Memory-Based Neuromorphic Hardware for Advanced Neural Network Models

D.B. Strukov

UC Santa Barbara

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

Sponsors:

SRC Semiconductor Research Corporation

SAMSUNG

NSF

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH UNITED STATES AIR FORCE

DARPA

SCIENCE TO SHAPE THE FUTURE OF THE ARMY ARMY RESEARCH OFFICE 1951

Google

DENSO

APPLIED MATERIALS

Part I. Introduction

(Brief overview of neurocomputing, mixed-signal hardware for simpler models)

NEURAL NETWORK MODELS

Resemblance to biology

Plasticity
On-line learning
Homeostasis

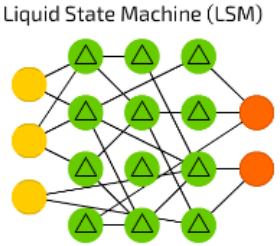
Stochasticity

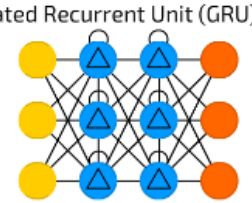
Task-specific
connectivity

Analog computing
In-memory computing
Massive parallelism

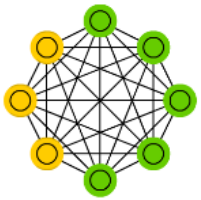
The most common operation for all: vector-by-matrix multiplication operation!

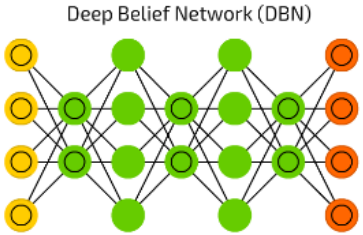
Spiking neural networks

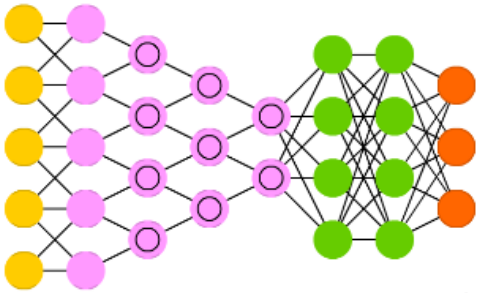


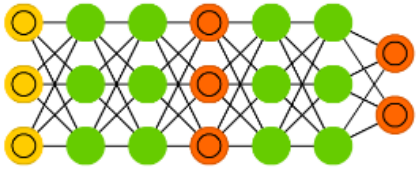


Stochastic neural networks

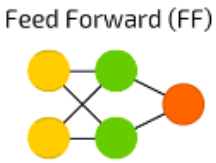


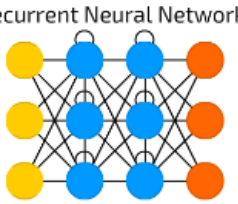


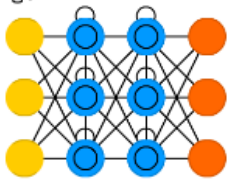


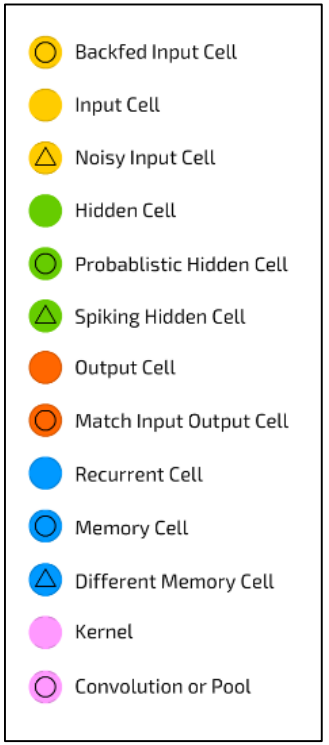


Recurrent and feed-forward deep networks









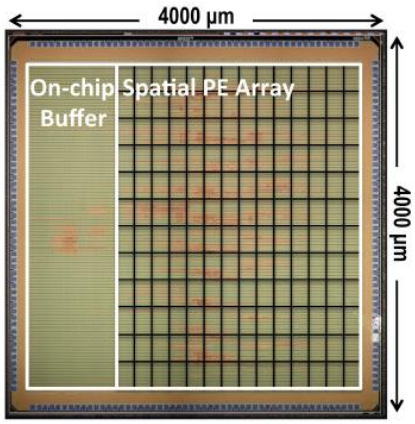
practically useful now (speech, image, video, recommendation, analytics applications etc.)

Neuron/synapse functionality

Network figures adapted from http://www.asimovinstitute.org/neural-network-zoo/

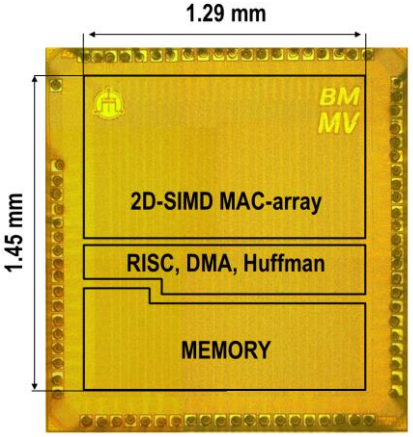
STATE-OF-THE-ART (ALEXNET) DEEP LEARNING HARDWARE: CUSTOM DIGITAL CIRCUITS

Eyeriss



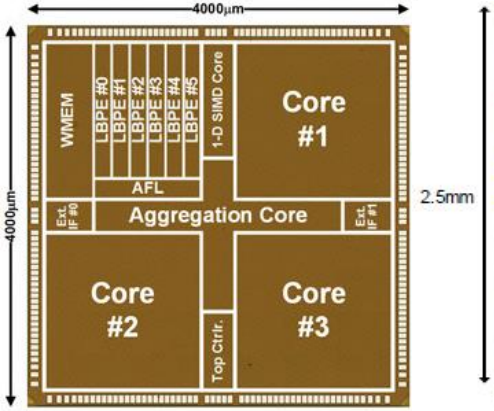
Y.H. Chen *et al., ISSCC'16*

Envision



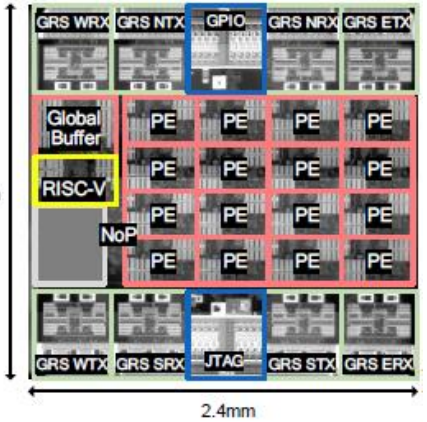
B. Moons *et al., ISSCC'17*

UNPU



J. Lee *et al., ISSCC'18*

NVidia MCM



B. Zimmer *et al., VLSISymp'19*

	Eyeriss (2016)	Envision (2017)	UNPU (2018)	NVidia (2019)
Technology (nm)	65 LP CMOS	28 UTBB FD-SOI	65	16
Peak performance [GOPS]	67	(1 to 4) × 102	1382 (4) /345 (16)	4010 (8 bit)
Active area [mm^2]	12.25	1.87	16	3.1
Filter size	1-1024	any	any	any
Precision [b]	16	4-16	1-16	8
Power [mW] @ frame rate [fps]	278 @ 34.7*	44 @ 47*	297 @ ? *	?
Min/max energy efficiency [TOps/J]	0.15 – 0.35	0.26 – 10 (~**2.5** for 4-bit)	50.6 (1 bit) / **11.6** (4 bit) / 3.08 (16 bit)	~ 9.09 (8 bit)

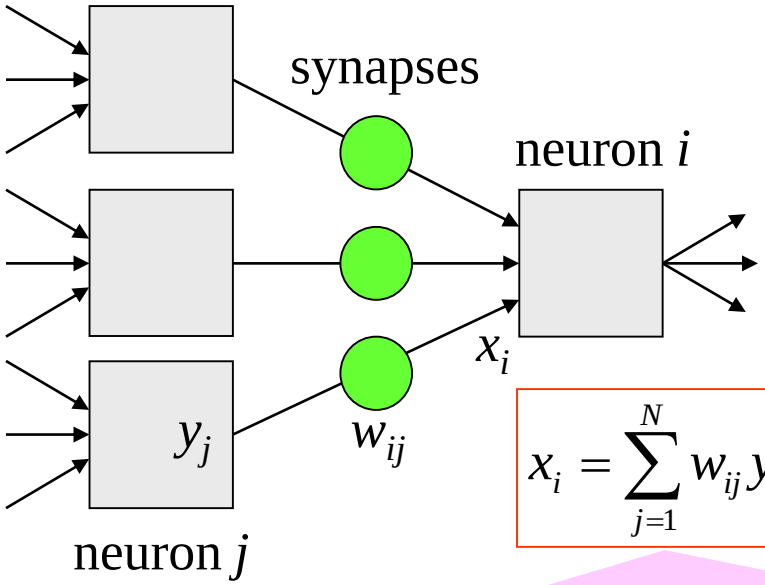
* AlexNet convolutional layers only

- Saturating improvements for purely digital implementations
- Biology is five-six orders of magnitude more energy efficient

BASIC IDEA FOR RADICAL IMPROVEMENT: ANALOG VECTOR-BY-MATRIX MULTIPLICATION

VMM:

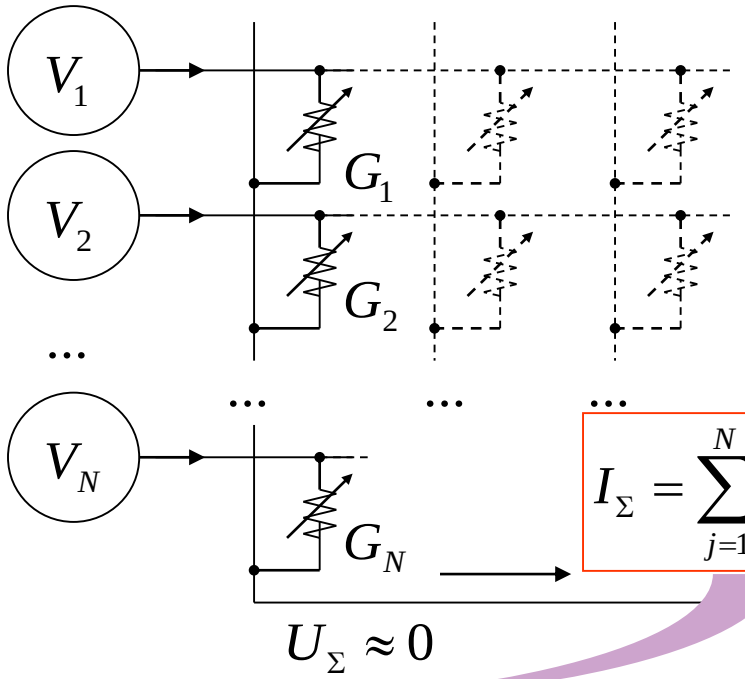
basic neuromorphic operation



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

Analog VMM:

using the Ohm & Kirchhoff laws



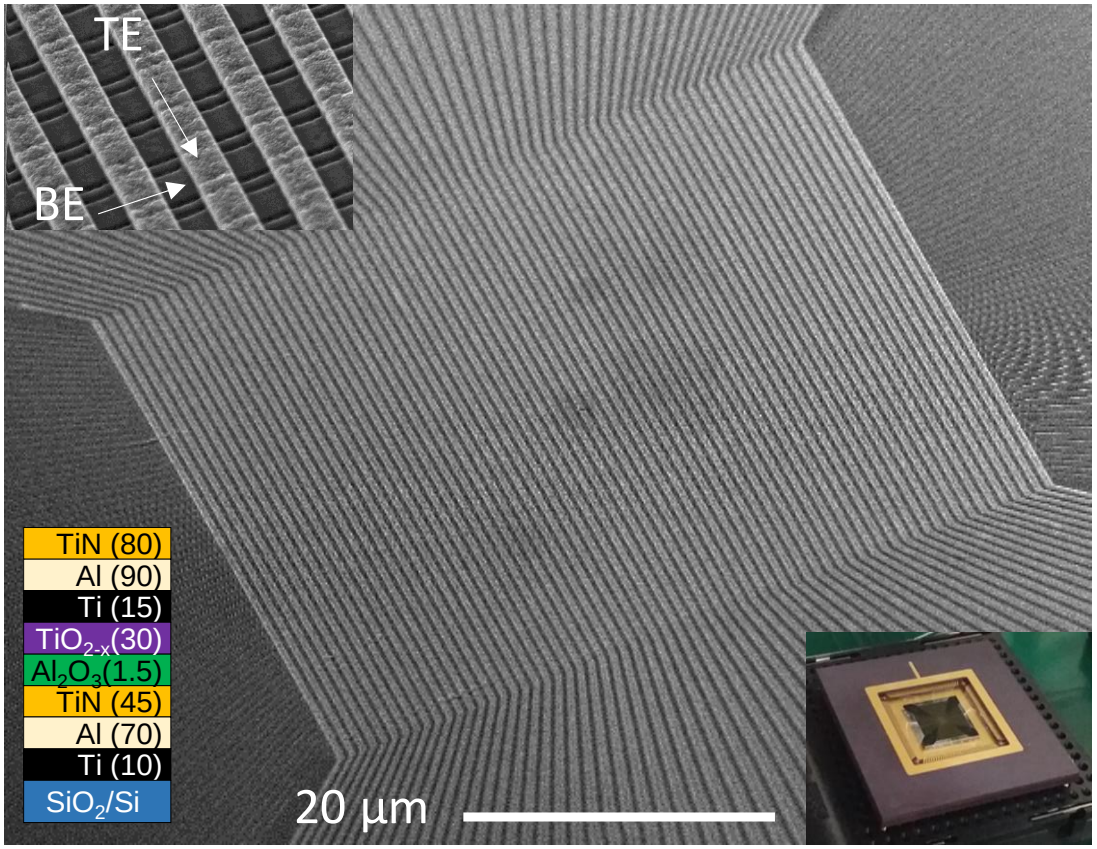
$$I_\Sigma = \sum_{j=1}^{N} G_j V_j$$

Features:

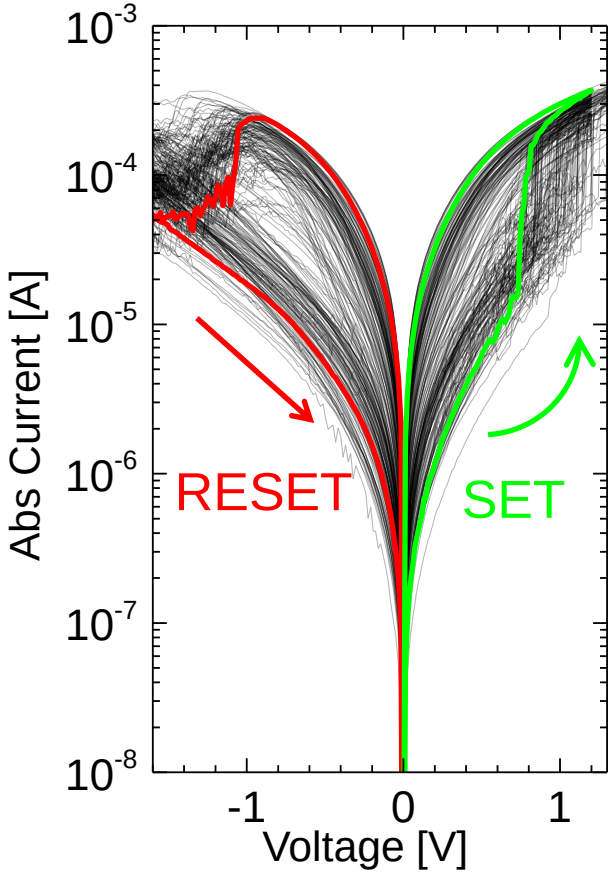
- physical-level, in-memory, fast, very energy-efficient
- proposed by Widrow in 1960s, popularized by Carver Mead and students (CalTech) in the 1980s
- no **dense** adjustable-conductance crosspoint devices - until recently

UC SANTA BARBARA'S MEMRISTORS

- **64 × 64 crossbar circuit**



- **Typical I-Vs**



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

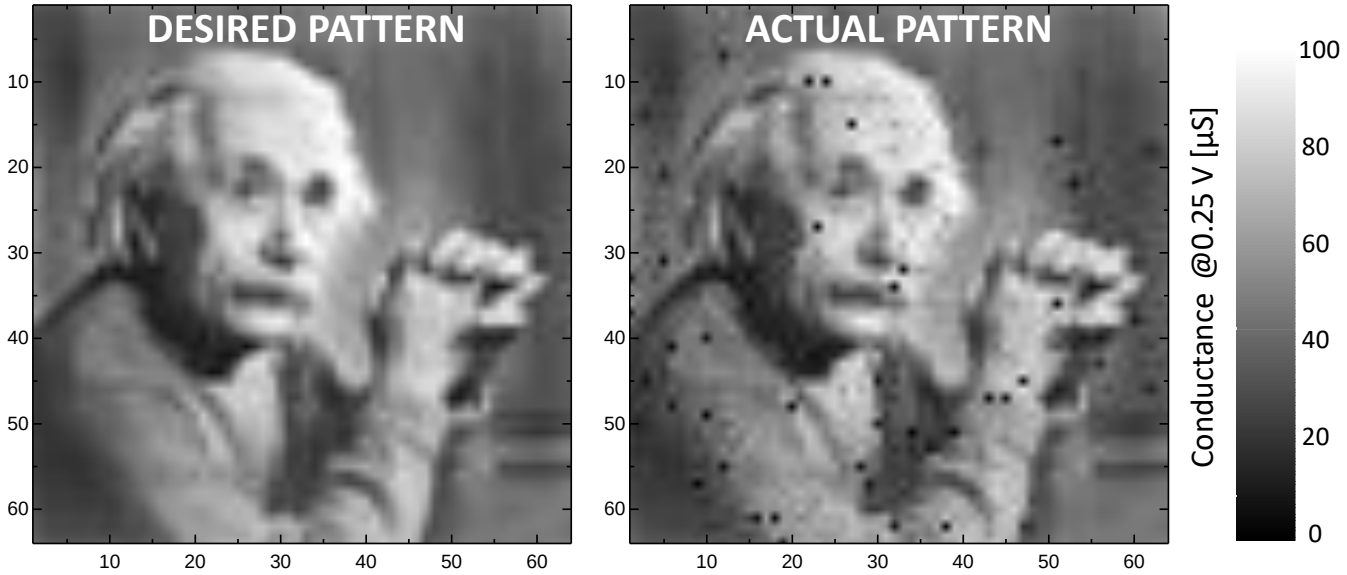
Details:

- Al_2O_3/TiO_{2-x} active bilayer by reactive sputtering
- ~250 nm wide lines, passive (0T1R) integration
- CMP/dry etching and TiN/Al electrodes for higher conductance
- Higher as-fabricated film conductance → low forming voltage → very uniform I-Vs

>250x/10,000x better memristor/memory cell density compared to 1T1R work from HPL/UMass collaboration at comparable complexity

ANALOG APPLICATION DEMO WITH 4K-DEVICE CROSSBAR

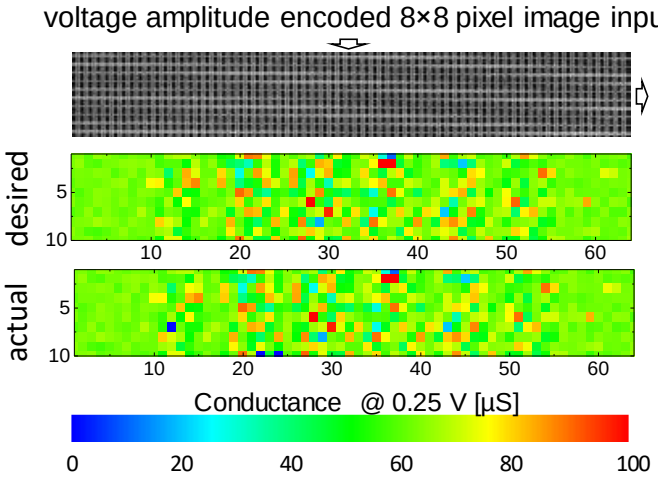
- **Conductance tuning**

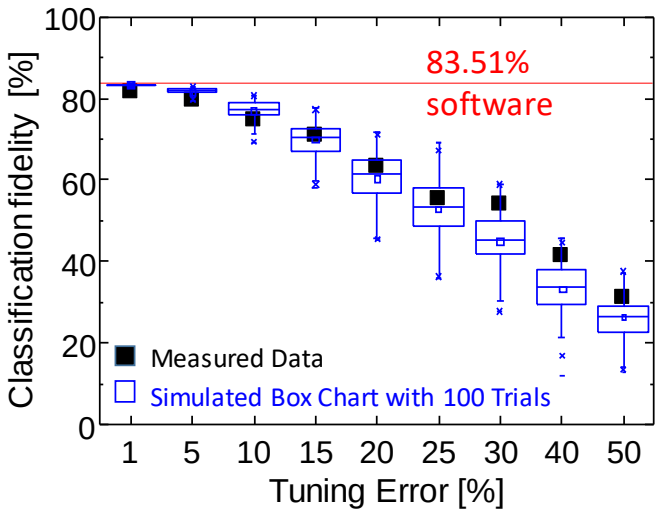


- Color encoding: 256 levels from white (10 µS) to black (100 µS) @ 0.2V
- <5% relative, <3% absolute tuning precision with automated algorithm
- Black dots: ~1% devices that cannot be switched

- **MNIST classification**

H. Kim *et al.* arXiv 2019

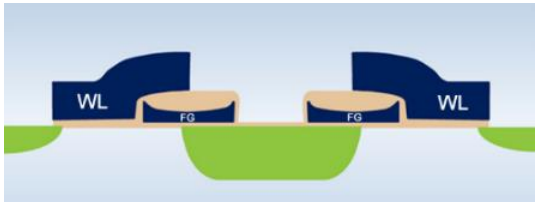




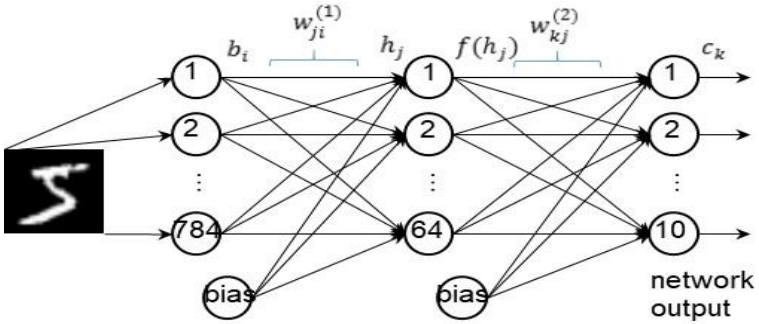
- Ex-situ trained single 64-10 single layer perceptron
- Emulated neuron functionality
- Very close to simulation measured fidelity (within 1.5%) for highest fidelity

NEUROCOMPUTING BASED ON FLOATING GATE DEVICES: ARHICTECTURE AND CHIP DEMO

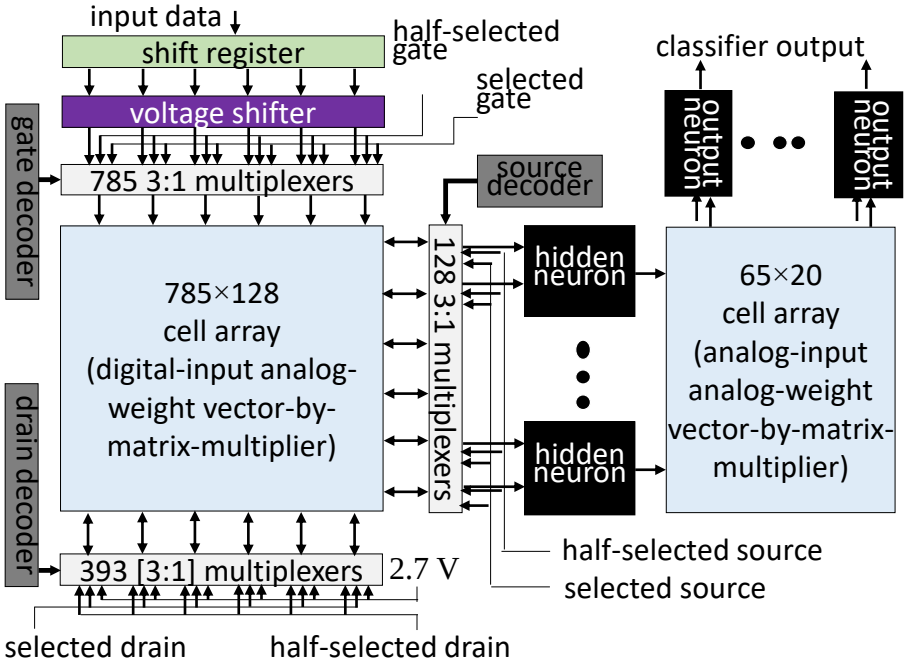
- **NOR eFlash**



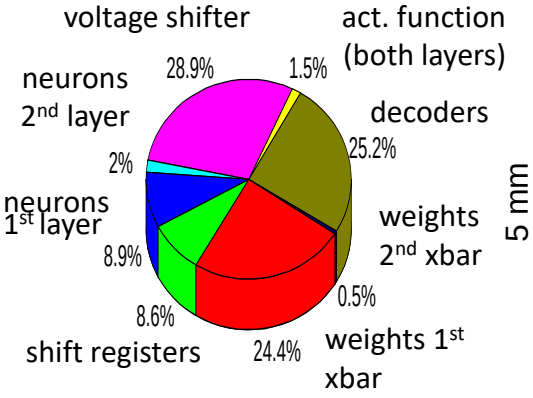
- **Multilayer perceptron circuit**



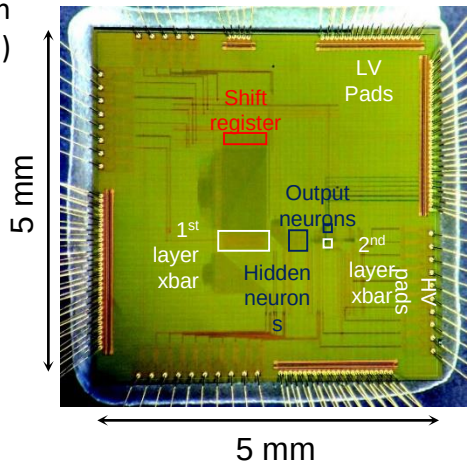
- **High-level architecture**



- **Area breakdown and chip layout**



Total area : 0.78 mm² (active components)



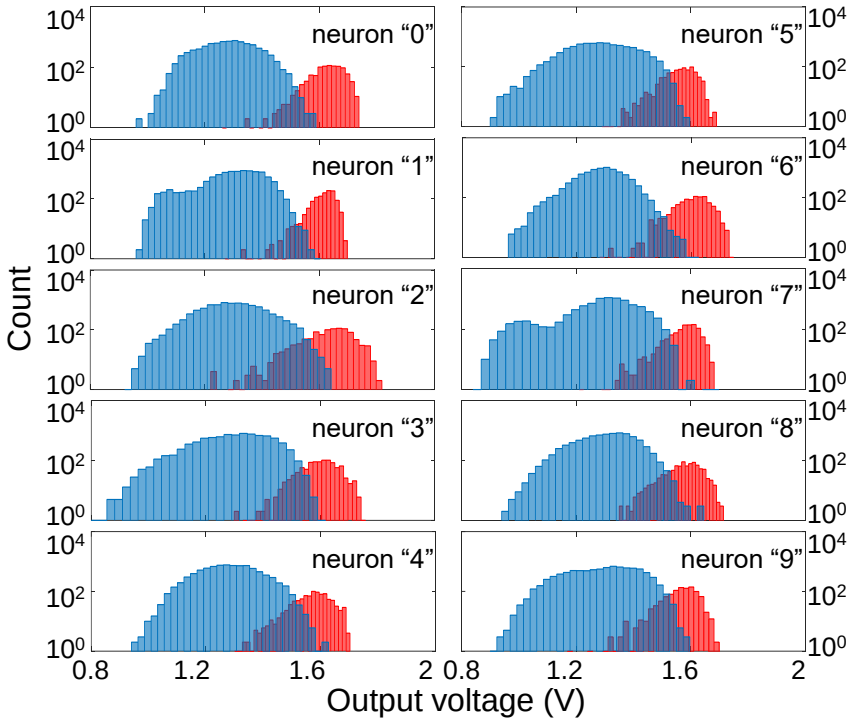
F. Merrikh Bayat et al., *TNNLS'18,* X. Guo et al., *IEDM'17*

Classifier features:

- 28x28 B/W input, 10-class output
- >100,000 NOR flash synapses, 64 hidden layer CMOS neurons
- 180-nm process with eFlash
- Differential implementation of synaptic weights
- High voltage circuitry for weight import

NEUROCOMPUTING BASED ON FLOATING GATE DEVICES: EXPERIMENTAL RESULTS

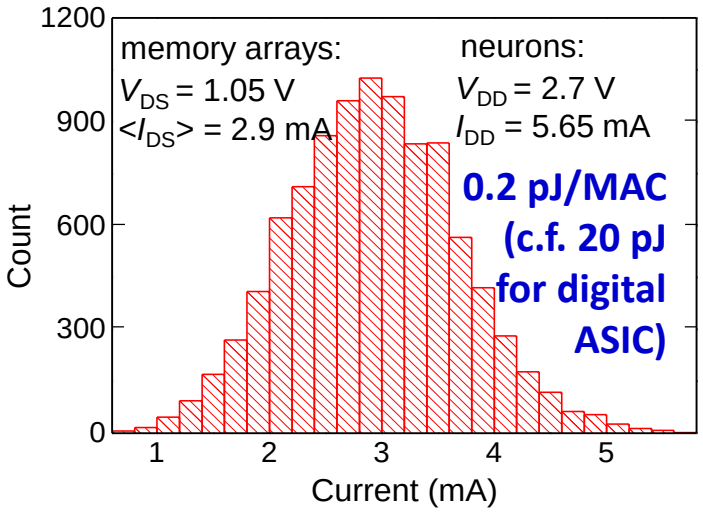
- **Classification performance**
  (10,000 MNIST test patterns)



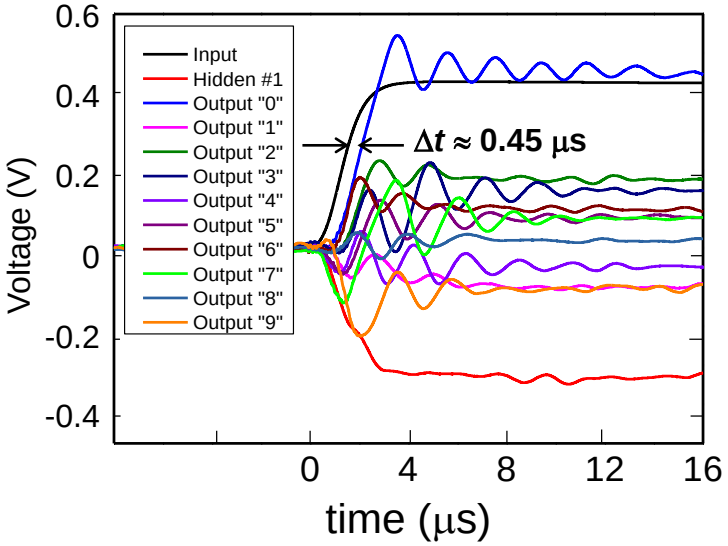
Experimental measured performance for MNIST:

- 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 purely digital circuits at comparable MNIST fidelity (6 orders of magnitude better energy-delay compare to IBM TrueNorth)
- Reproducible, temperature insensitive, no change in performance after 7 months

- **Power consumption**



- **Latency (one pattern)**



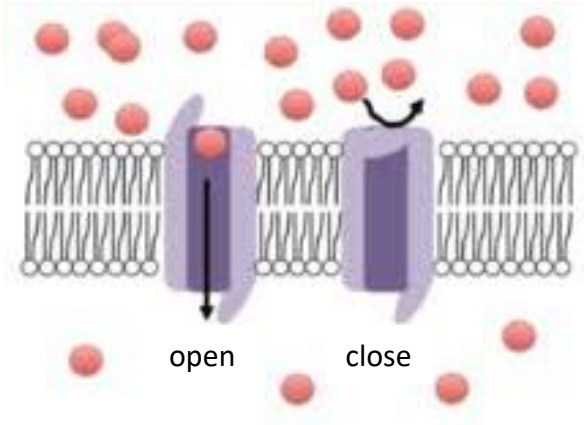
F. Merrikh Bayat et al., *TNNLS'18,* X. Guo et al., *IEDM'17*

Part II. Hardware for Stochastic Neural Networks

STOCHASTIC NEUROCOMPUTING

Molecular-level operations in the brain are stochastic, e.g.

Voltage-gating of ion channels



Neurotransmitter release at synaptic cleft

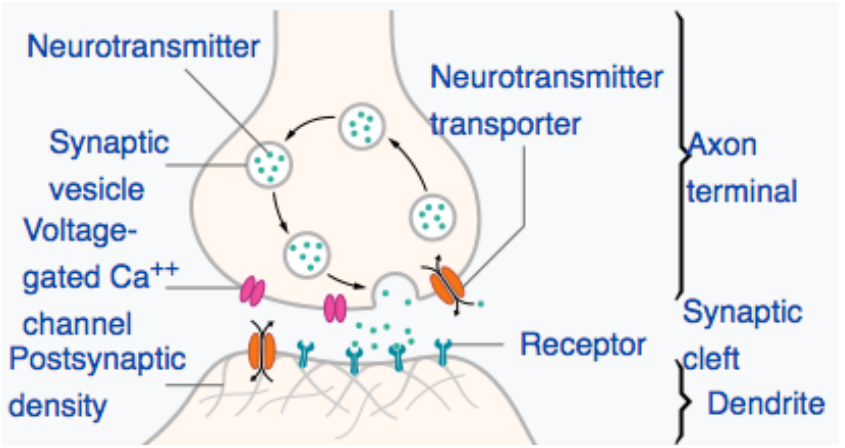
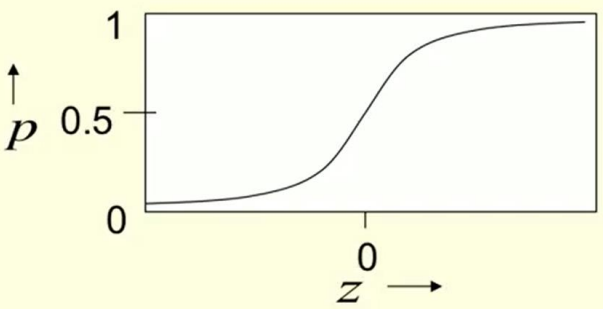


Image source: Wikipedia

- **Stochastic (binary) neuron**

$$z = b + \sum_i x_i w_i$$

$$p(s=1) = \frac{1}{1+e^{-z}}$$



- **Stochastic networks**

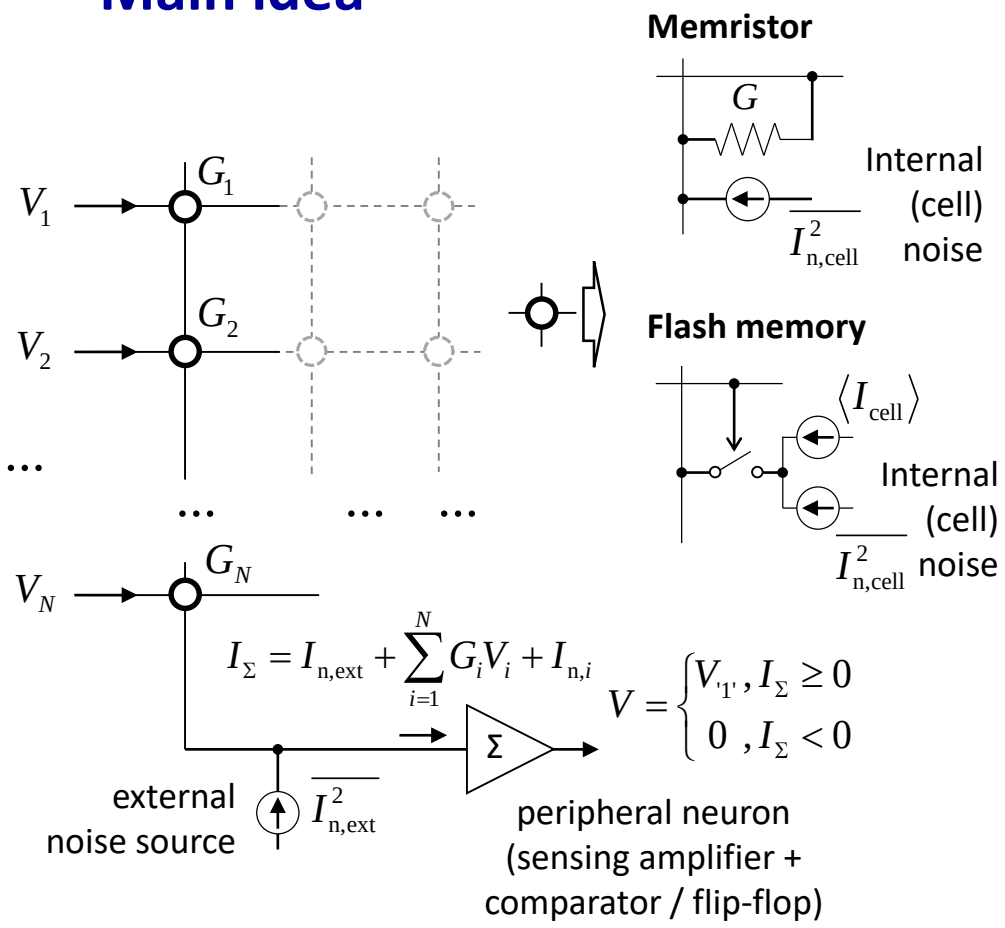
- Boltzmann machines
- Restricted Boltzmann machines
- Stochastic Hopfield networks
- Deep Believe networks
- Bayesian networks

...

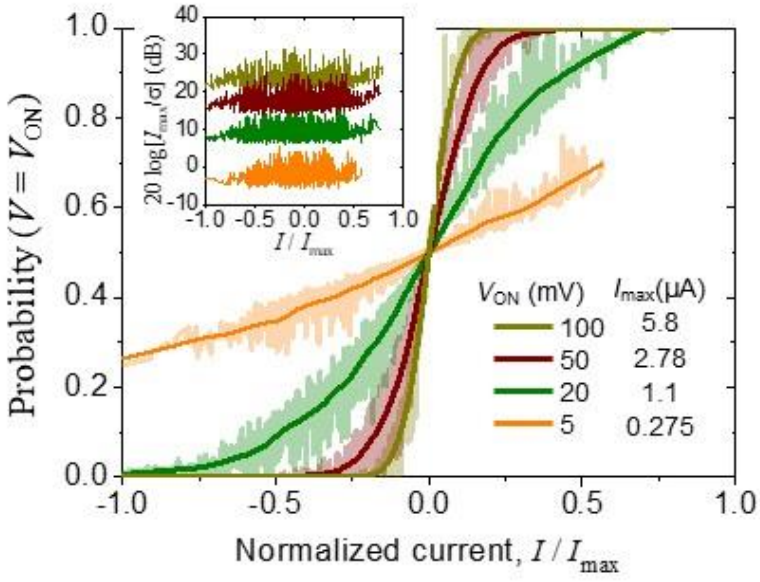
Need efficient implementations of both dot-product and stochastic functionality

STOCHASTIC DOT PRODUCT CIRCUIT

- **Main idea**



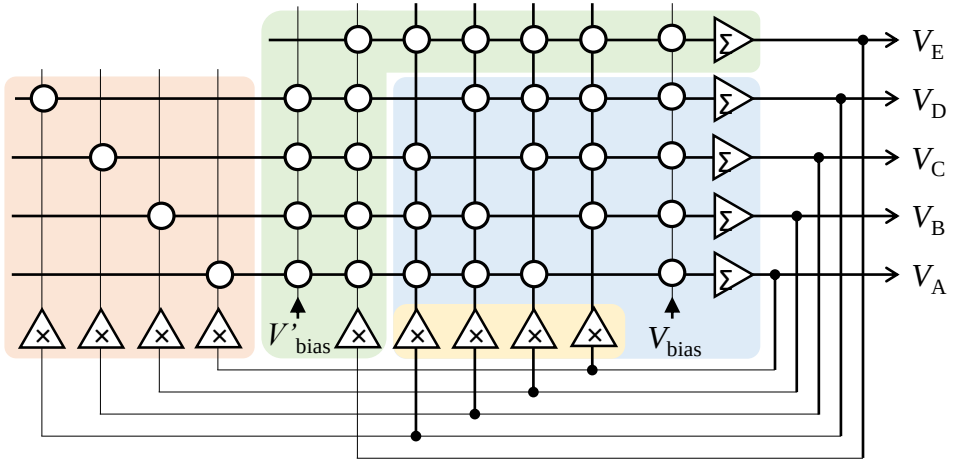
- **Experimental results**
  (memristor-based circuits, externally-injected noise from readout circuitry)



- Rely on intrinsic (from memory cells) and/or externally injected noise
- Sigmoid slope (computing temperature) controlled by the applied (V_{ON}) voltage

NEUROOPTIMIZATION WITH STOCHASTIC HOPFIELD NETWORKS

Hopfield network with annealing



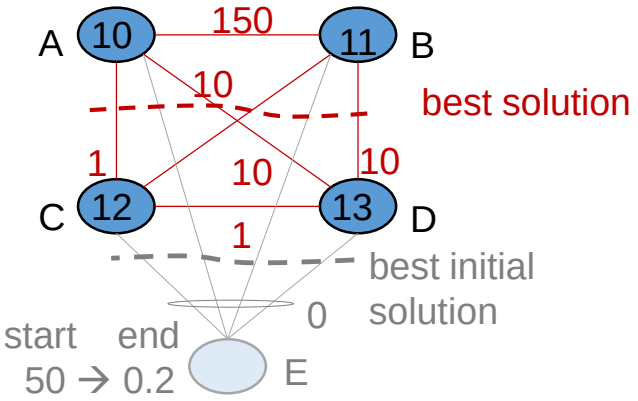
Color background highlights circuitry for:

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

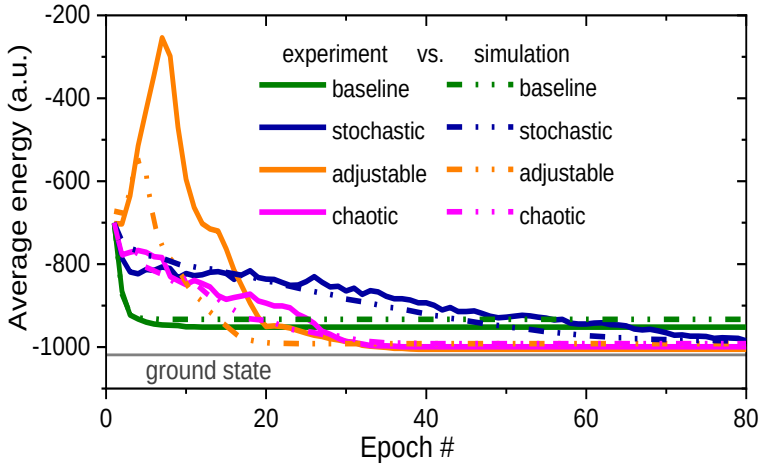
× = scaling Σ = summing

M. Mahmoodi et al., *submitted* April 2019; Background work: 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

Graph partitioning problem



Experimental results



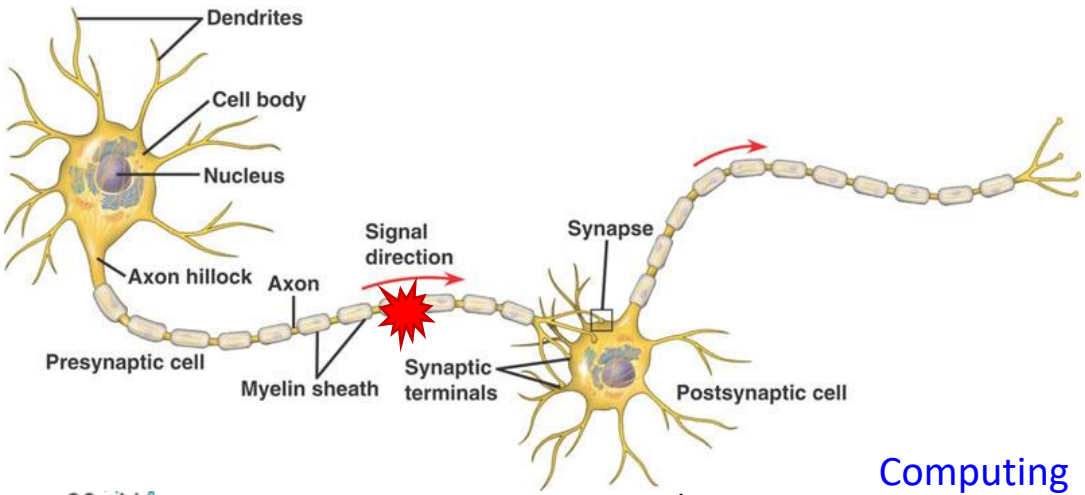
Comparison with other approaches

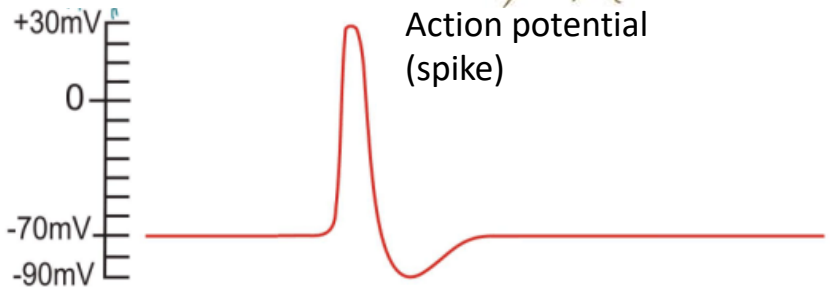
Adapted from ArXive 1903.11194

	Conventional		Emerging technology		Our 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

Part III. Neuromorphic Hardware for Spiking Neural Networks

SPIKING NEURAL NETWORKS

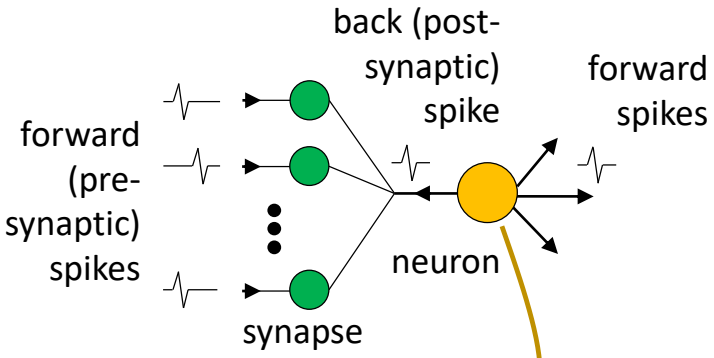


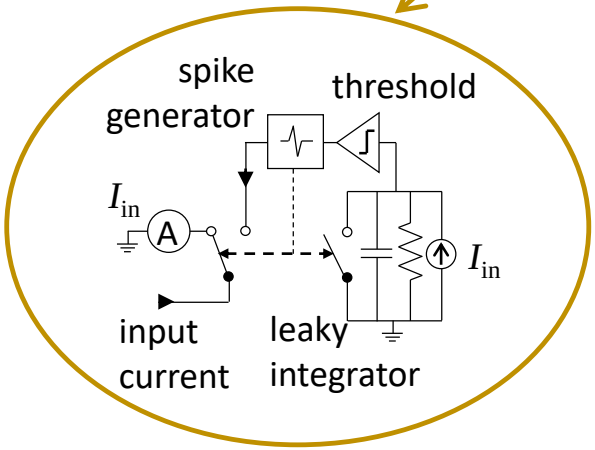


Action potential (spike)

Computing model

Leaky integrate & fire neuron

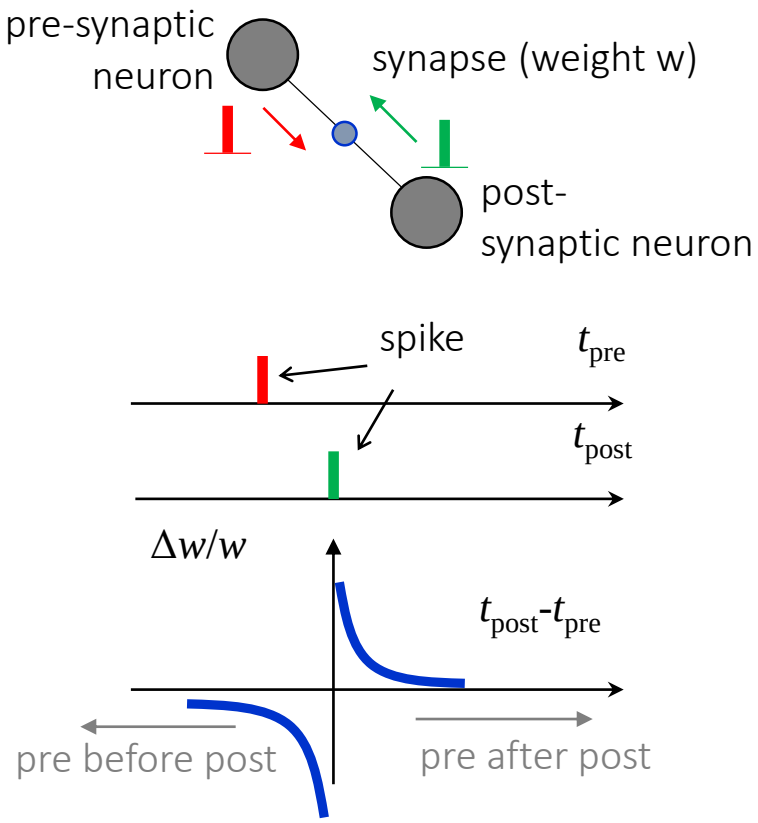




- Information encoded in timing of spikes (rate vs. temporal)
  - coordinated processing of spatial-temporal information
  - believed to be more energy efficient
- Local learning rules for synaptic weight update → suitable for online training and HW friendly
- More biologically plausible (but so far outperformed by firing-rate ANNs in virtually all machine learning tasks)

SPIKE-TIMING-DEPENDENT PLASTICITY

- **Main idea**



STDP is essential feature of spiking neural networks

- **STDP in cultured hippocampal neurons**

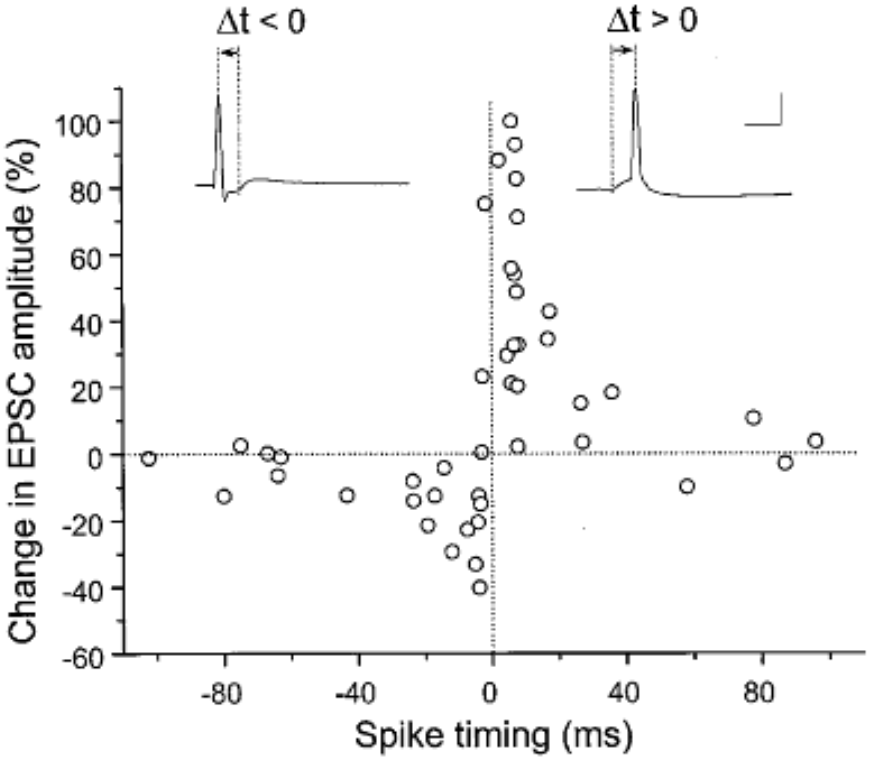
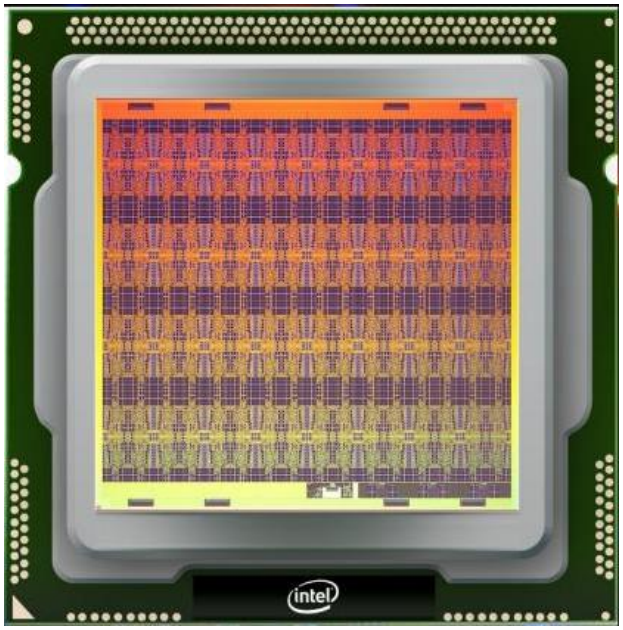


Figure 7. Critical window for the induction of synaptic potentiation and depression. The percentage change in the EPSC amplitude at 20–30 min after the repetitive correlated spiking (60 pulses at 1 Hz) was plotted against the spike timing. Spike timing was defined by the time interval (Δt) between the onset of the EPSP and the peak of the postsynaptic action potential during each cycle of repetitive stimulation, as illustrated by the traces above. For this analysis, we included only synapses with initial EPSC amplitude of <500 pA, and all EPSPs were subthreshold for data associated with negatively correlated spiking. Calibration: 50 mV, 10 msec.

Bi and Poo, Annual Rev. Neurosci. 2001. 24:139–66

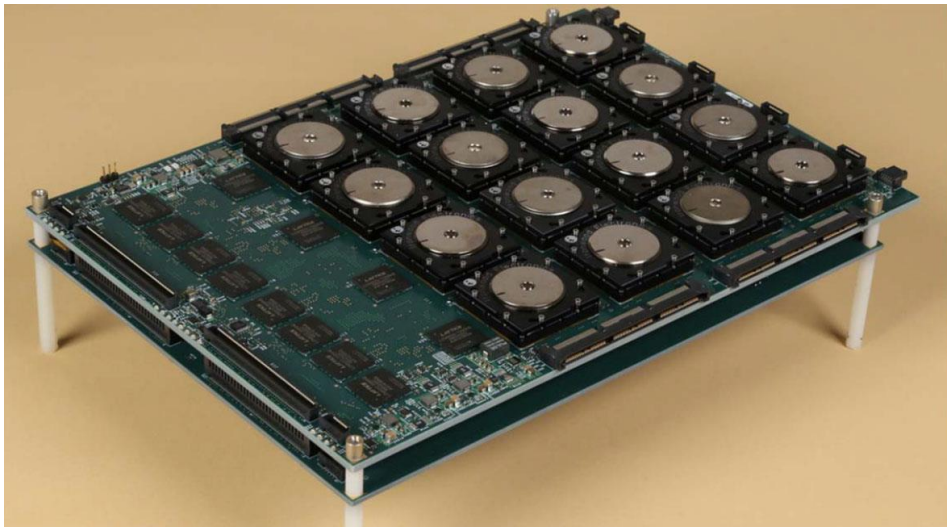
STATE-OF-THE-ART DIGITAL SPIKING NEUROMOPRHIC HARDWARE

Intel Loihi 2018



14 nm
128 M synapses
128 M neurons
2.07 B transistors
on-chip learning

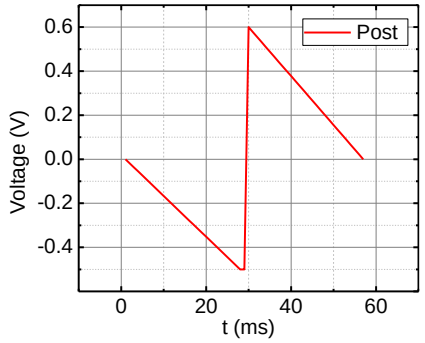
IBM TrueNorth 2014

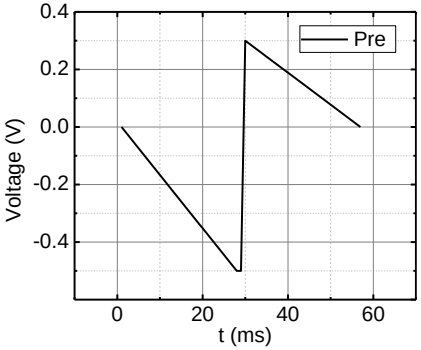


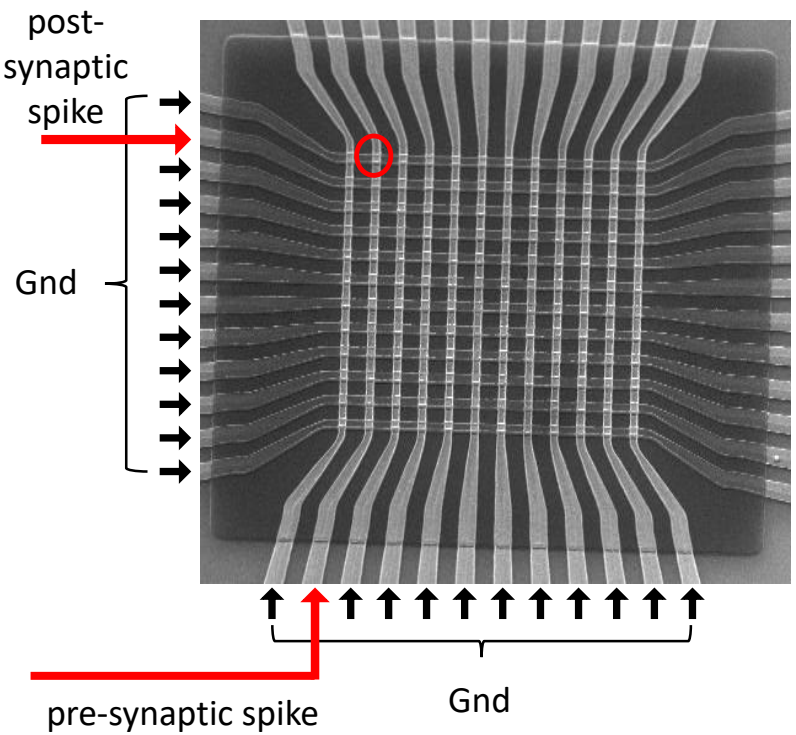
28 nm
256 M synapses
1 M neurons
5.4 B transistors
inference only

EXPERIMENTAL DEMONSTRAITON OF STDP BASED ON MEMRISTIVE CROSSBAR CIRCUITS

Applied voltage pulses

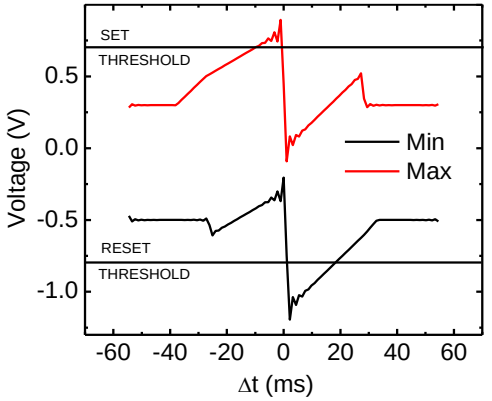




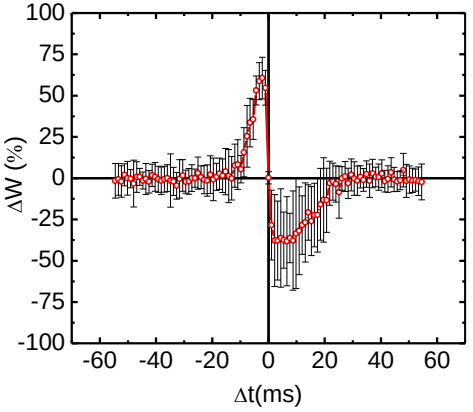


M.Prezioso et al., Nat. Sci. Rep. 2016

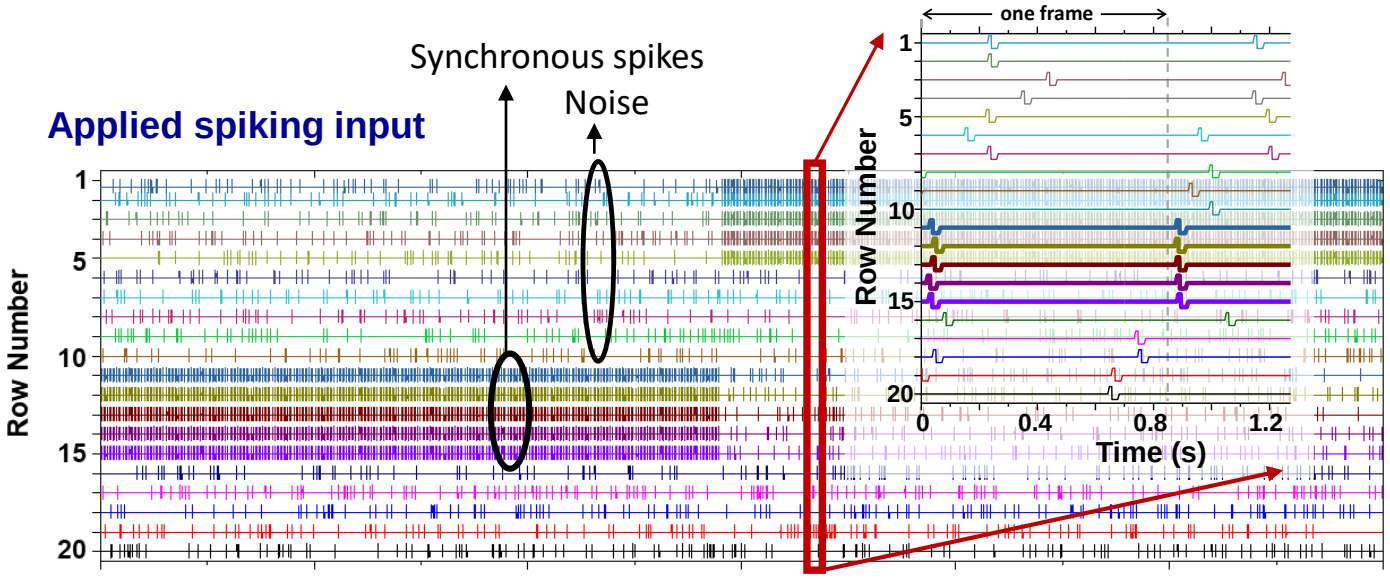
Voltage across memristor

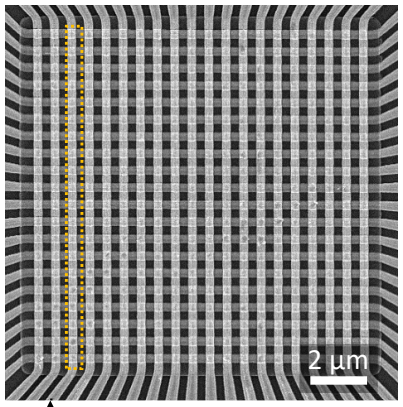


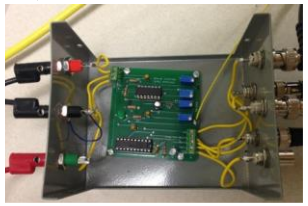
Measured STDP



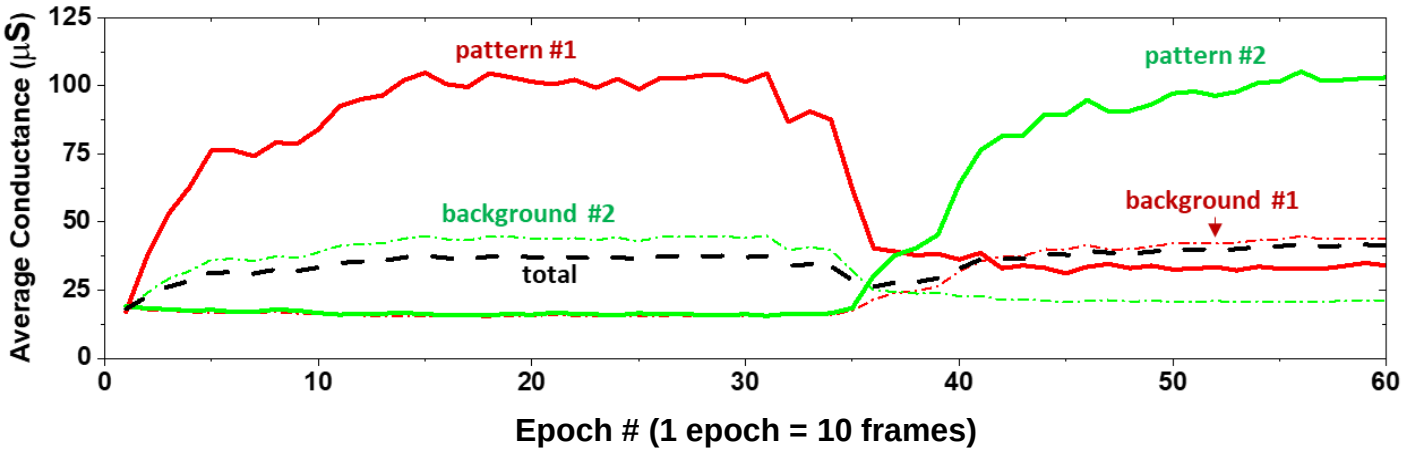
COINCIDENCE DETECTION BY PASSIVE MEMRISTOR-BASED SPIKING NEURAL NETWORK







CMOS IC LIF neuron

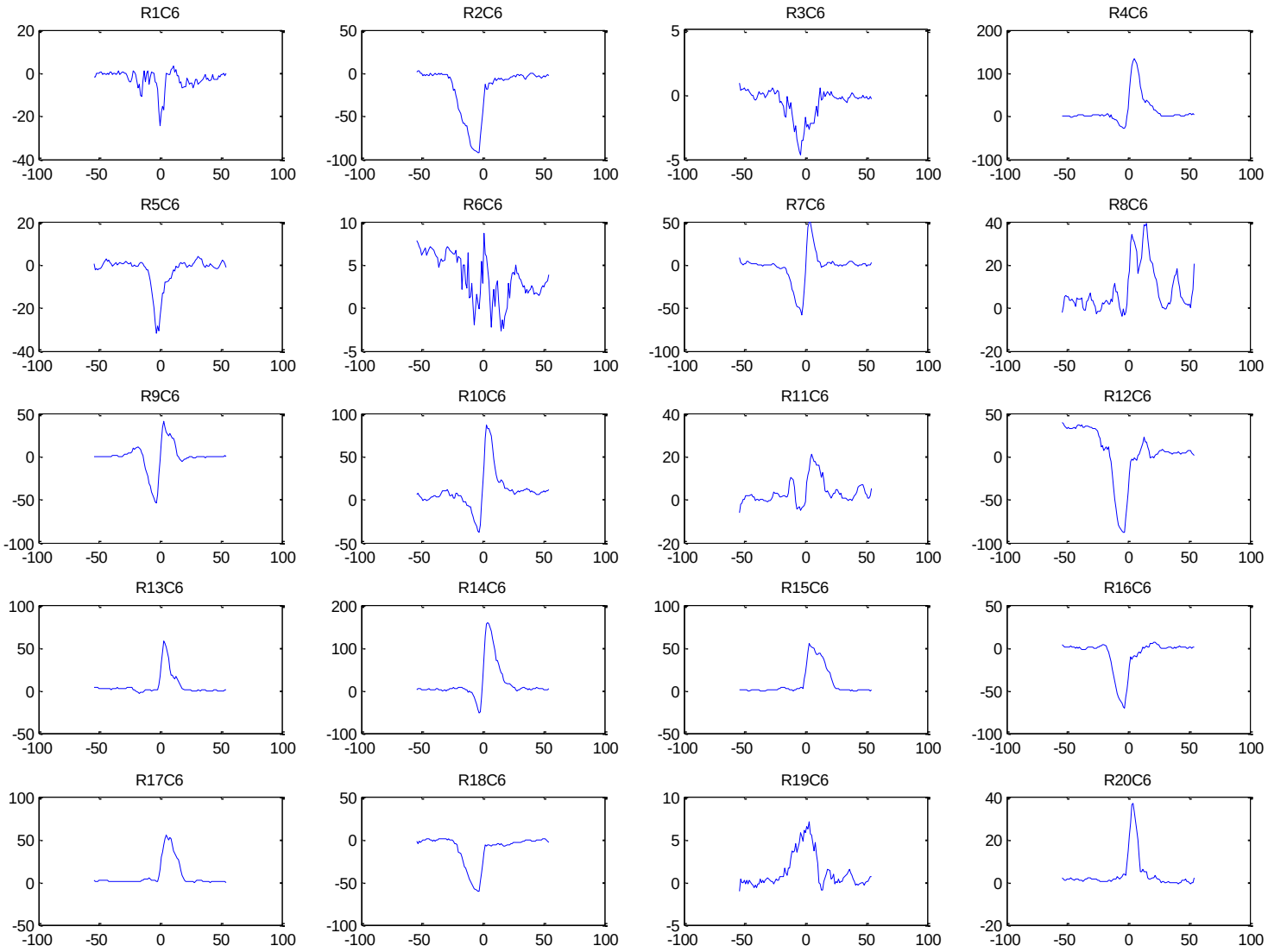


Main results: Demonstration of unsupervised learning of spike coincidence (i.e. spiking on synchronous input) via STDP mechanism

M. Prezioso et al. *Nature Comm* (2018)

CHALLENGES FOR SPIKING NEURAL NETWORKS

STDP variations for 20 xbar devices



M. Prezioso et al. *Nature Comm* (2018)

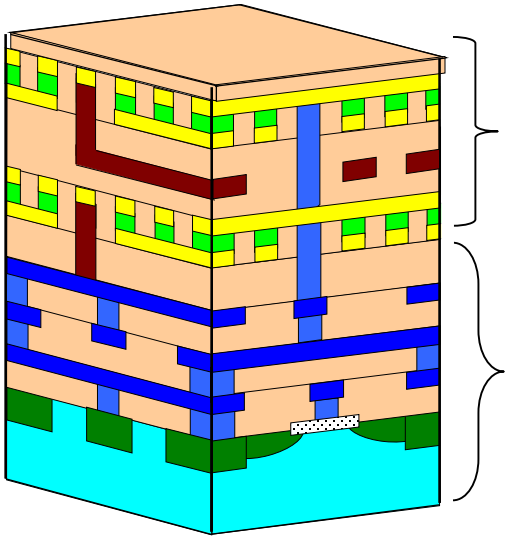
More severe impact of d2d + c2c variations than for ex-situ-trained systems

Require higher switching endurance

FUTURE WORK AND CHALLENGES

Important future work

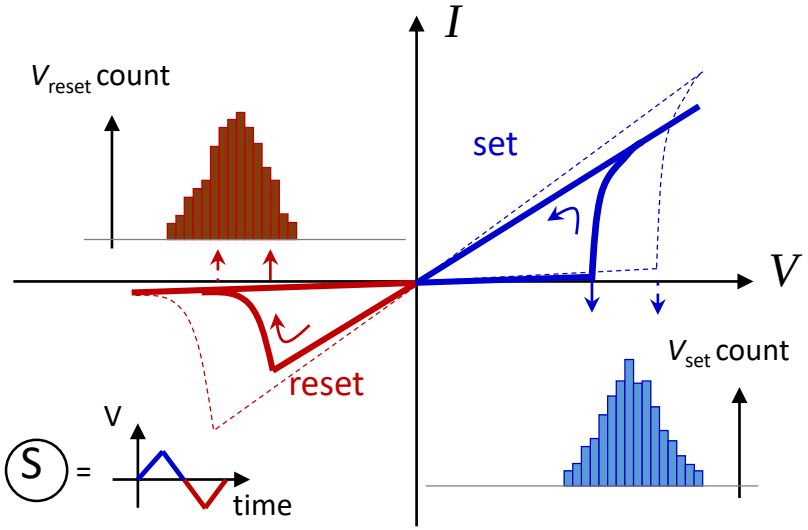
- Improving memristor technology, monolithic integration with CMOS, e.g. 3D CMOL



- >1000x better in energy-delay over purely digital system (experiment for smaller scale, sim for larger scale systems)
- 10^{13} synapses per cm^2 for 100-layer 10-nm memristive crossbar circuits (~30x less compared to human brain)

Challenges

- Device yield, device to device and cycle to cycle variations in I-V, switching endurance (but more tolerant to defects than logic)



- Economical and confidence barriers
- Lack of algorithms for higher intelligence

THANK YOU!