Neuromorphic Computing with NanoCrossbar Circuits

Dmitri Strukov

University of California at Santa Barbara

Acknowledgments:

Research group: G. Adam, B. Chakrabarti, X. Guo, B. Hoskins, F. Merrikh Bayat, M. Prezioso

Collaborators: P. Auroux, J. Edwards, M. Graziano (BAE), I. Kataeva (DENSO), K. K. Likharev (SBU), N. Do (SST), L. Sengupta (NGC)

Funding support: AFOSR MURI, ARO, DARPA UPSIDE, DENSO CORP., NSF

RECENT SURGE OF A.I.

(you could not avoid the buzz…)

The Washington Post

Tuesday, March 1, 2016 Edition: U.S. & World | Regional

How artificial intelligence is moving from the lab to your kid's playroom

The New York Times

Google Car Exposes Regulatory Divide on Computers as Drivers

The New York Times

A Learning Advance in Artificial Intelligence Rivals Human Abilities

By JOHN MARKOFF DEC. 10, 2015

The New York Times

Start-Up Lessons From the Once-Again Hot Field of A.I.

Bits

By STEVE LOHR FEB. 28, 2016

Los Angeles Times

Toyota invests $1 billion in artificial intelligence in U.S.

The Washington Post

Tuesday, March 1, 2016 Edition: U.S. & World | Regional

Can AI fix the world? IBM, TED and X Prize will give you $5 million to prove it.

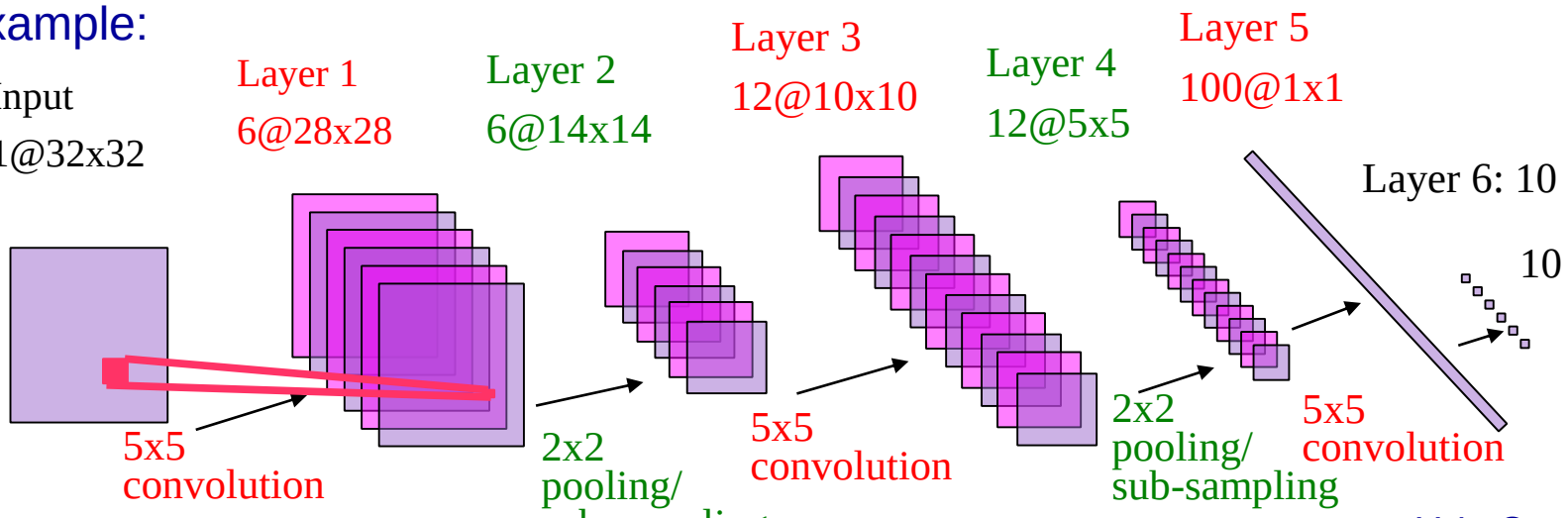
Drumpf Twitterbot learns to imitate Trump via deep-learning algorithm

"OK, it's amazing right now with ISIS, I tell you what? I don't want them to vote, the worst very social people. I love me"

Donald TrumpBot
@thetrumpbot

PATTERN CLASSIFICATION IN CONVOLUTIONAL (A.K.A. "DEEP") NETWORKS

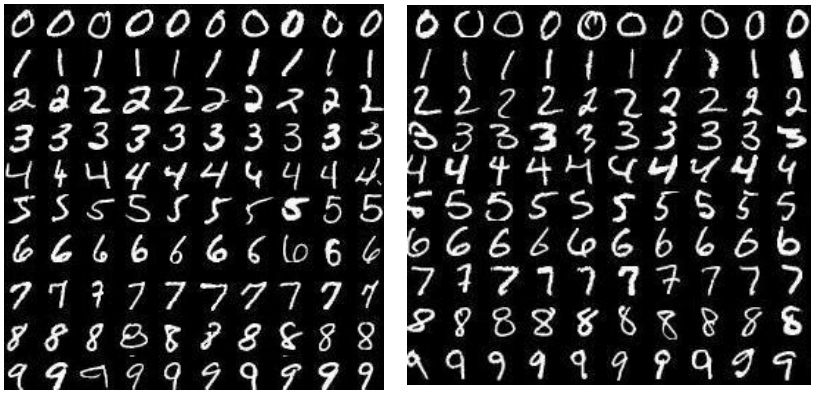
Example:



Y. LeCun (2007)

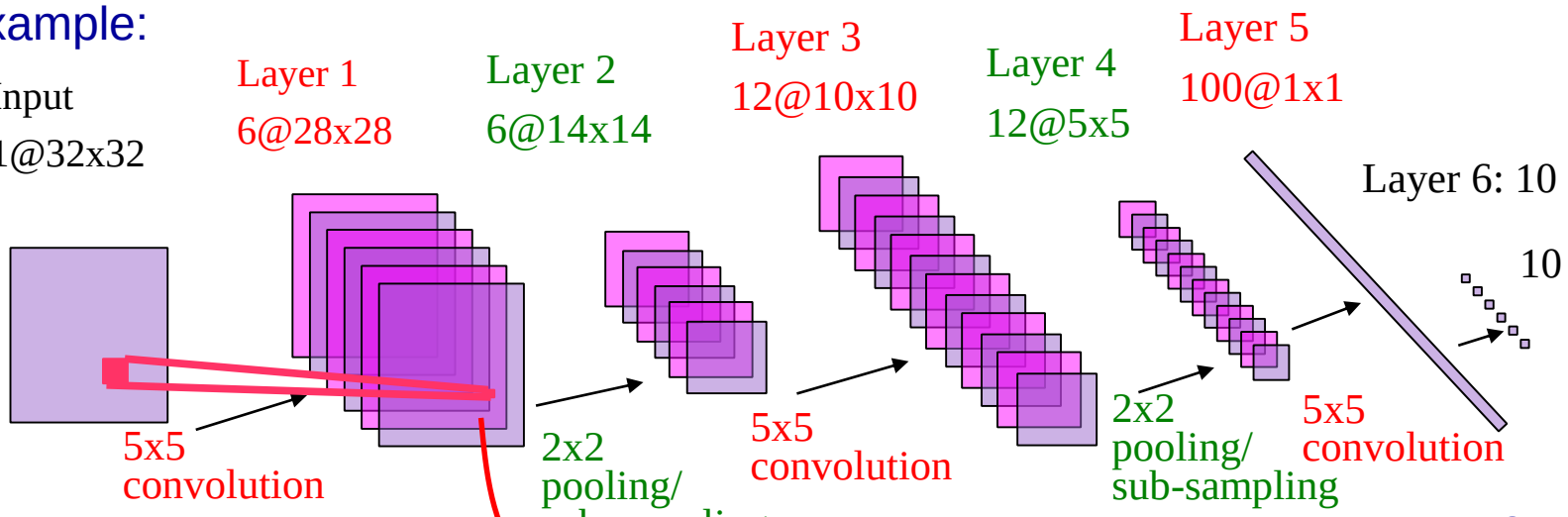
- MLP with limited bio-inspired connectivity
- the best method for hand-writing recognition
- used by NCR for check reading machines
- used by Microsoft for OCR
- 0.62% error on the MNIST set

MNIST set (60,000-image database)



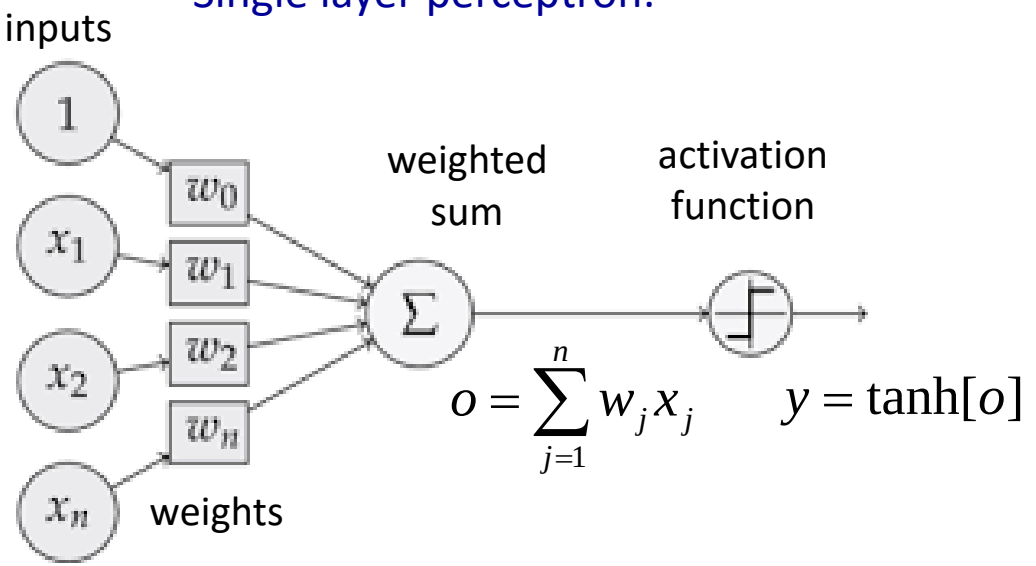
PATTERN CLASSIFICATION IN CONVOLUTIONAL (A.K.A. "DEEP") NETWORKS

Example:



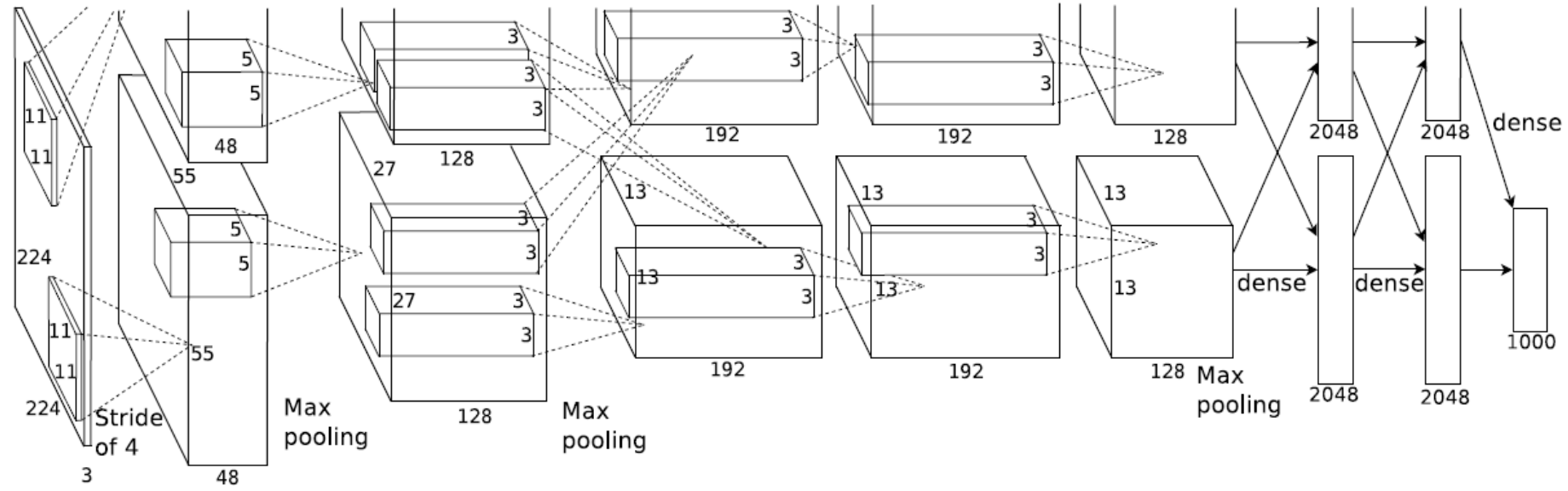
Y. LeCun (2007)

Single layer perceptron:



- supervised error backpropagation training
- weights determine functionality
- neurons learn "features"
- most computationally intensive operation: dot product
- 4 to 5-bit synapses (weights) accuracy sufficient

U. TORONTO'S NETWORK

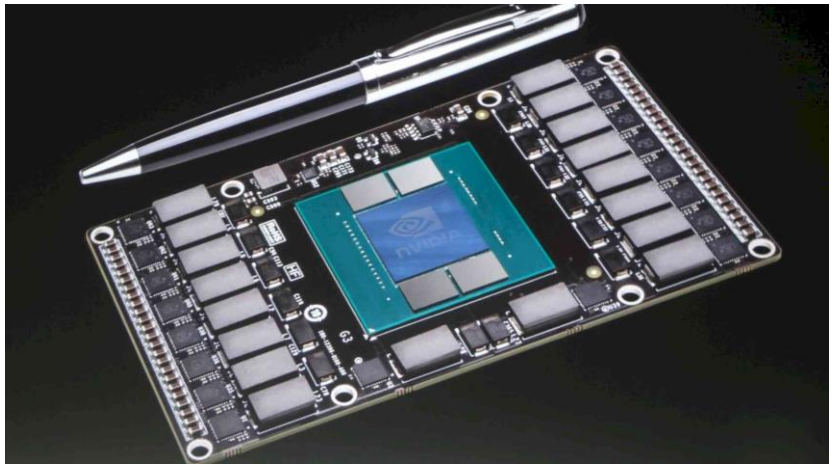


- 650,000 neurons, 0.63×10^9 synapses
- Image Net LSVRS-2010 benchmark set
- 1.2 M images; 1,000 classes
- error rates: top-1 37.5%, top-5 17%

Bottleneck: massive number of dot product (vector-by-matrix) computations between analog inputs and analog (fixed) weights

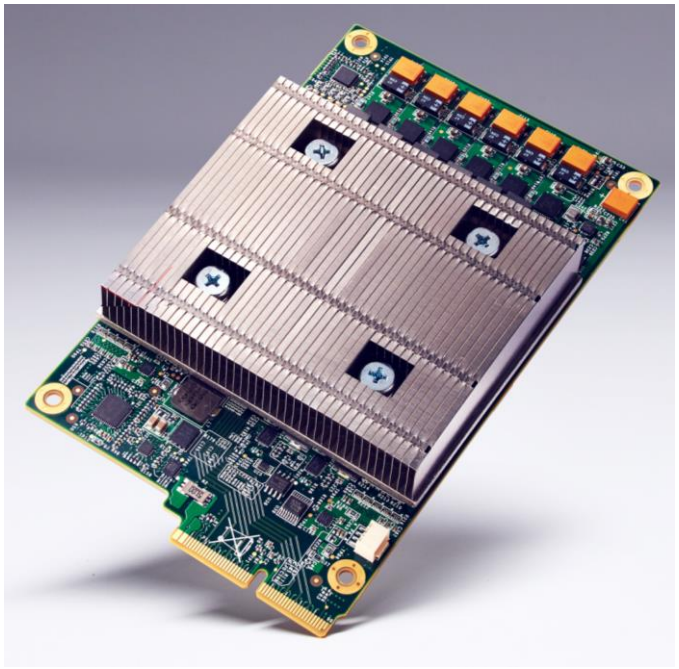
DIGITAL CIRCUITS FOR DEEP LEARNING

Nvidia's Pascal

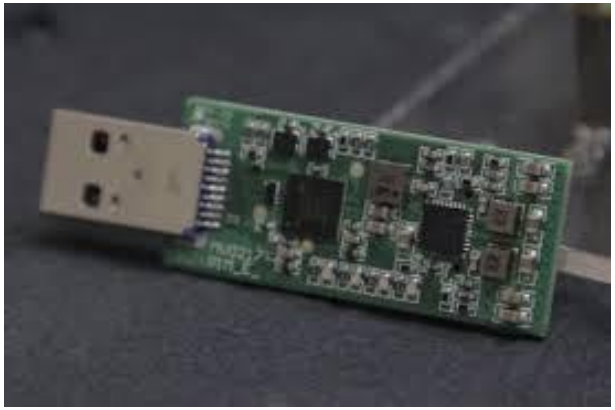


21 TFLOPS for deep learning performance

Google's Tensor Processing Unit

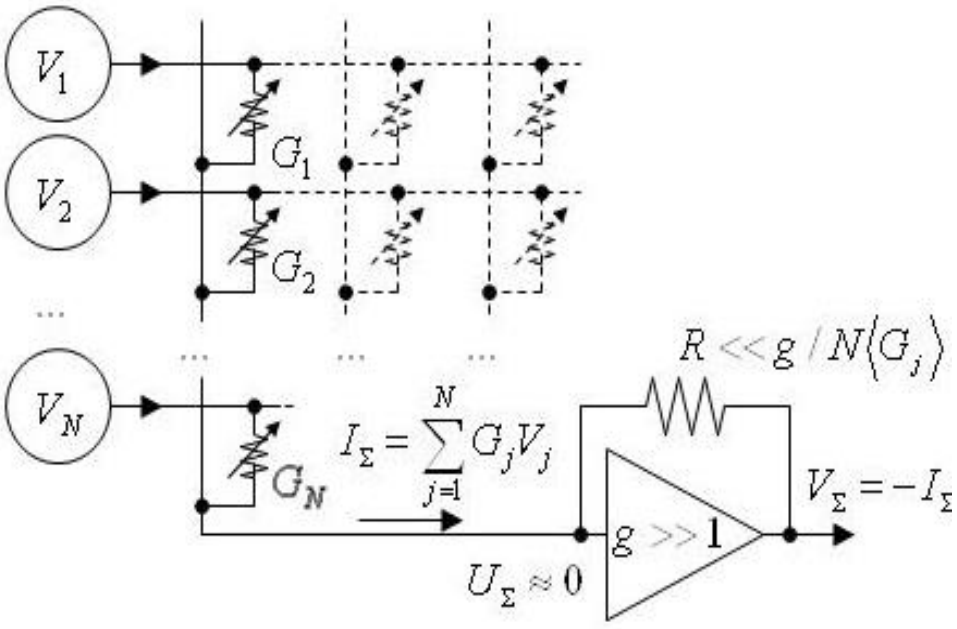


Movidius's fanthom



15 inferences /sec @ 16-bit FP precision for ImageNet @ <2W

ANALOG VECTOR-BY-MATRIX COMPUTATION



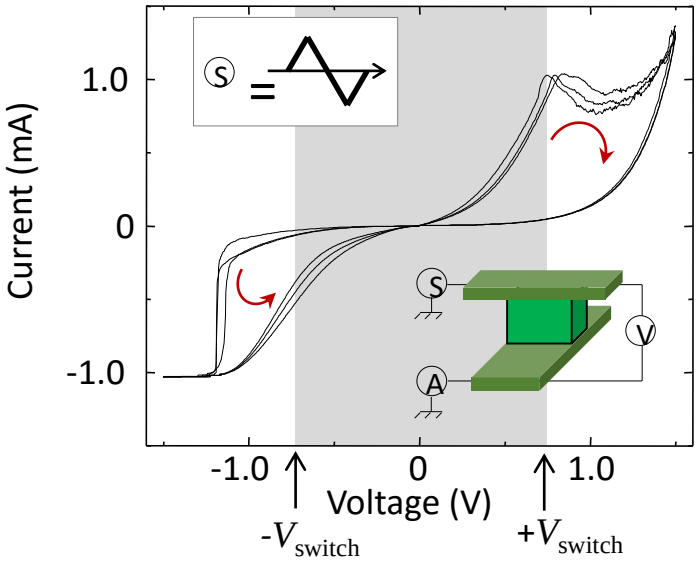
- Proposed by Carver Mead and his students 25+ years ago
- Exact analog-domain dot-product due to Ohm's and Kirchhoff's law
- No need to waste energy on memory bits movement (in-memory computing)
- Major challenge: adjustable cross-point devices
- Two very promising recent options:
  - Custom-built metal-oxide memristors
  - Redesigned commercial NOR flash
- Other (not discussed) options: phase change, ferroelectric, and magnetic devices

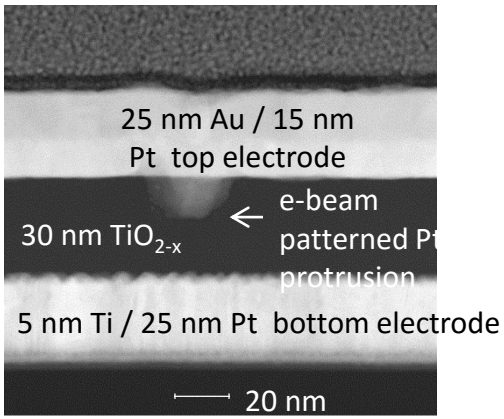
	Digital				Analog				Human Brain
	CPU 2.66 GHz 45 nm	GPU 1 GHz 33 nm	FPGA 200 MHz 40 nm	ASIC 400 MHz 65 nm	NOR ESF-1 180 nm	NOR ESF-3 55 nm	2D memristors 200 nm	3D memristors 10 nm	
Time (s)	~8×10^{-3}	~3×10^{-4}	~1.5×10^{-4}	~5×10^{-5}	~2×10^{-6}	~7×10^{-7}	~5×10^{-8}	~10^{-8}	~3×10^{-2}
Power (W)	~30 to 40	~40	~10	~3	~1	~1	~1	~0.1	~10^{-5}
Energy (J)	~3×10^{-1}	~10^{-2}	~10^{-3}	~10^{-4}	~2×10^{-6}	~7×10^{-7}	~5×10^{-8}	~10^{-9}	~3×10^{-7}

Strukov et al., DRC'16

MEMRISTORS

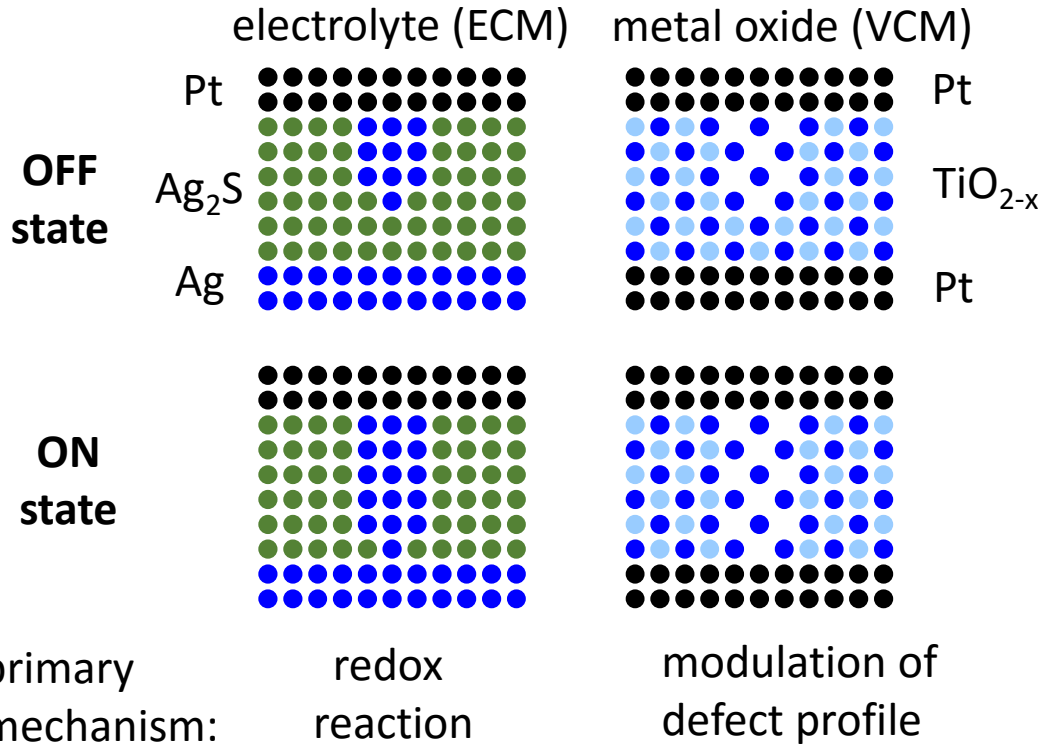
- **Typical I-V for $Pt/TiO_{2-x}/Pt$ devices**





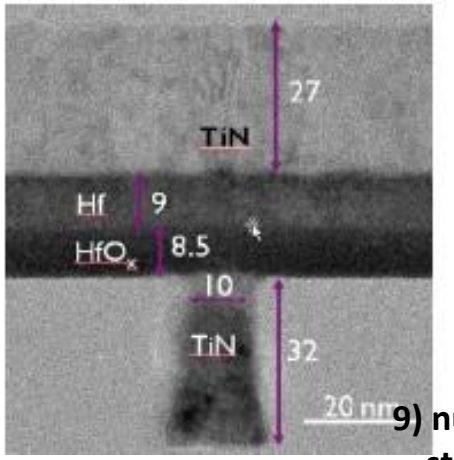
Alibart et al., Nature Comm, 2013

- **Two major types of memristors**

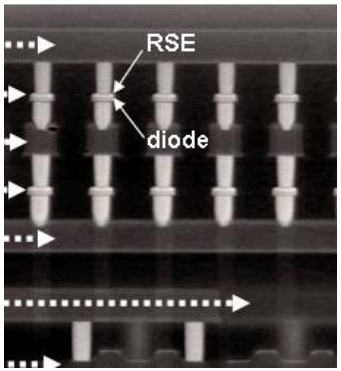


- Analog switching: Any state between ON and OFF
- Strongly (superexp) nonlinear switching dynamics
- Gray area = no change
- Memory state defined as current measured within gray area

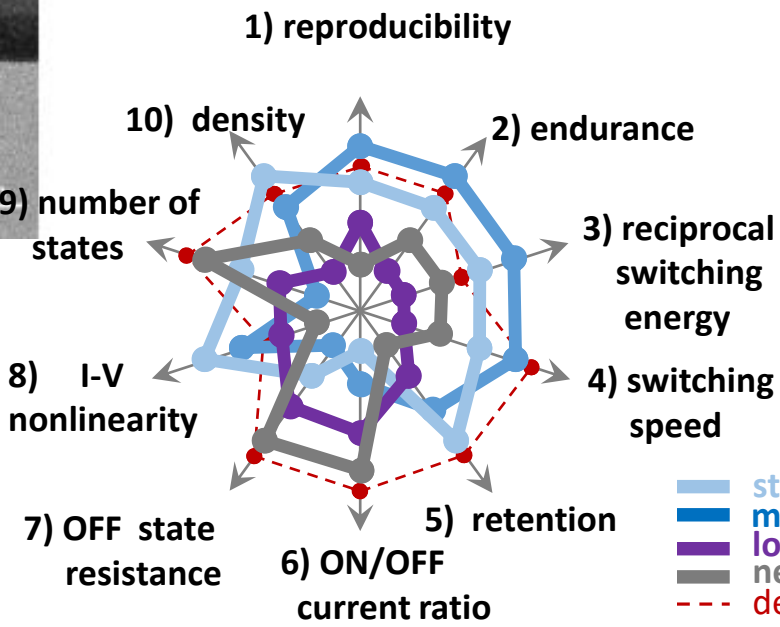
STATE-OF-THE-ART PERFORMANCE

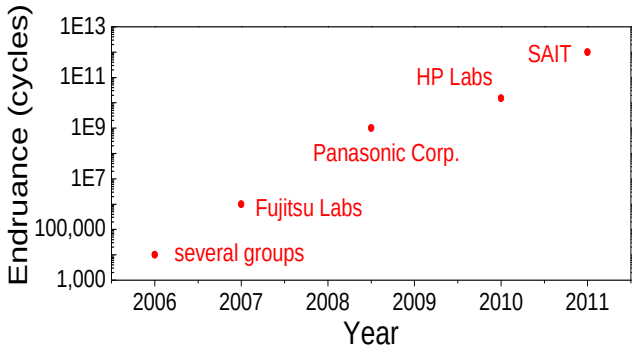


Govoreanu,et all IEDM, 2012

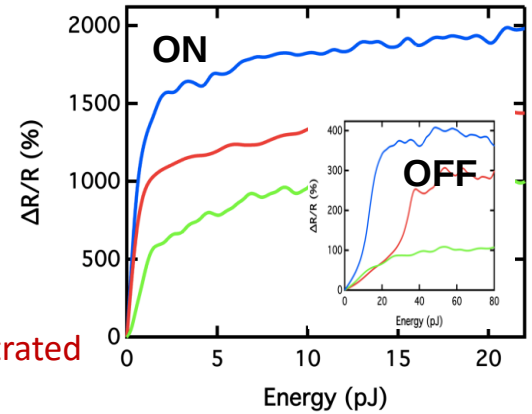


Kawahara et al. Panasonic, 2012

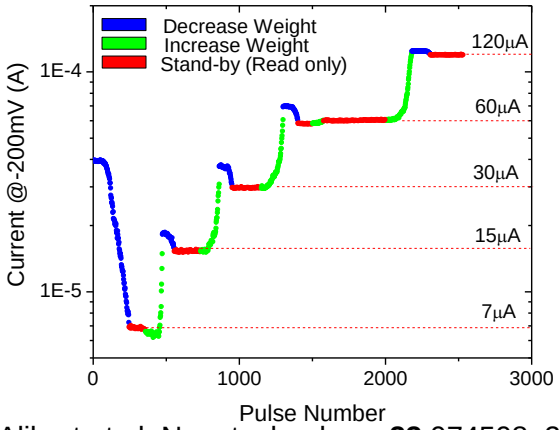




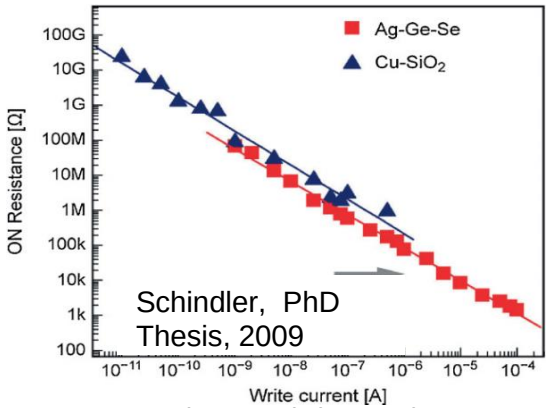
J. Yang, DBS, and D. Stewart Nature Nano **8** 13-24 (2013)

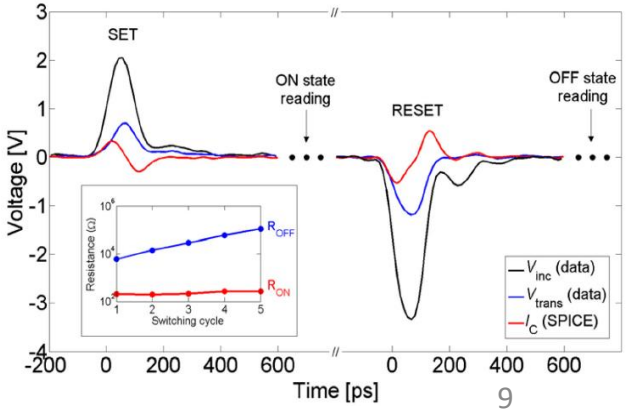


Strachan et al, Nanotechnology **22** 505402 2011



Alibart et al, Nanotechnology **23** 074508, 2012

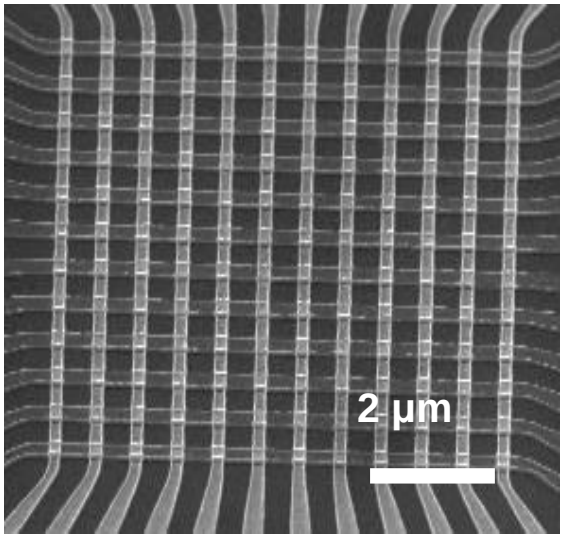




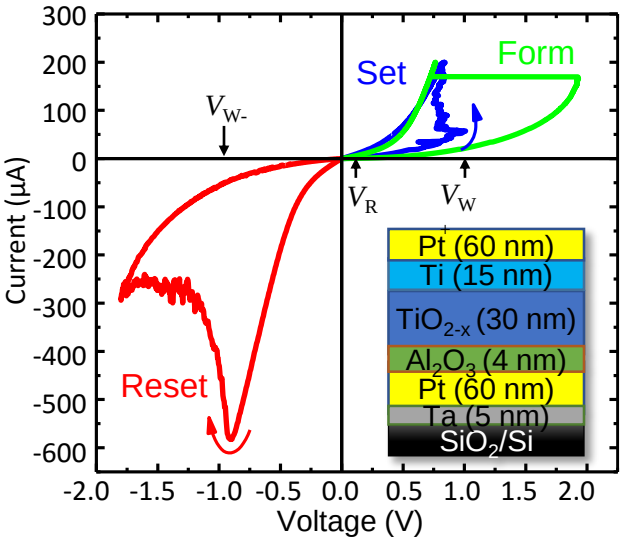
Torrezan et al, Nanotechnology **22** 485203 2011

PASSIVE MEMRISTIVE CROSSBAR CIRCUIT

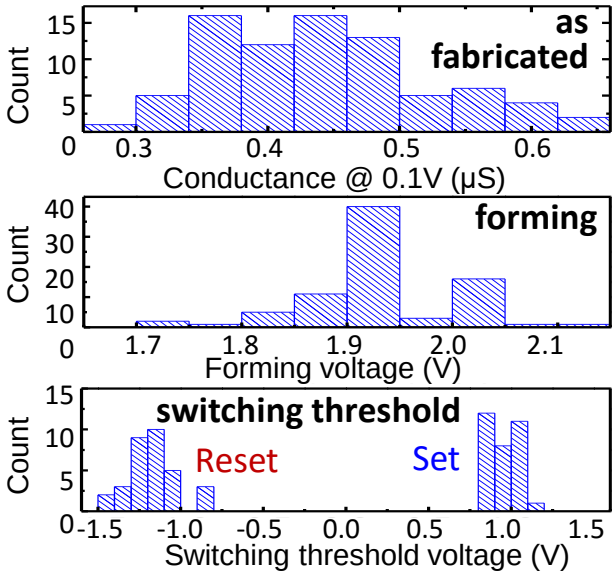
- **Crossbar circuit**



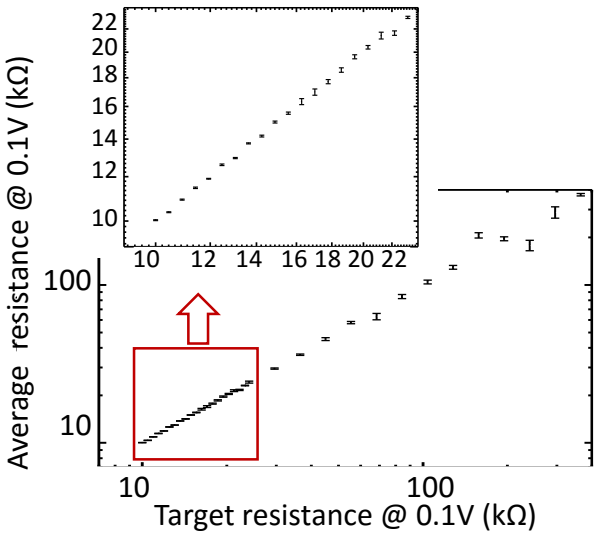
- **Typical I-Vs**

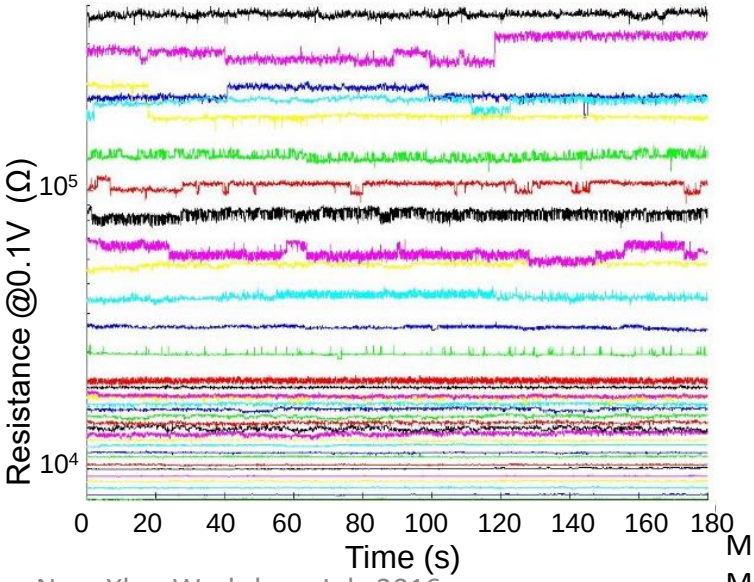


- **Statistics**



- **Analog properties and state tuning**



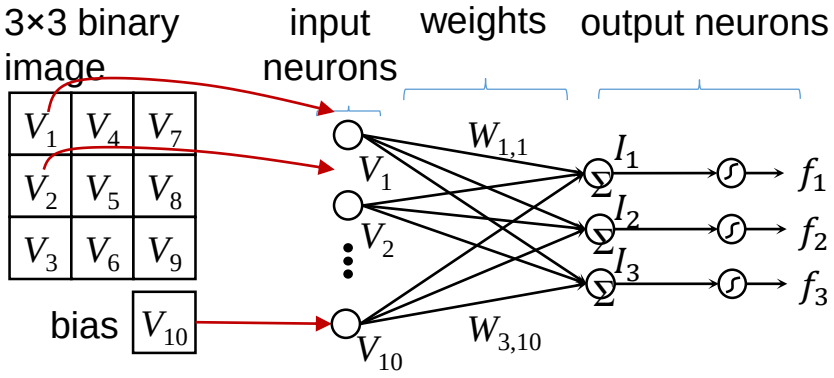


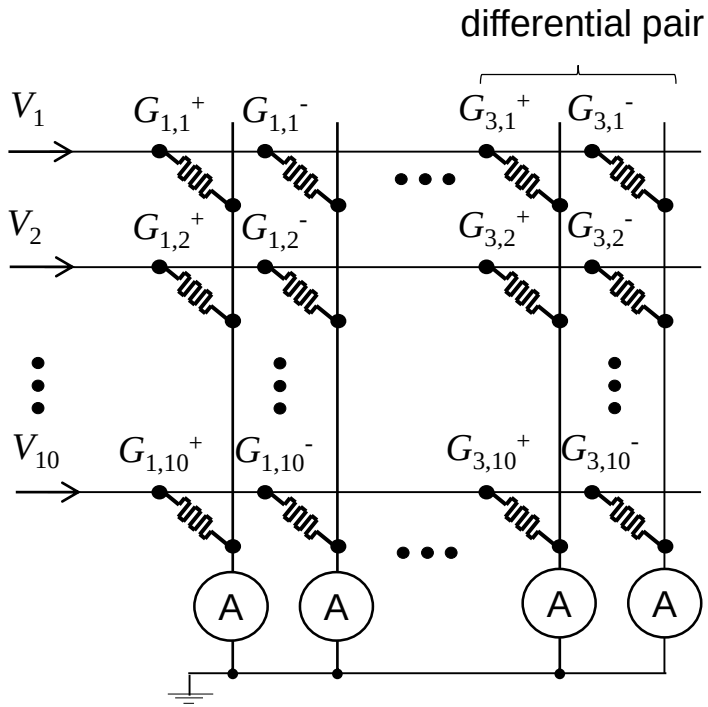
- **Major features**

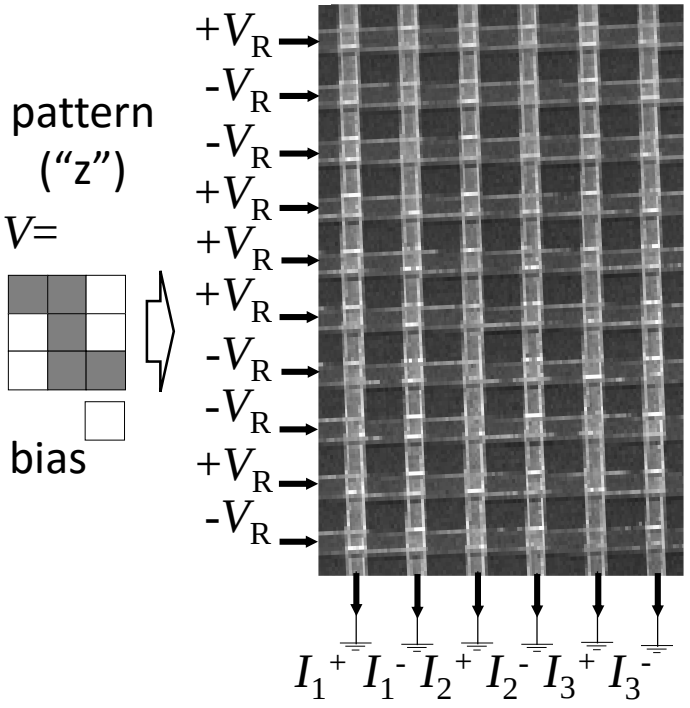
- 0T1R
- 200 nm wide lines
- Al_2O_3 and TiO_{2-x} by sputtering
- Very uniform (~17%) norm. RMS of R@0.1V for 8x10 virgin array
- >500K stress pulses without much degradation

M. Prezioso et al., *Nature* May 2015
M. Prezioso et al., *IEDM'15*

CLASSIFIER OPERATION (INFERENCE)



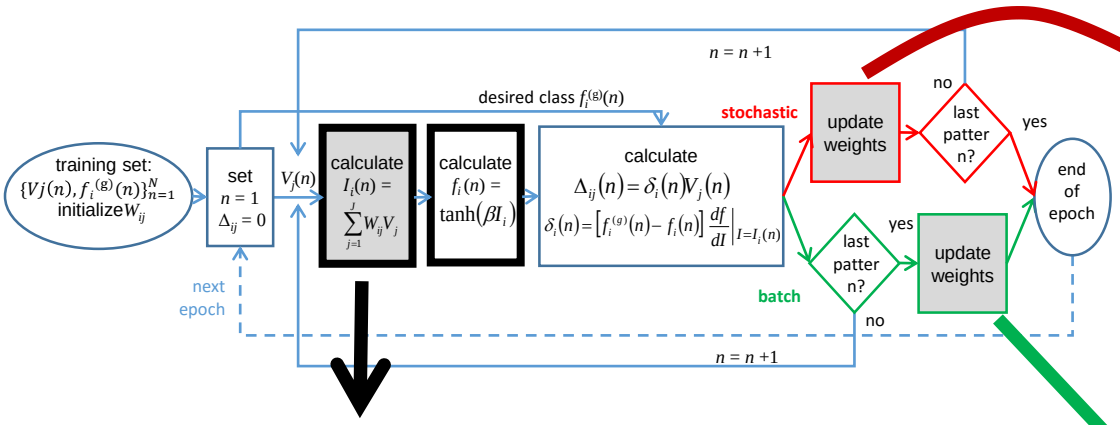




- Neurons functionality (opamp) is emulated in software
- Differential pair of memristors per weight

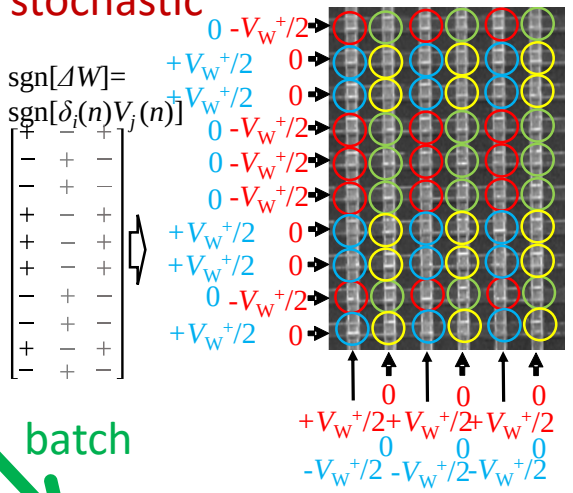
M. Prezioso et al., *Nature* May 2015
M. Prezioso et al., *IEDM'15*

CLASSIFIER IN-SITU TRAINING (WEIGHT UPDATE)

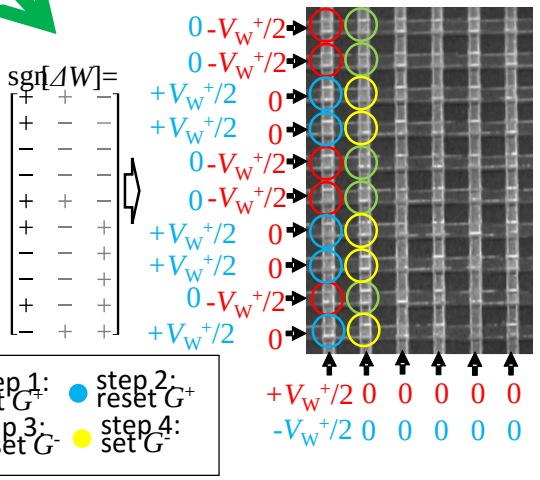


- **Weight update**

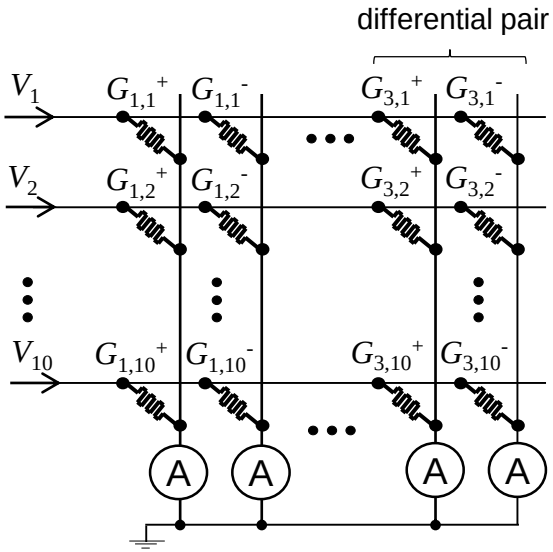
stochastic

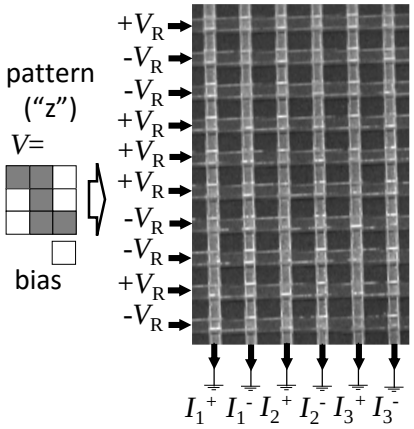


batch



- **Analog vector-by-matrix multiplier**





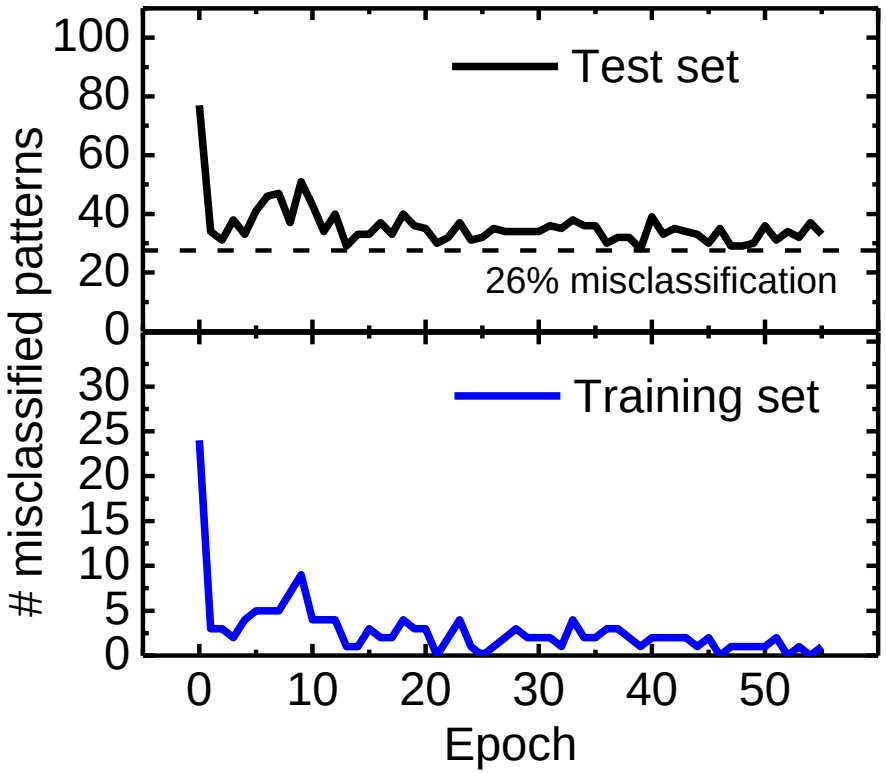
- Neurons functionality (opamp) is emulated in software
- Differential pair of memristors per weight

- Half-biasing technique
- One column at a time (fully parallel possible with stochastic training mode)

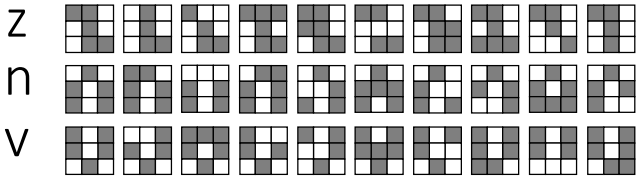
M. Prezioso et al., *Nature* May 2015
M. Prezioso et al., *IEDM'15*

EXPERIMENTAL RESULTS

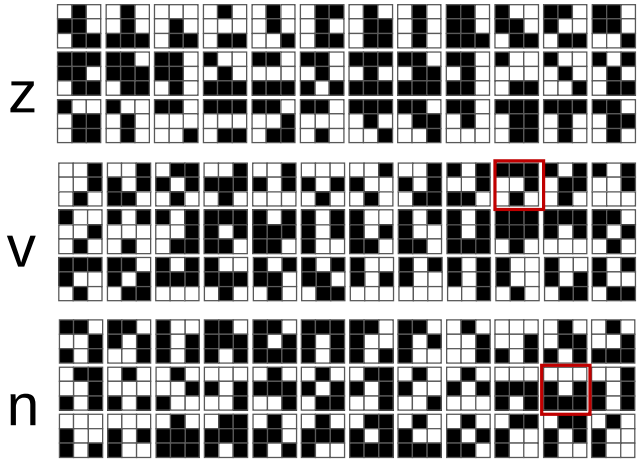
- **Classification performance (batch)**



- **Batch training and …**



- **… test sets**

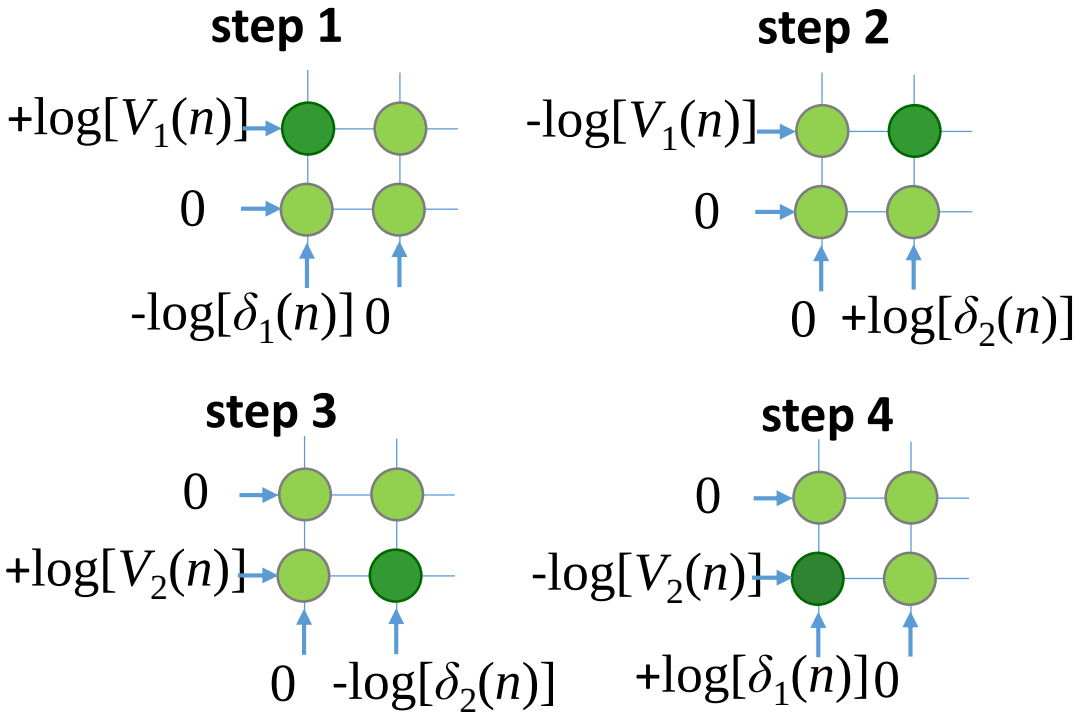


Batch Manhattan rule in-situ training:

- Trained on the original training set
- Test set formed by flipping two pixels
- **Perfect classification for multiple runs on training set**
- Perfect classification on test set hardly possible (e.g. see pattern highlighted with red)

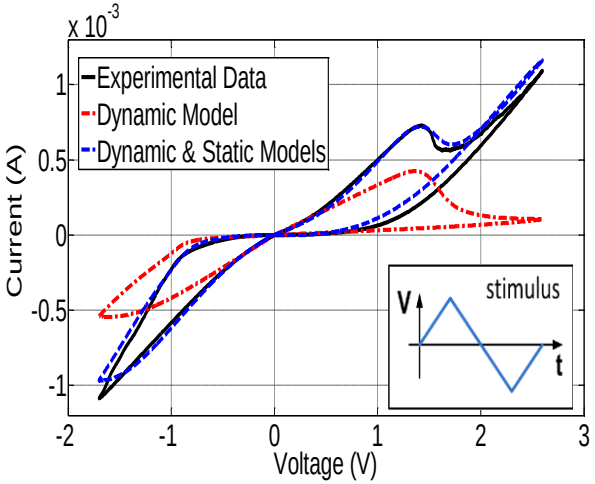
M. Prezioso et al., *Nature* May 2015
M. Prezioso et al., *IEDM'15*

OVERCOMING NONLINEAR SWITCHING KINETICS



MODELING OF LARGE-SCALE CLASSIFIERS

- **Experimentally-verified memristor device models**

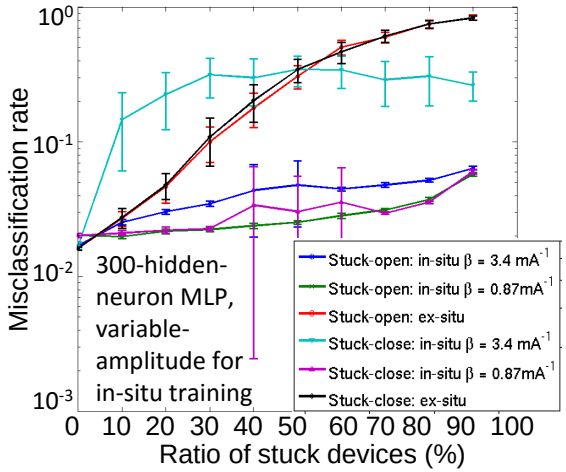


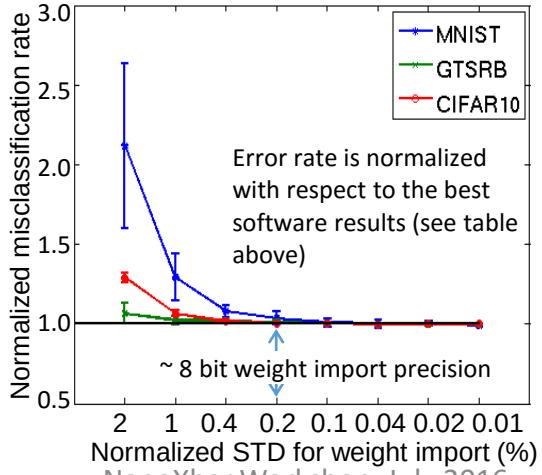
F. Merrikh Bayat et al., *Applied Physics A*, 2015

- **Classification performance results for large-scale deep learning convolutional neural networks**

Data set	Software		Xbar in-situ (var. ampl.)		Xbar ex-situ 2%		Xbar ex-situ 0.2%	
	best	average	best	average	best	average	best	average
MNIST	0.40	0.47±0.05	0.4	0.48±0.024	0.61	0.89±0.22	0.41	0.42±0.01
GTSRB	1.36	1.53±0.18	1.26	1.56±0.27	1.42	1.56±01	1.46	1.47±0.01
CIFAR10	15.63	15.91±0.22	15.67	15.87±0.22	19.77	20.29±0.43	15.5	15.8±0.01

- **Performance sensitivity to defects and ex-situ training precision**

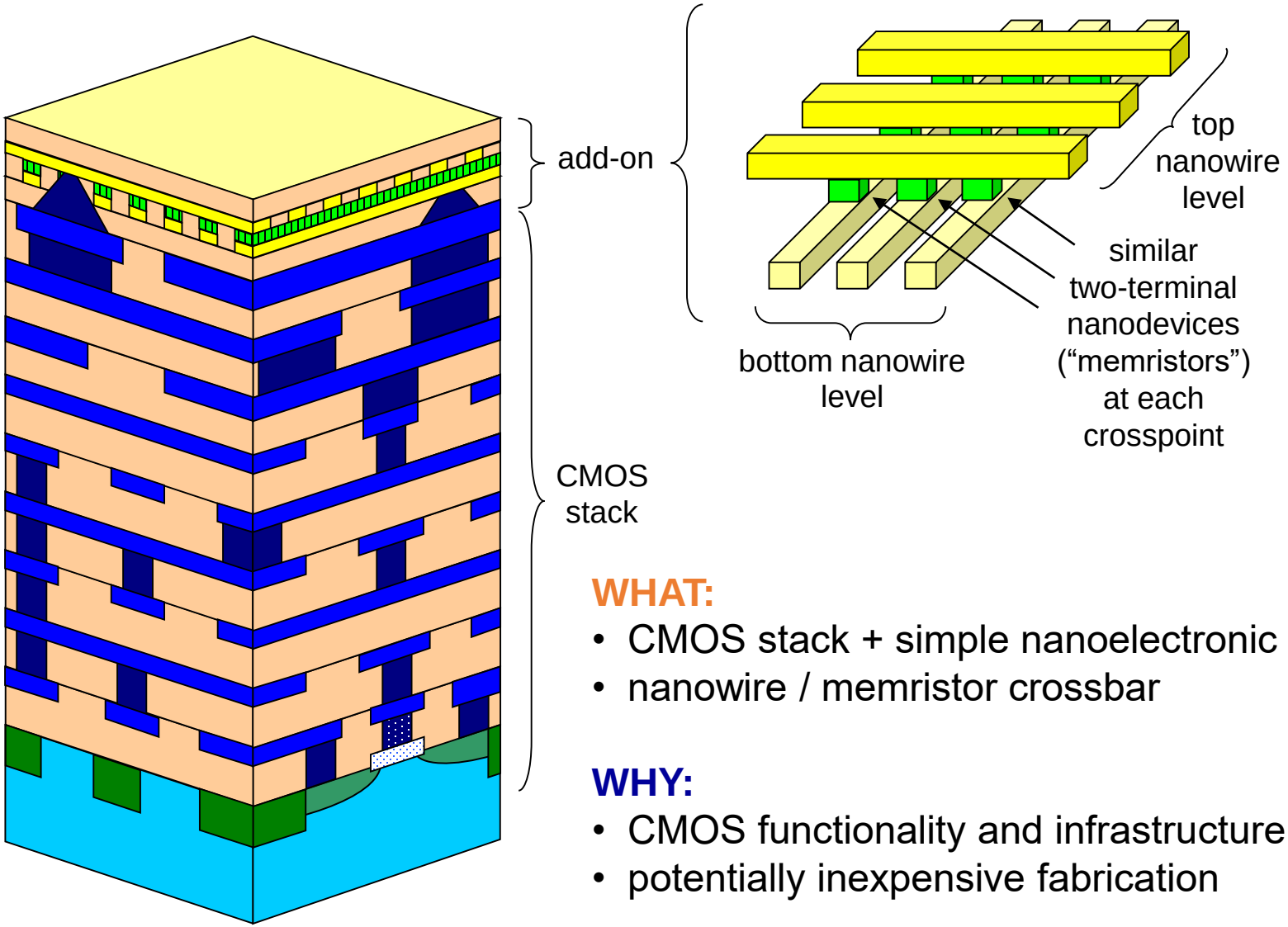




Main result: Comparable to the state-of-the-art classification performance for MNIST, GTSRB, and CIFAR benchmarks when using accurate models of hardware

I. Kataeva et al., *IJCNN'15*,
M. Prezioso et al., *IEDM'15*

CMOS/NANO HYBRIDS: THE IDEA



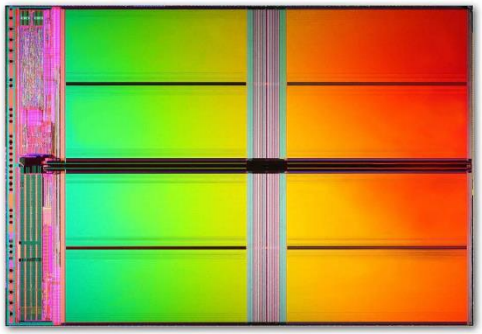
WHAT:

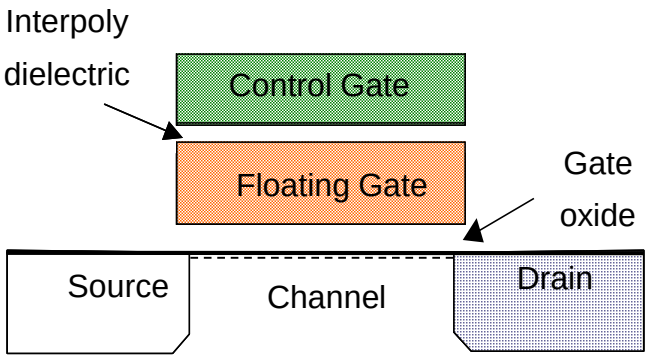
- CMOS stack + simple nanoelectronic add-on
- nanowire / memristor crossbar

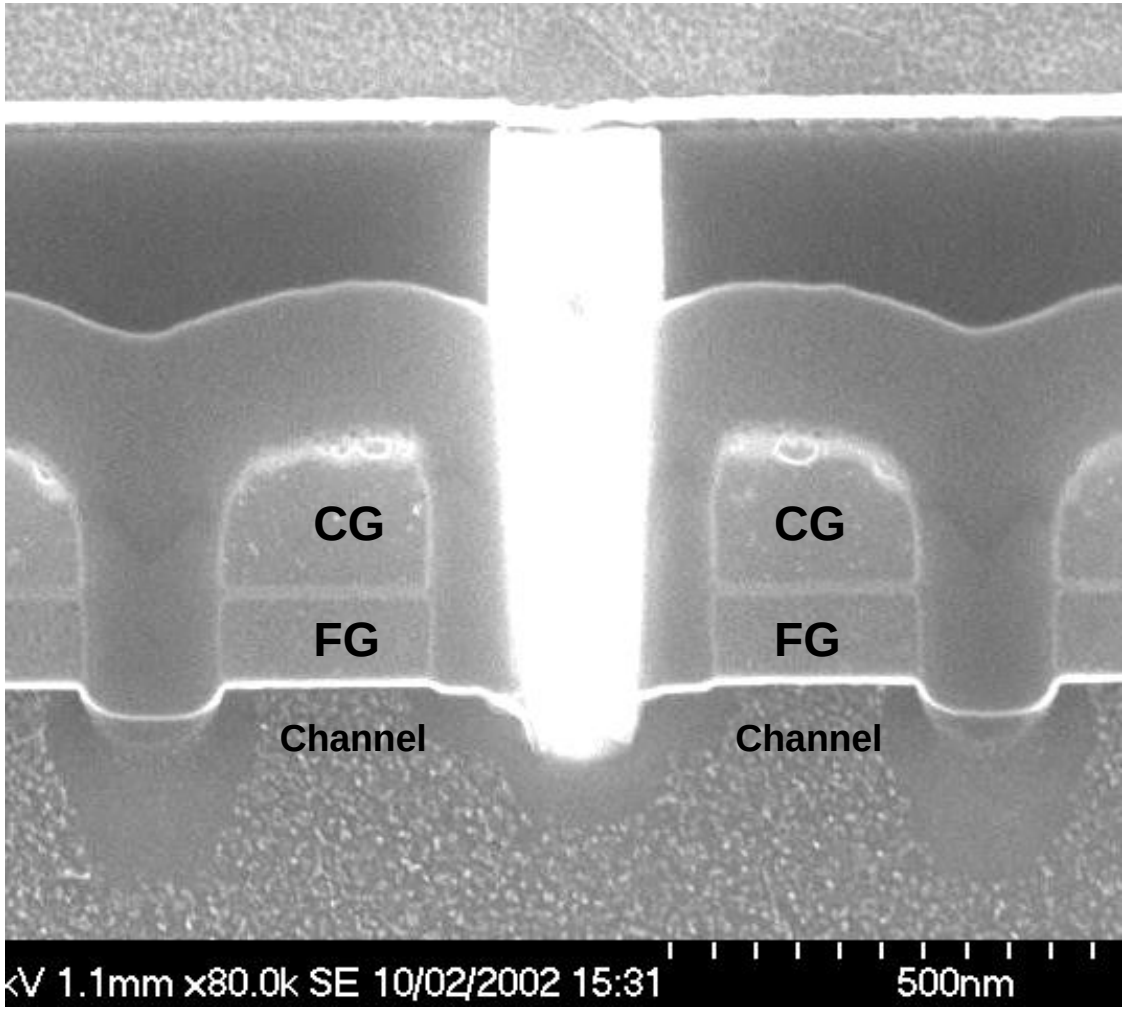
WHY:

- CMOS functionality and infrastructure intact
- potentially inexpensive fabrication

NVM ("FLASH") MEMORY TECHNOLOGY



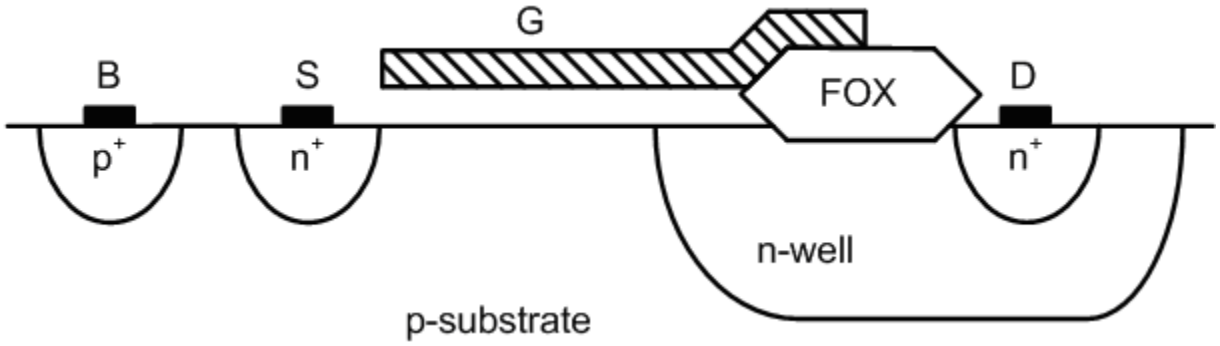




NVM CELLS FOR ANALOG APPLICATIONS

(from late 1990s: C. Mead, C. Diorio, P. Hasler,…)

Example: “extended drain” NMOS structure



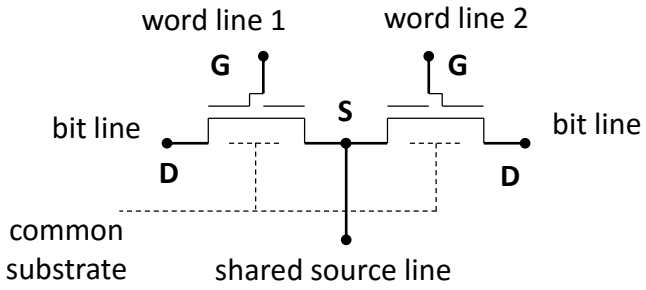
Hasler’s group at Georgia Tech (http://www-old.me.gatech.edu/mist/gokce.htm)

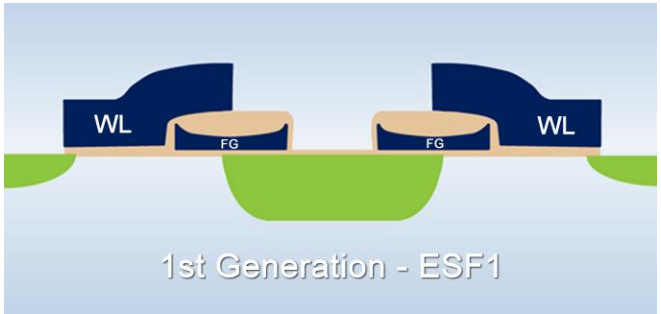
Chip built	Process node (nm)	Die area (mm^2)	No of synapses	Synapse area (μm^2)	Syn density	Synapse storage resolution and complexity
GT neuron1d (Brink et al., 2012)	350	25	30,000	133	**1088**	>10 bit, STDP
FACETs chip (Schemmel et al., 2006, 2008b)	180	25	98,304	108	3338	4 bit register
Stanford STDP	250	10.2	21,504	238	3810	STDP, no storage
INI chip (Indiveri et al., 2006)	800	1.6	256	4495	7023	1 bit w/learning dynam
ISS + INI chip (Camilleri et al., 2007)	350	68.9	16,384	3200	26,122	2.5 w/learning dynam

Bold value indicates synapse density as the synapse area normalized by the square of the process node.

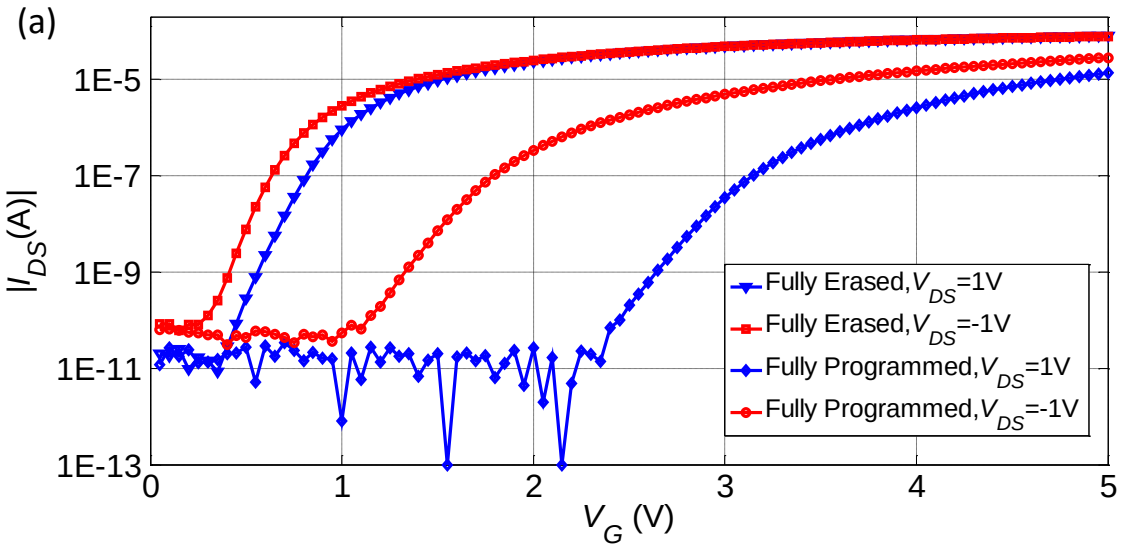
J. Hasler and B. Marr (2013)

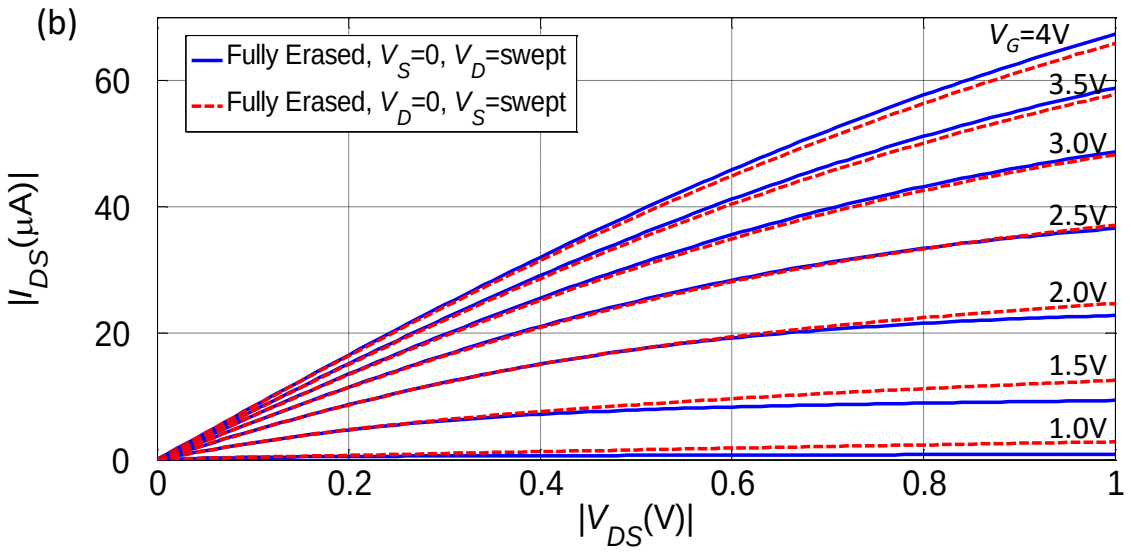
SILICON STORAGE TECHNOLOGY, INC. (SST): ESF1





Output current as a function of applied voltages:

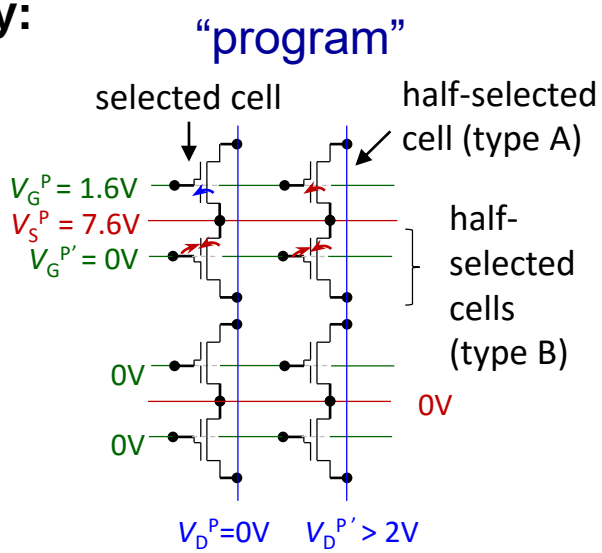


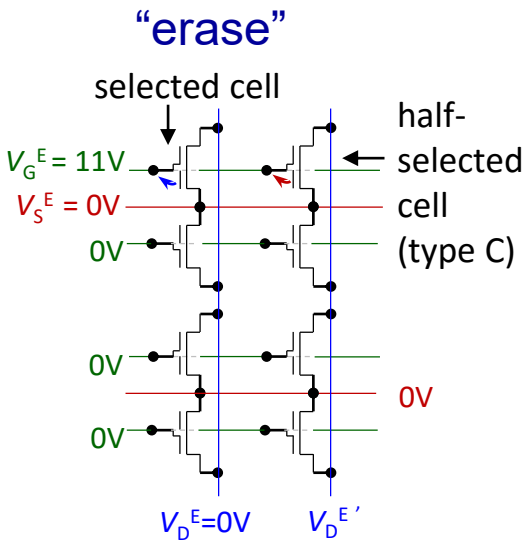


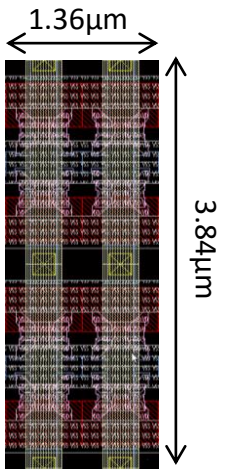
F. Merrikh-Bayat *et al.* (2015)

FLASH ARRAY REDESIGN FOR ANALOG APPLICATIONS

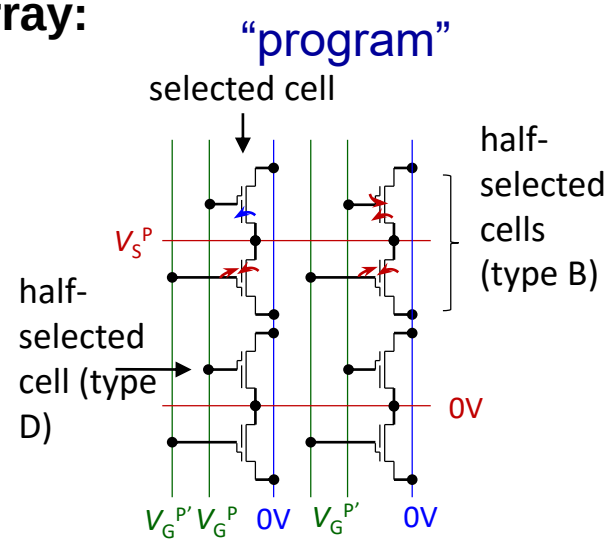
Old array:

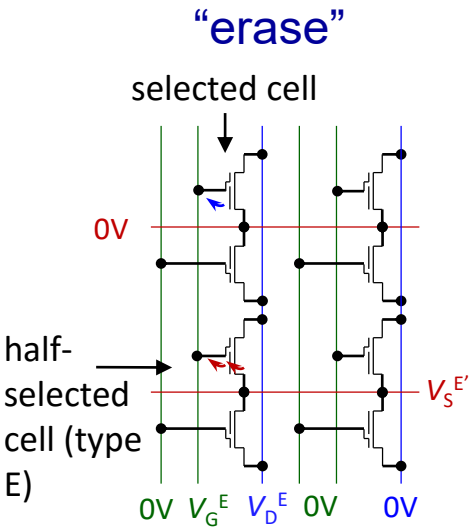


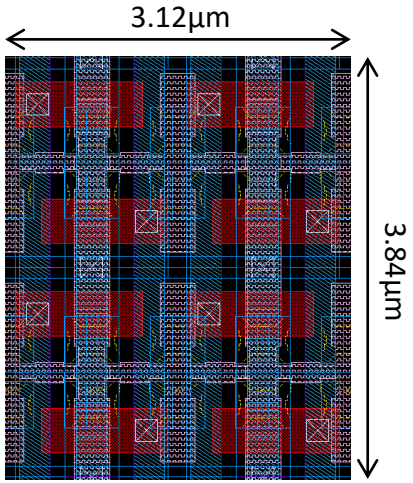




New array:



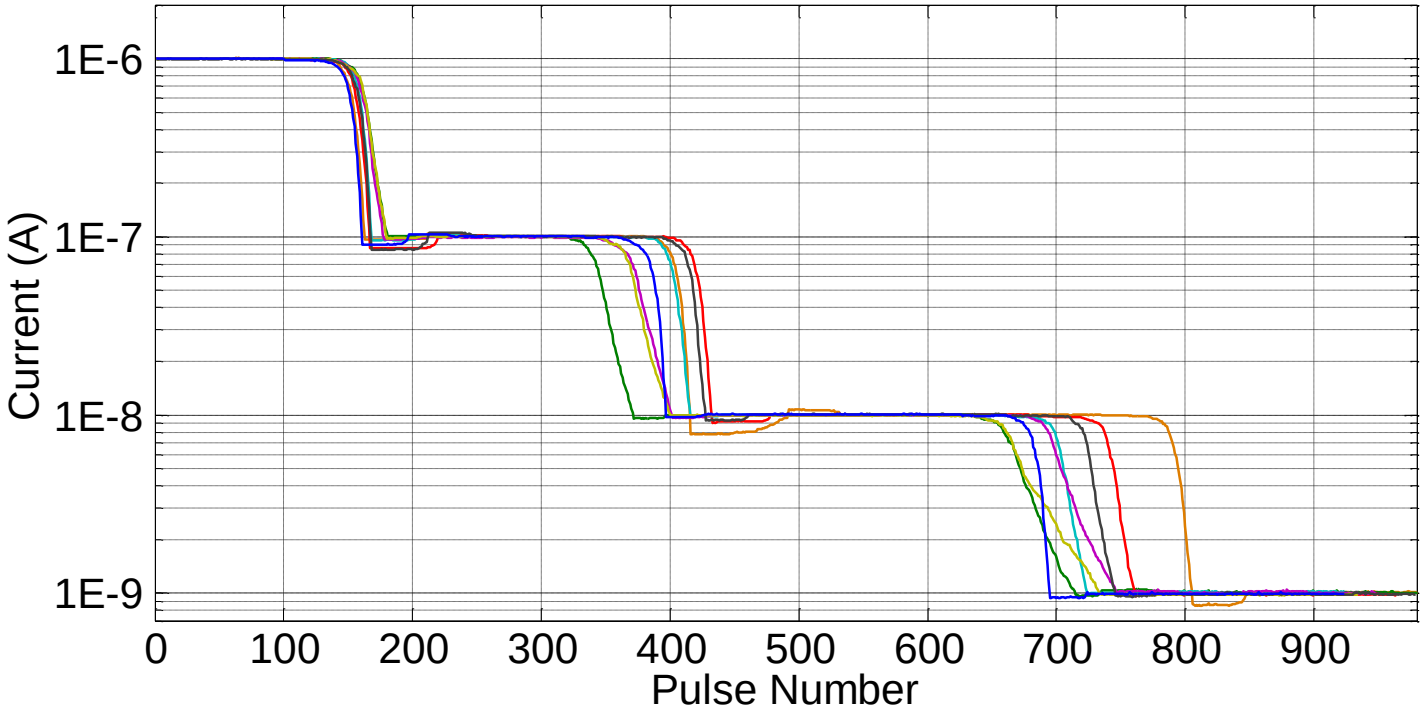


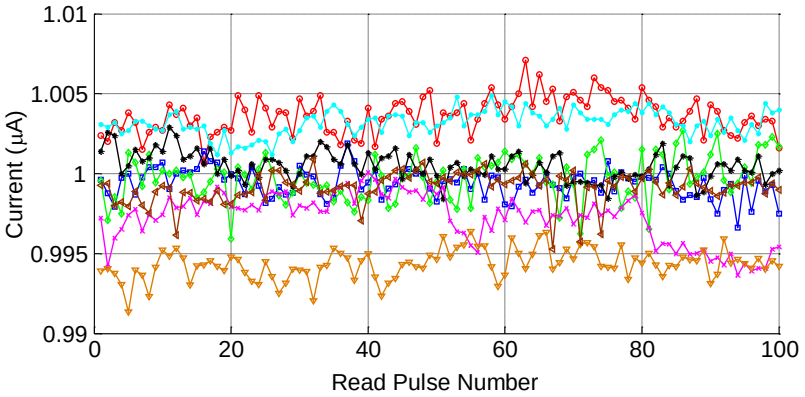


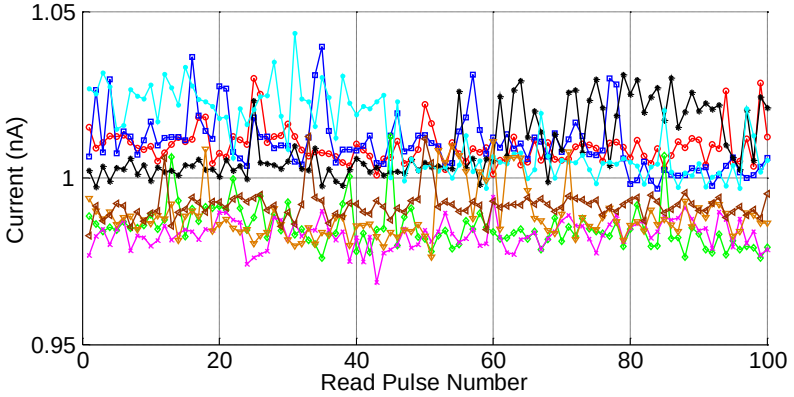
<100F^2 area per synapse

F. Merrikh-Bayat *et al.* (2015)

TUNING (OF EACH CELL!) TO PRE-SET VALUES

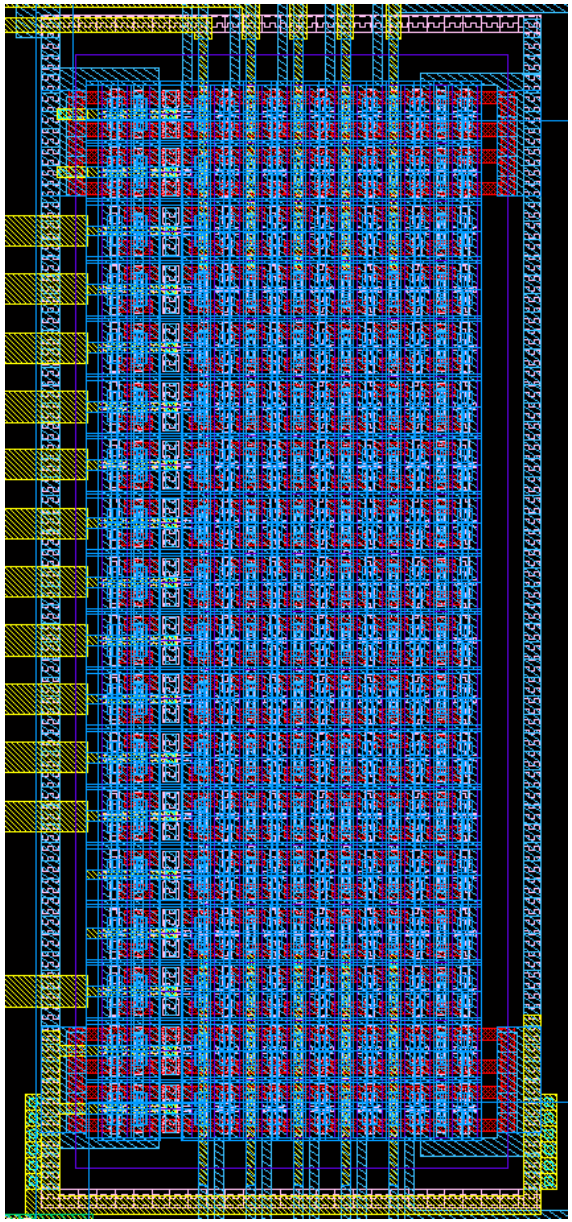


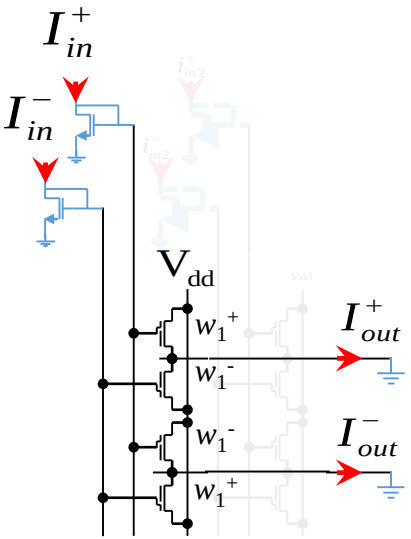


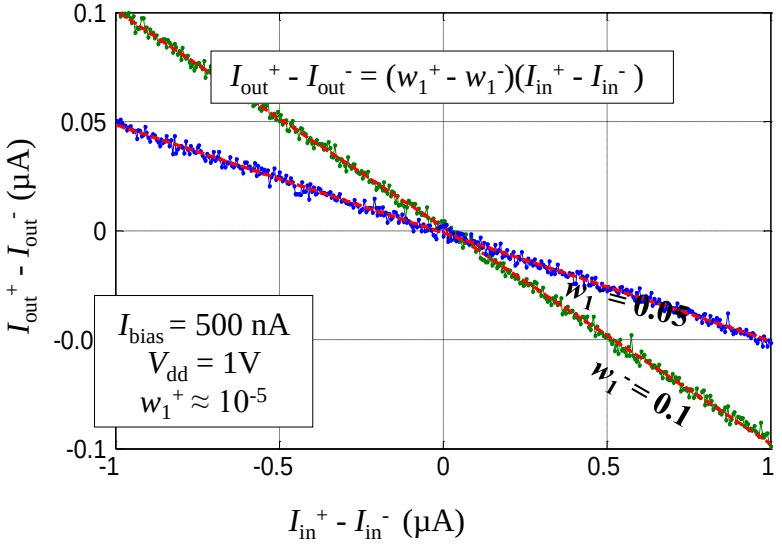


F. Merrikh-Bayat *et al.* (2015)

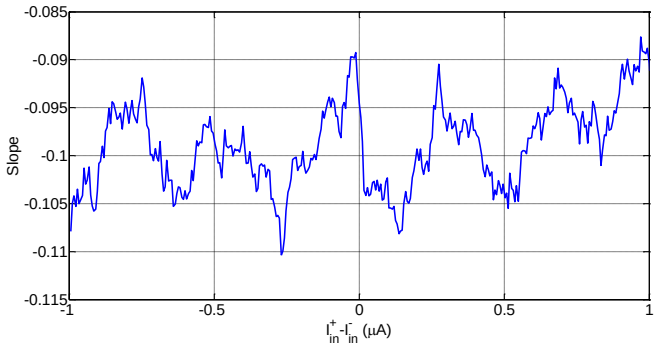
VECTOR-BY-MATRIX MULTIPLIER (VMM) DEMO







VMM linearity:



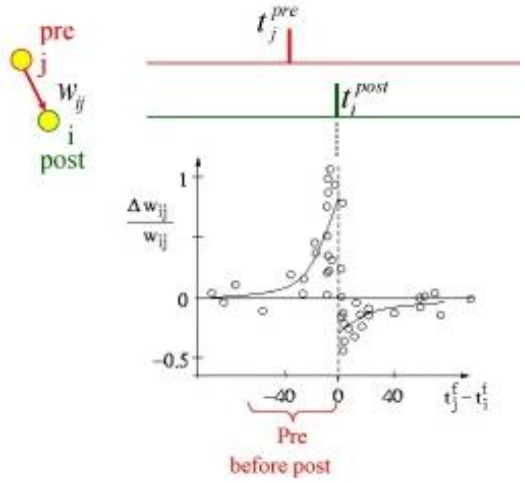
Transfer to a better NVM 2D technology would bring synapse area well below ~ 1 µm^2.

F. Merrikh-Bayat *et al.* (2015)

SPIKING NEURAL NETWORKS

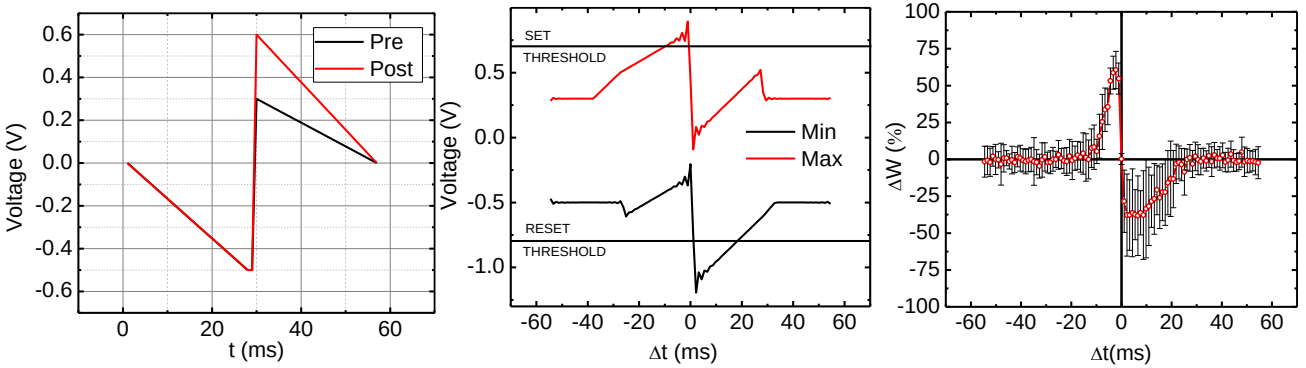
Motivation

- Richer functionality (spatial and temporal processing) and better energy efficiency of spiking networks as compared to firing rate
- Local (Hebbian) training → more efficient hardware
- Essential feature to demonstrate: Spike-timing dependent plasticity (STDP)



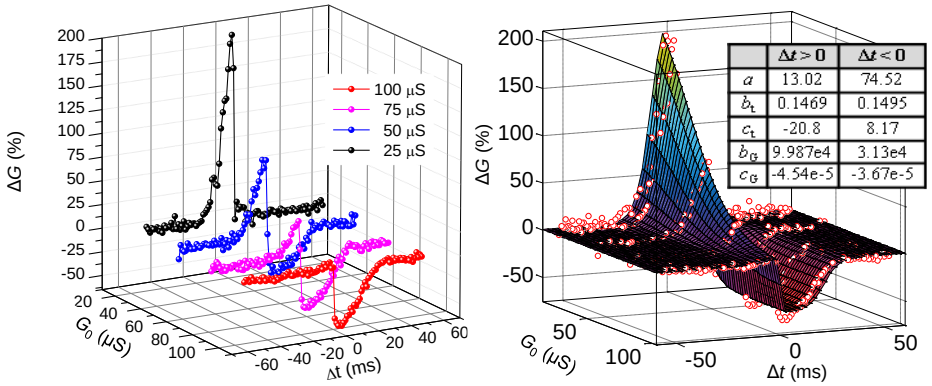
- Three STDP windows demonstrated using crossbar
- The most accurate STDP demonstration to date

Experimental demonstration of STDP



Experimentally-verified analytical model STDP

M.Prezioso et al., *Nature Scientific Report*, 2016



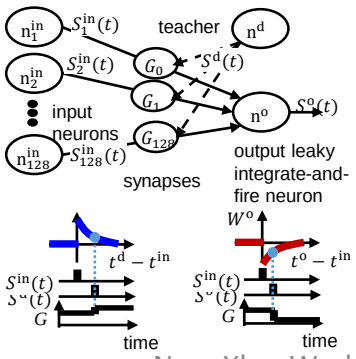
	$\Delta t > 0$	$\Delta t < 0$
a	13.02	74.52
b_t	0.1469	0.1495
c_t	-20.8	8.17
b_G	9.987e4	3.13e4
c_G	-4.54e-5	-3.67e-5

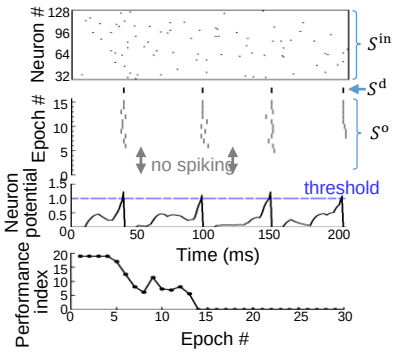
$$\Delta G = \Lambda_t \Lambda_G,$$

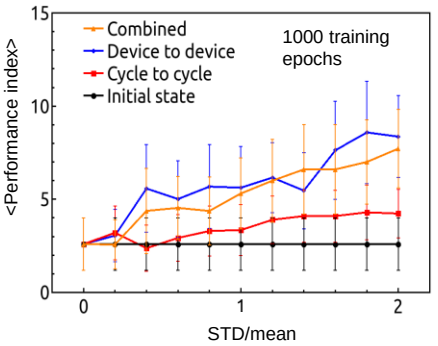
$$\Lambda_t \equiv \begin{cases} a^+\left\{\tanh\left[b_t^+\left(\Delta t + c_t^+\right)\right]-1\right\}, \Delta t > 0, \\ a^-\left\{\tanh\left[b_t^-\left(\Delta t + c_t^-\right)\right]+1\right\}, \Delta t < 0, \end{cases}$$

$$\Lambda_G \equiv \begin{cases} 1+\tanh\left[b_G^+\left(G_0 + c_G^+\right)\right], \Delta t > 0, \\ 1-\tanh\left[b_G^-\left(G_0 + c_G^-\right)\right], \Delta t < 0. \end{cases}$$

Simulation of memristor-based spiking neural networks







M. Prezioso et al., *ISCAS* 2016

SUMMARY

- Emerging nonvolatile memories enable (for the first time?) efficient analog neural network implementations and could challenge human brain in energy efficiency and speed
  - Experimental demonstration of key hardware block for both memristor and flash-based artificial neural networks
  - Small scale demonstrations of firing-rate feedforward/recurrent and spiking memristor-based artificial neural networks with comparable to state-of-the-art functional performance for large scale NVM-based networks via simulation with data-verified device models
  - Estimated >100x / >1000x improvement in energy efficiency as compared to ASICs for flash / memristor based implementations
- Need industry involvement to develop large-scale memristor circuit
  - no such issue with flash memory–based circuit

	Digital				Analog				Human Brain
	CPU 2.66 GHz 45 nm	GPU 1 GHz 33 nm	FPGA 200 MHz 40 nm	ASIC 400 MHz 65 nm	NOR ESF-1 180 nm	NOR ESF-3 55 nm	2D memristors 200 nm	3D memristors 10 nm	
Time (s)	$\sim 8\times10^{-3}$	$\sim 3\times10^{-4}$	$\sim 1.5\times10^{-4}$	$\sim 5\times10^{-5}$	$\sim 2\times10^{-6}$	$\sim 7\times10^{-7}$	$\sim 5\times10^{-8}$	$\sim 10^{-8}$	$\sim 3\times10^{-2}$
Power (W)	~30 to 40	~40	~10	~3	~1	~1	~1	~0.1	$\sim 10^{-5}$
Energy (J)	$\sim 3\times10^{-1}$	$\sim 10^{-2}$	$\sim 10^{-3}$	$\sim 10^{-4}$	$\sim 2\times10^{-6}$	$\sim 7\times10^{-7}$	$\sim 5\times10^{-8}$	$\sim 10^{-9}$	$\sim 3\times10^{-7}$

Strukov et al., DRC'16

THANK YOU!

strukov@ece.ucsb.edu

SELECTED RECENT PUBLICATIONS

- F. Merrikh-Bayat, X. Guo, M. Klachko, N. Do, K. Likharev, and D. Strukov, “Model-based high-precision tuning of NOR flash memory cells for analog computing applications”, to appear in Device Research Conference (DRC’16), Newark, DE, June 2016 (NOR flash)
- M. Prezioso, Y. Zhong, D. Gavrilov, F. Merrikh Bayat, B. Hoskins, G. Adam, K.K. Likharev, and D.B. Strukov, “Spiking Neuromorphic Networks with Metal-Oxide Memristors”, to appear in International Symposium on Circuits and Systems (ISCAS'16), Montreal, Canada, May 2016 (Memristor spiking neural networks)
- M. Prezioso, F. Merrikh Bayat, B. Hoskins, K. Likharev, and D. Strukov, "Self-adaptive spike-time-dependent plasticity of metal-oxide memristors", Nature Scientific Reports 6, art. 21331, Jan. 2016. (Memristor spiking neural networks)
- F. Merrikh Bayat, M. Prezioso, X. Guo, B. Hoskins, D.B. Strukov, and K.K. Likharev, "Memory technologies for neural networks", in: Proc. IMW'15, Monterey, CA, May 2015, pp. 1-4. (brief review)
- F. Merrikh Bayat, X. Guo, H.A. Om'mani, N. Do, K.K. Likharev, and D.B. Strukov, "Redesigning commercial floating-gate memory for analog computing applications", in: Proc. ISCAS'15, Lisbon, Portugal, May 2015, pp. 1921-1924. (NOR flash)
- M. Prezioso, F. Merrikh Bayat, B.D. Hoskins, G.C. Adam, K.K. Likharev, and D.B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors", Nature 521, pp. 61-64, May 2015. (Memristor firing-rate MLP networks)
- X. Guo, F. Merrikh-Bayat, L. Gao, B. D. Hoskins, F. Alibart, B. Linares-Barranco, L. Theogarajan, C. Teuscher, and D.B. Strukov, "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. (Memristor recurrent networks)
- F. Merrikh Bayat, B. Hoskins, and D.B. Strukov, "Phenomenological modeling of memristive devices", Applied Physics A 118 (3), pp. 770-786, 2015. (Memristor model)
- M. Prezioso, I. Kataeva, F. Merrikh-Bayat, B. Hoskins, G. Adam, T. Sota, K. Likharev, and D. Strukov, "Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al2O3/TiO2-x/Pt memristors", IEDM'15, Dec. 2015. (Memristor firing-rate MLP networks)
- M. Payvand, A. Madhavan, M. Lastras-Montaño, A. Ghofrani, J. Rofeh, K.-T. Cheng, D. Strukov, L. Theogarajan, "A configurable CMOS memory platform for 3D-integrated memristors", in: Proc. ISCAS'15, Lisbon, Portugal, May 2015, pp. 1378-1381. (Memristor integration)
- I. Kataeva, F. Merrikh Bayat, E. Zamanidoost, and D.B. Strukov, "Efficient training algorithms for neural networks based on memristive crossbar circuits", in: Proc. IJCNN'15, Killarney, Ireland, July 2015, pp. 1-8. (Memristor firing-rate MLP networks modeling)
- F. Alibart, E. Zamanidoost, and D.B. Strukov, "Pattern classification by memristive crossbar circuits with ex-situ and in-situ training", Nature Communications 4, art. 2072, 2013 (Memristor firing-rate MLP networks)
- J.J. Yang, D.B. Strukov and D.R. Stewart, “Memristive devices for computing”, Nature Nanotechnology 8, pp. 13-24, 2013 (review)