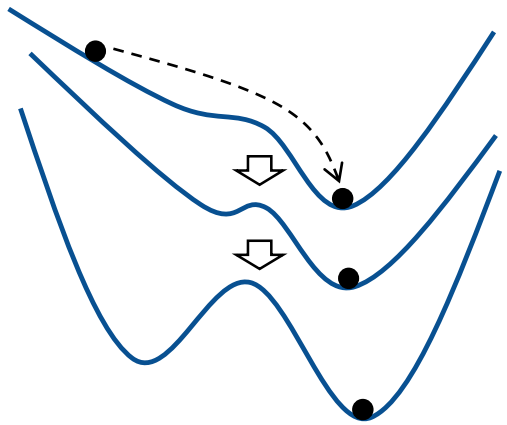
Experimental results for solving max-cut problems with up to 2,000 nodes with Ising network based on degenerate optical parametric oscillators

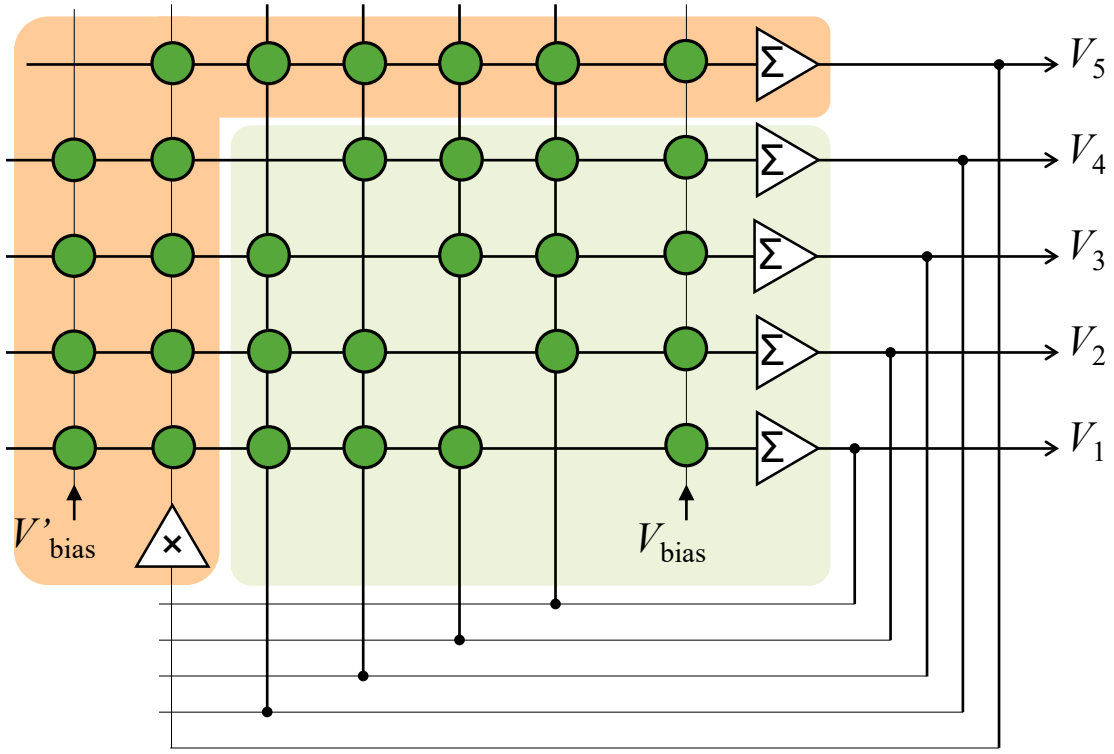
- **Slow due to high overhead of the electronic feedback used for updating spatial light modulator**

Inagaki, *et al. Science* 2016

Adjustable Energy Function Annealing

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





Another solution inspired by quantum annealers: Dynamically adjustable energy function

Color background:

Baseline Hopfield neural network

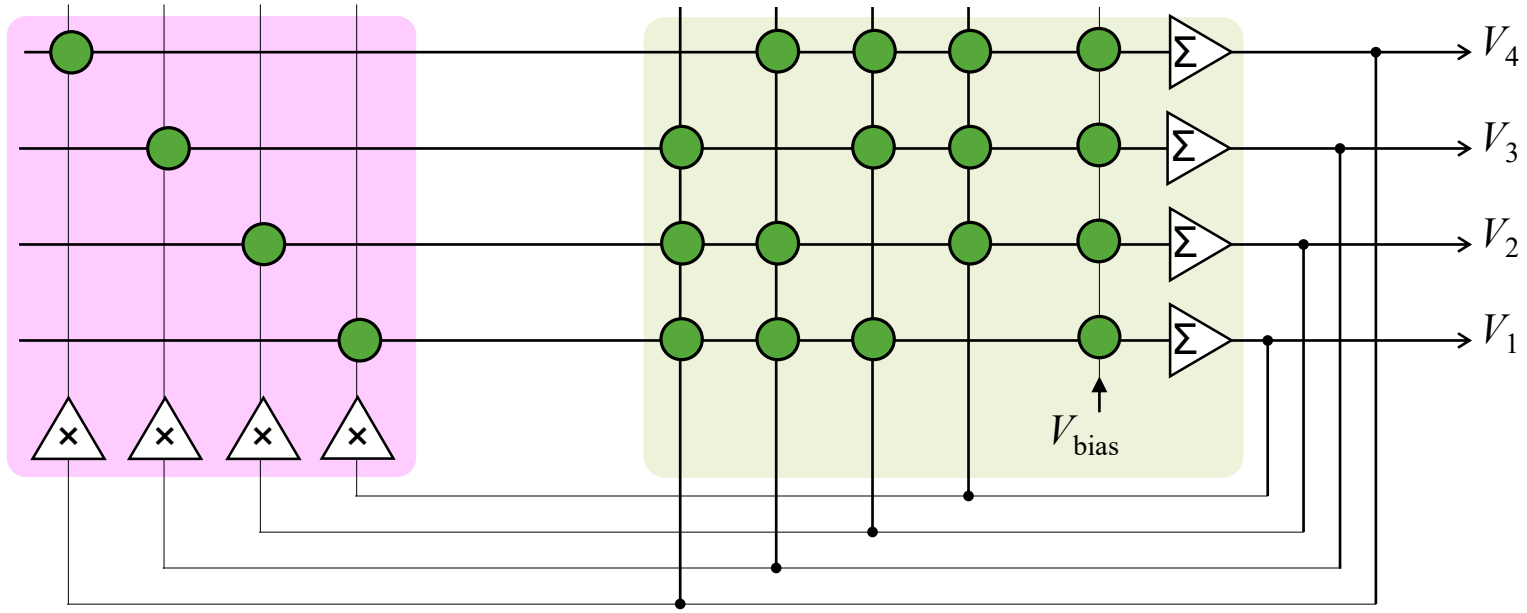
Adjustable energy function / weight annealing

∑ = sum amp & comparator

× = scaling

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

Yet Another Approach: Chaotic Annealing



Color background:

Baseline Hopfield neural network

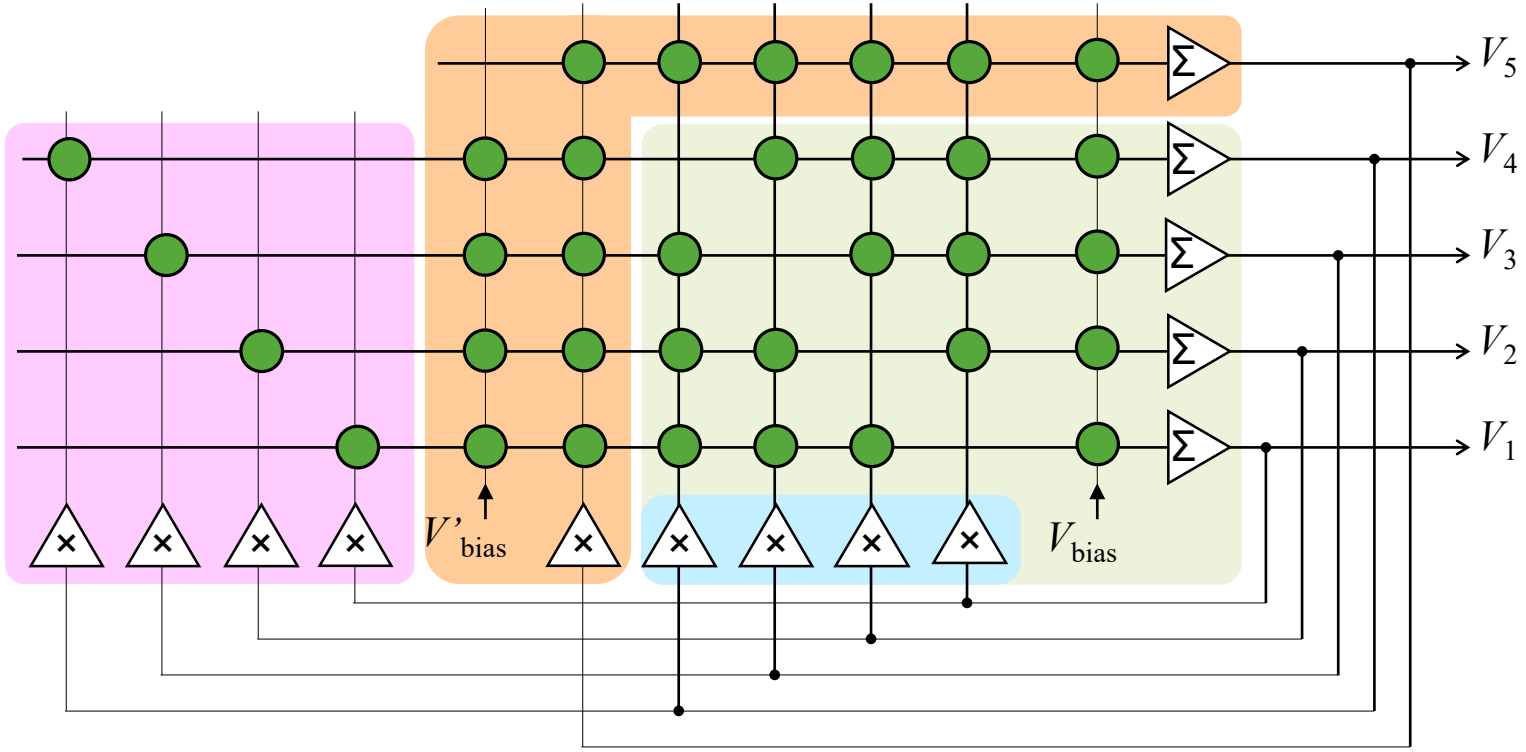
Chaotic annealing

∑ = sum amp & comparator

× = scaling

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

New Result #3: 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

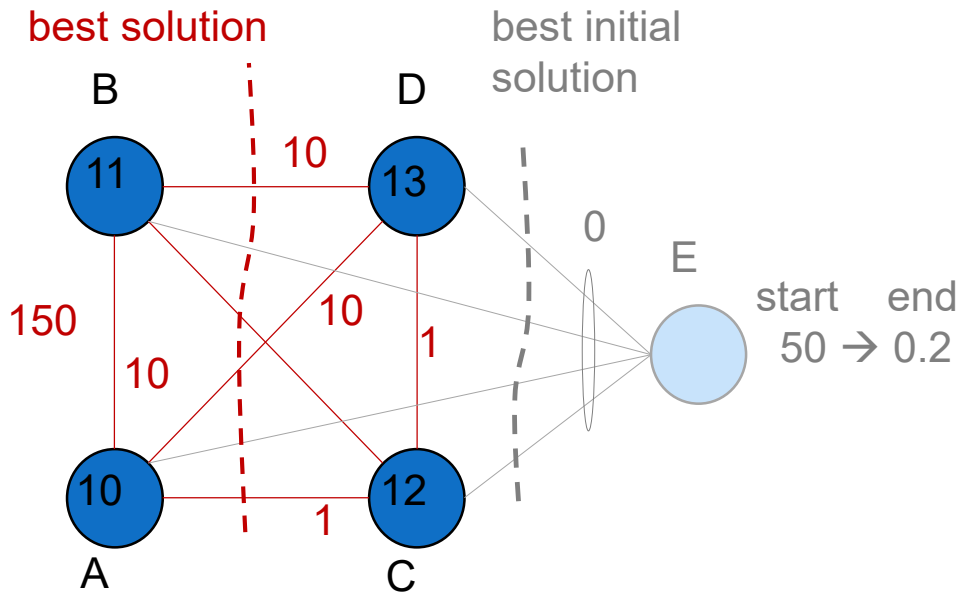
∑ = sum amp & comparator

× = scaling

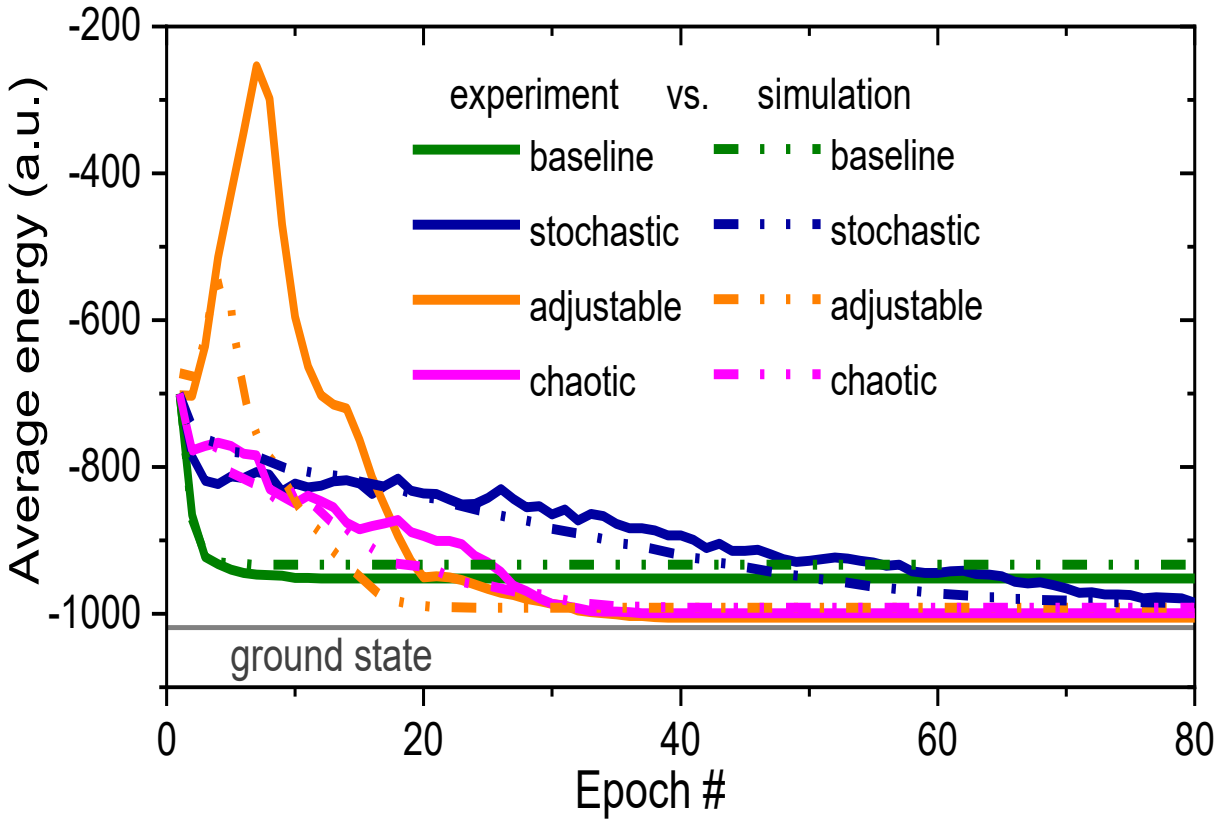
New Result #3: Combinatorial Optimization Demo with FG

- **Weighted graph partitioning problem...**

  (finding two mutually exclusive, set of nodes with maximally balanced node weights and minimized edge weights between two sets)



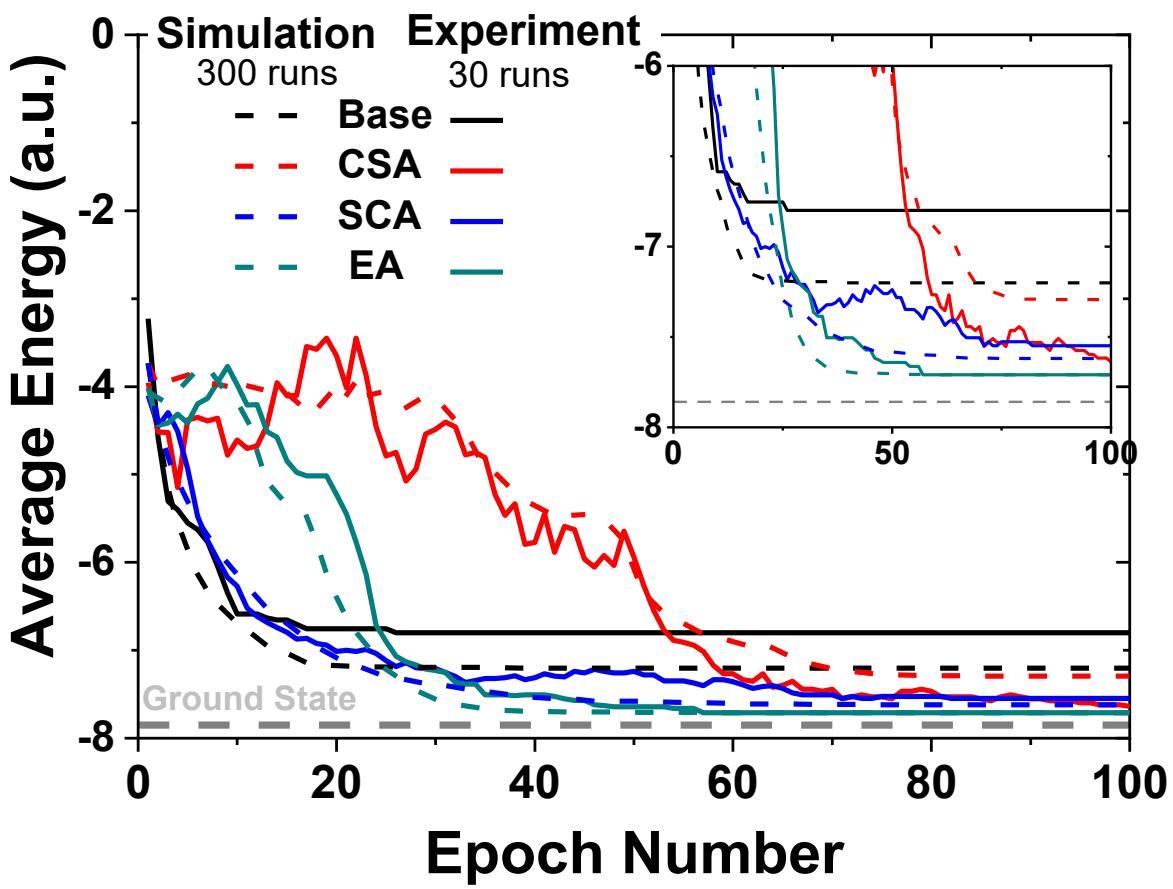
… and experimental results using 10×20 180-nm NOR flash memory array



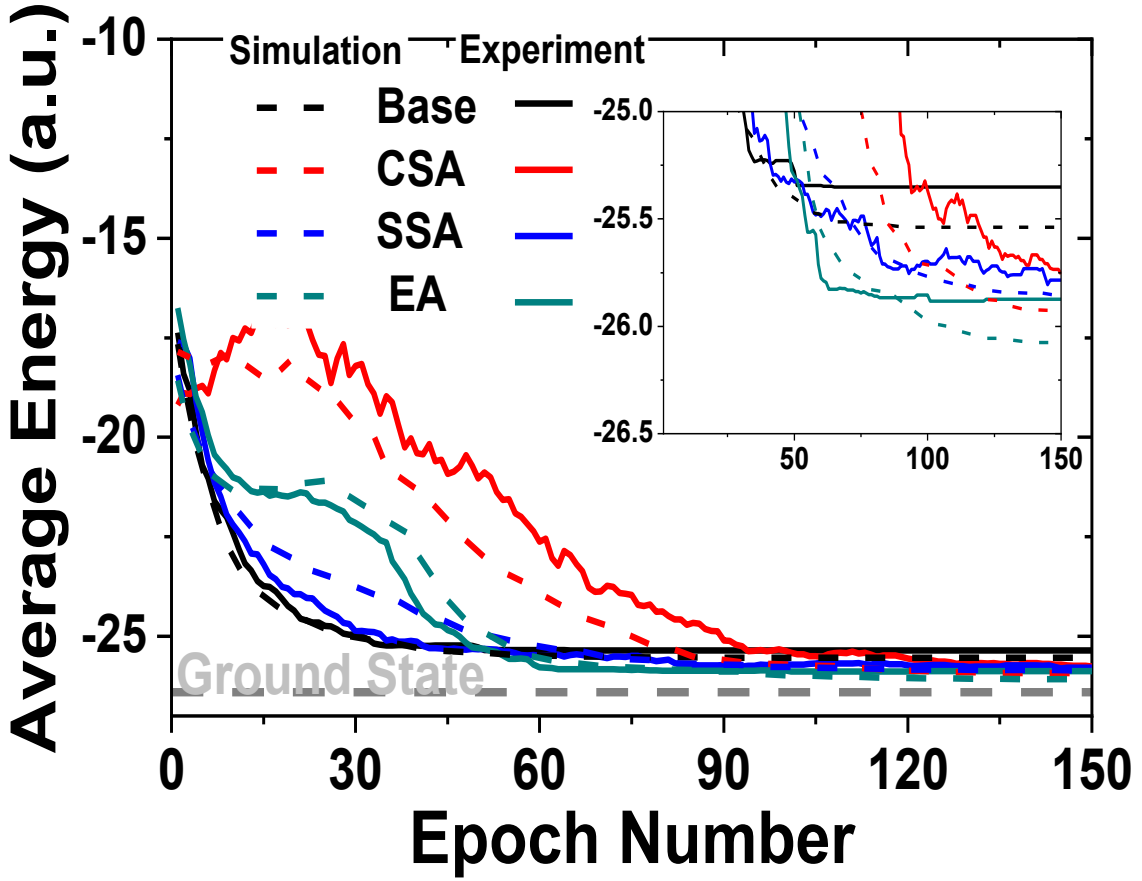
M.R. Mahmoodi et al. Nature Comm, 2019

New Result #3: Combinatorial Optimization Demo with Passive 64×64 Metal-Oxide Memristive Crosbar Circuits

- 5-node maximum-weighted clique problem



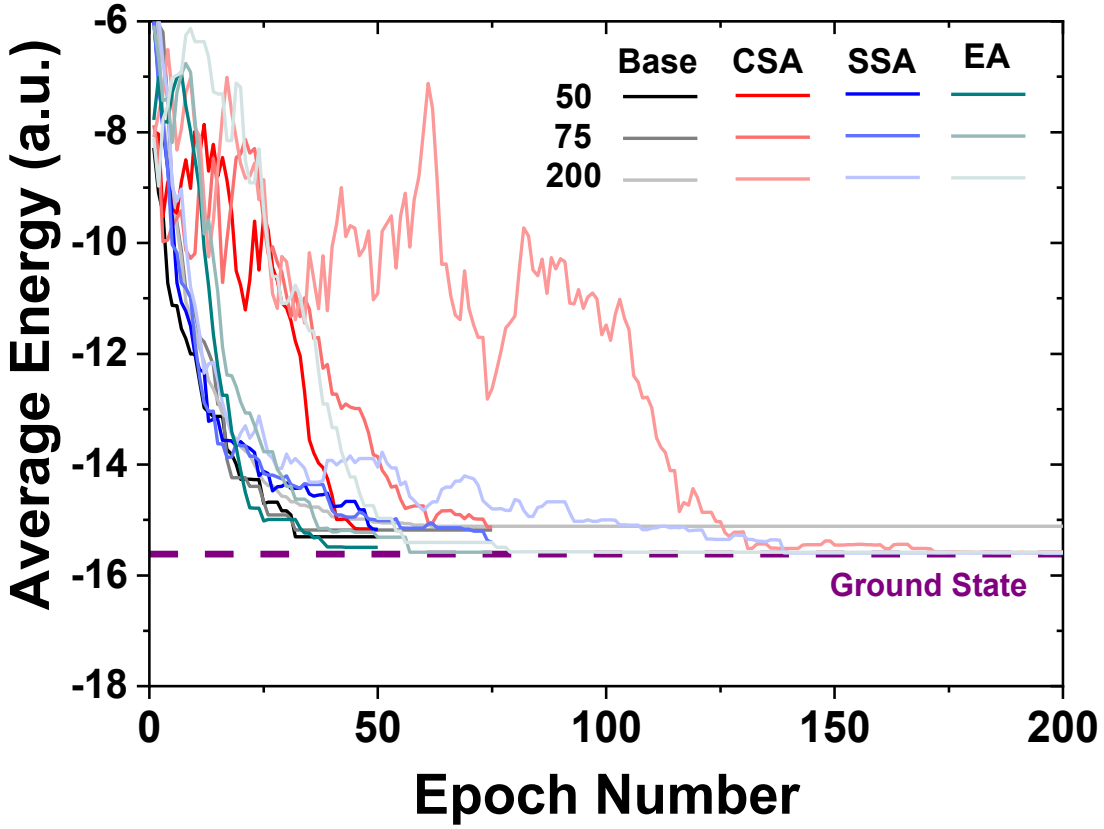
- 12-node maximum-weighted vertex cover problem



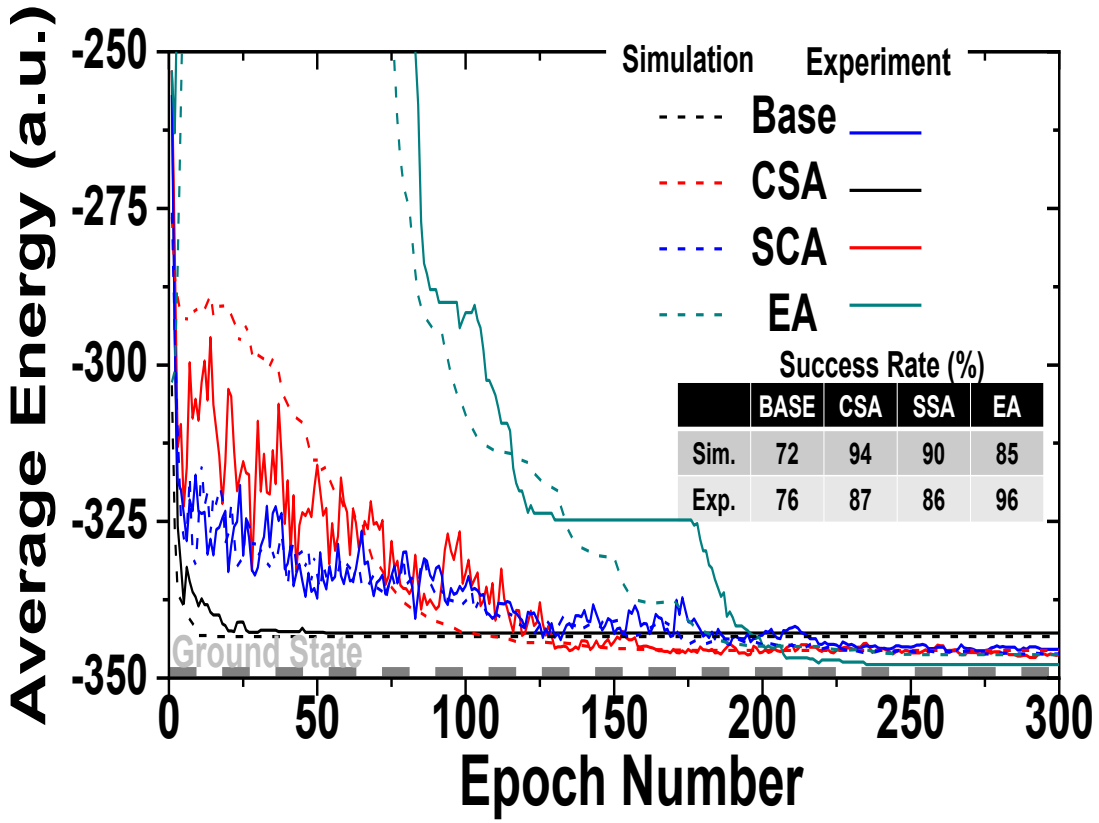
M.R. Mahmoodi et al., *Proc. IEDM'19*

New Result #3: Combinatorial Optimization Demo with Passive 64×64 Metal-Oxide Memristive Crosbar Circuits

- 10-node maximum-weight independent set problem



- 6-node maximum-weight graph partitioning problem



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

Summary

- In-memory analog computing based on emerging analog grade memory devices to enable very energy-efficient, compact, and fast analog VMMs
  - Near term: Metal oxide memristors (the most dense though least mature)
  - Long term: Embedded NOR floating gate memories (available at foundries now)
- Intrinsic noise of memory devices to implement stochastic transfer function or stochastic vector-by-matrix multiplication

Performance estimates & comparison to competition*

	Conventional		Emerging technology		This work	
	CPU	GPU	D-Wave	Fiber optics	Memristor	NOR flash
Time to solution (μs)	220	10	10^{10}	600	3	10
Energy to solution (μJ)	4000	2500	250×10^{12}	?	0.2	0.6

* benchmarked on noisy mean-field algorithm, adapted from *ArXiv:1903.11194*)

- Experimental demonstration of Bolztmann machines based on small-scale stochastic VMMs circuits with applications in deep believe networks and combinatorial optimization
- Major memristor challenges: poor yield, device uniformity, high cell currents

Relevant References

- **Stochastic neurocomputing and neuro-optimization demos**
  - M.R. Mahmoodi et. al.,“Combinatorial optimization by weight annealing in memristive Hopfield networks”, *to appear in Scientific Reports’20*
  - M.R. Mahmoodi et al., "An analog neuro-optimizer with adaptable annealing based on 64×64 0T1R crossbar circuit", *Proc. IEDM’19*
  - M.R. Mahmoodi et al., "Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization", *Nature Communications* 10, art. 5113, 2019
  - X. Guo et al., "Modeling and experimental demonstration of a Hopfield network analog-to-digital converter with hybrid CMOS/memristor circuits", *Frontiers in Neuroscience* 9, art. 488, Dec. 2015
  - L. Gao et al., "Digital-to-analog and analog-to-digital conversion with metal oxide memristors for ultra-low power computing", *Proc. NanoArch’13*, pp. 19-22
- **Passive metal-oxide memristors**
  - H. Kim et al. arXiv 2019
  - F. Merrikh Bayat et al., "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits", *Nature Communications* 9, art. 2331, 2018
  - G.C. Adam et al., “3-D memristor crossbars for analog and neuromorphic computing applications”, *IEEE TED 64* (1), pp. 312-318, 2017
  - M. Prezioso et al., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors", *Nature* 521, pp. 61-64, 2015
  - M. Prezioso et al., "Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al2O3/TiO2-x/Pt memristors", *Proc. IEDM’15*, pp. 17.4.1 – 17.4.4
- **NOR flash VMM-level experimental demos**
  - F. Merrikh Bayat et al. “High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cells”, *IEEE TNNLS* **29** 4782-4790 2018
  - X. Guo et al. “Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology”, *Proc. IEDM’17,* pp. 6.5.1-6.5.4
  - X. Guo et al., “Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells”, *Proc. CICC’17*, pp. 1-4

Questions?!
strukov@ece.ucsb.edu

Paper co-authors:

UC Santa Barbara: Zahra Fahimi, Hyungjin Kim, Hussein Nili, M. Reza Mahmoodi

Linkoping U., Norrkoping, Sweden: Leo Sedov and Val Polishchuk

Acknowledgments: G. Adam, F. Alibart, M. Bavandpour, B. Chakrabarti, N. Do, J. Edwards, M. Graziano, X. Guo, B. Hoskins, I. Kataeva, M. Klachko, K. Likharev, F. Merrikh Bayat, M. Prezioso, S. Sahay, A. Vincent

Sponsors (past and present):