Learning with Memristors

Shahar Kvatinsky

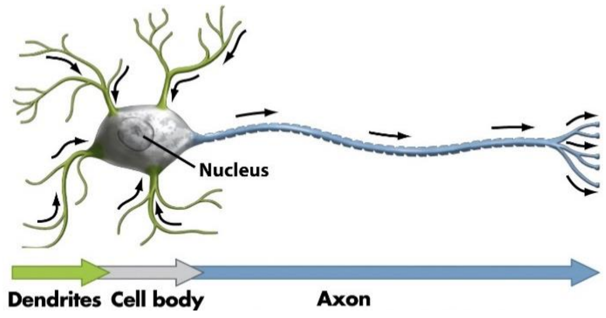
Viterbi Faculty of Electrical Engineering
Technion – Israel Institute of Technology

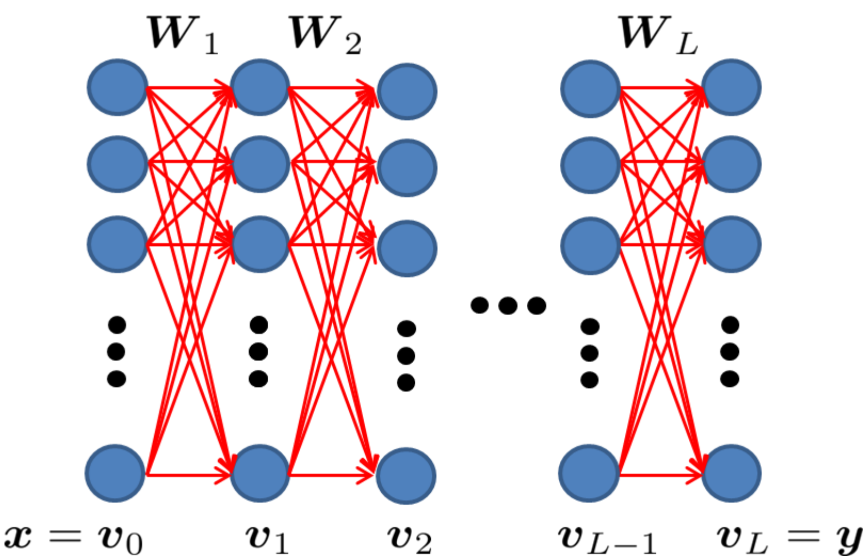
ICSEE November 2016

ASIC^2
ARCHITECTURES
SYSTEMS
INTELLIGENT COMPUTING
INTEGRATED CIRCUITS

Deep/Multilayer Neural Networks

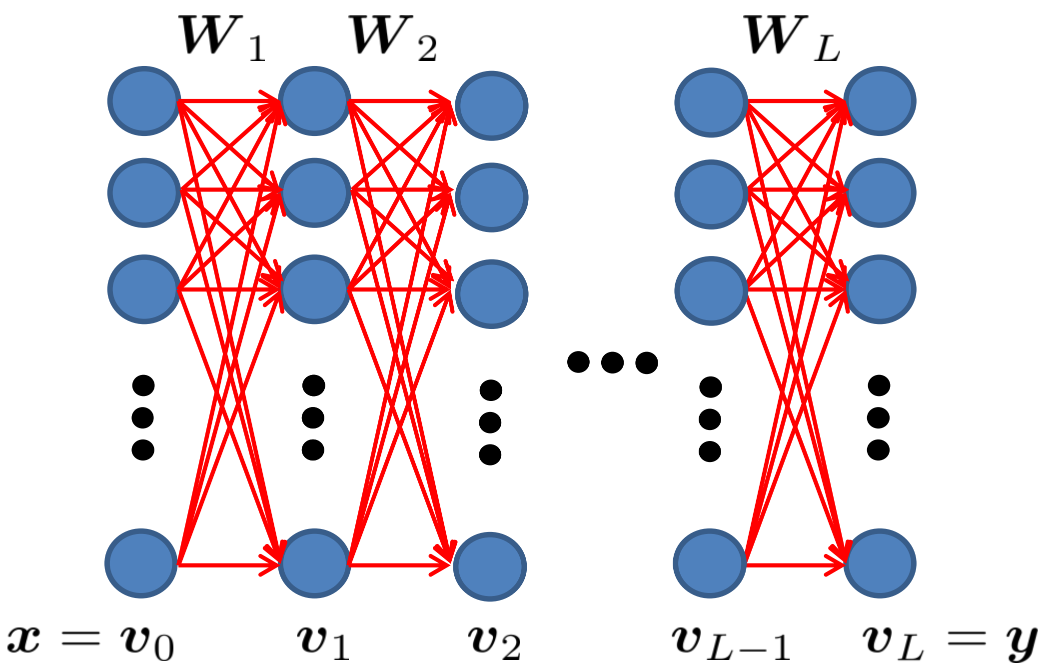
- Useful, robust, computationally intensive
- Many applications:
  - Pattern recognition
  - Natural Language Processing
  - Signal processing
  - Data Mining

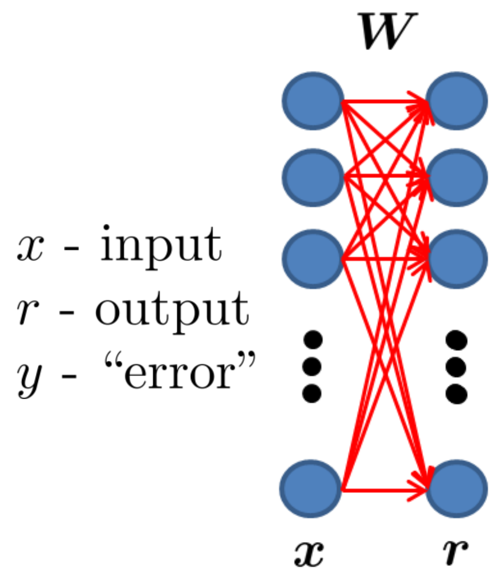




Computational Bottlenecks

$$\boldsymbol{v}_l = f_l(\boldsymbol{W}_l \boldsymbol{v}_{l-1})$$





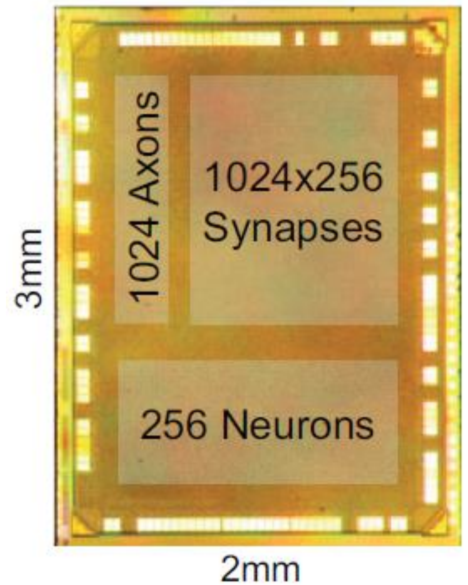
$$\Delta \mathbf{W} \propto \mathbf{y}\mathbf{x}^\top$$

- Propagation $\mathbf{r} = \mathbf{W}\mathbf{x}$ costs $O(N^2)$ operations
- Training each layer also costs $O(N^2)$ operations

Common NN Hardware

- Offline training in CPU/GPU
- Dedicated hardware (TrueNorth, DianNao, TPU)
- Online training – hard with CMOS

Design	#Transistors	Comments
Proposed design	2 (+1 memristor)	
[54]	2	Also requires UV light + Weights decay ~ minutes
[55]	6	Weights only increase (unusable)
[56, 57]	39	Must keep training
[58]	52	Must keep training
[59]	92	Weights decay ~ hours
[60]	83	Also requires a "weight unit"
[61]	150	

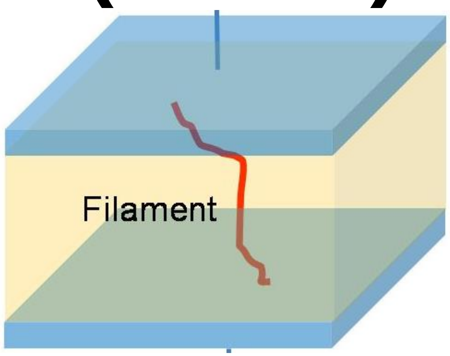


IBM TrueNorth

Memristors to the Rescue

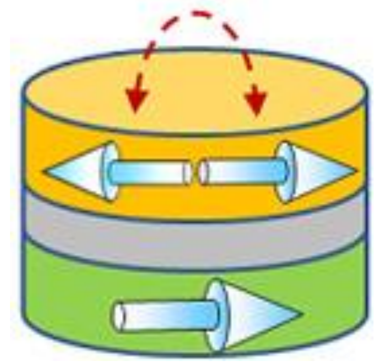
Emerging Nonvolatile Memory Technologies

Resistive RAM (RRAM)



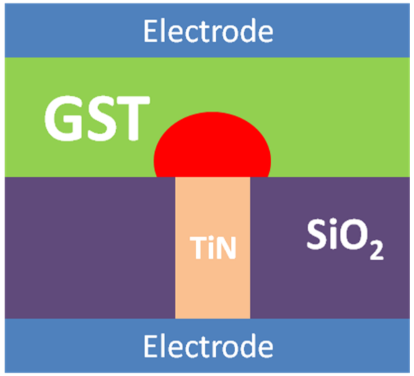
SanDisk SONY
hp SAMSUNG
Panasonic
SK hynix TOSHIBA Crossbar

STT MRAM



EVERSPIN TECHNOLOGIES
HITACHI
CROCUS Technology Blossoming future
TOSHIBA
QUALCOMM
SAMSUNG

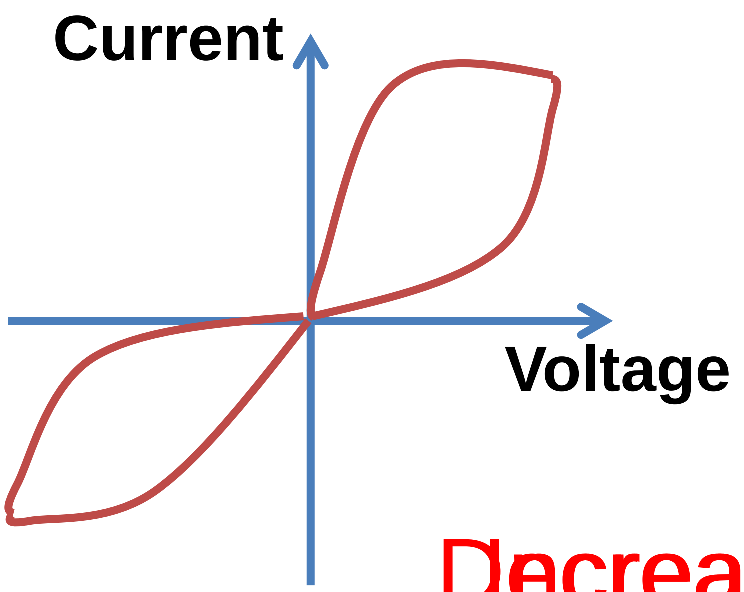
Phase Change Memory (PCM)



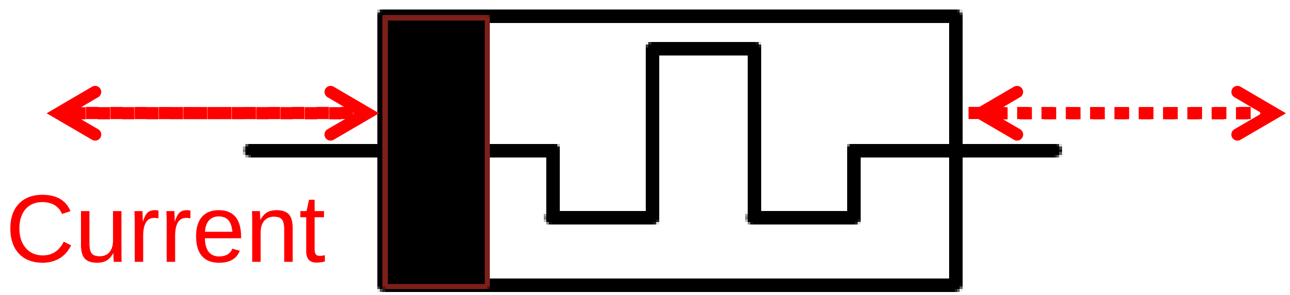
SAMSUNG intel
IBM ST
Micron SK hynix

Memristor – Memory Resistor

Resistor with Varying Resistance

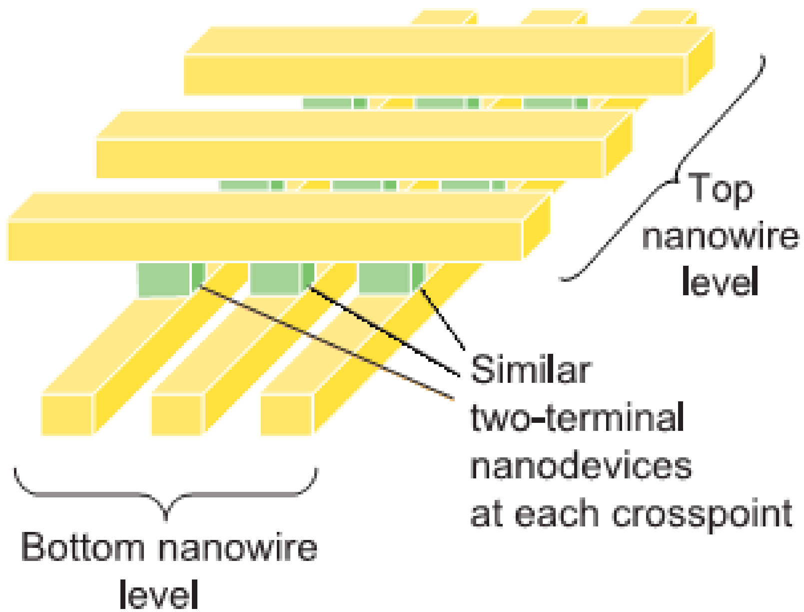


Decrease resistance

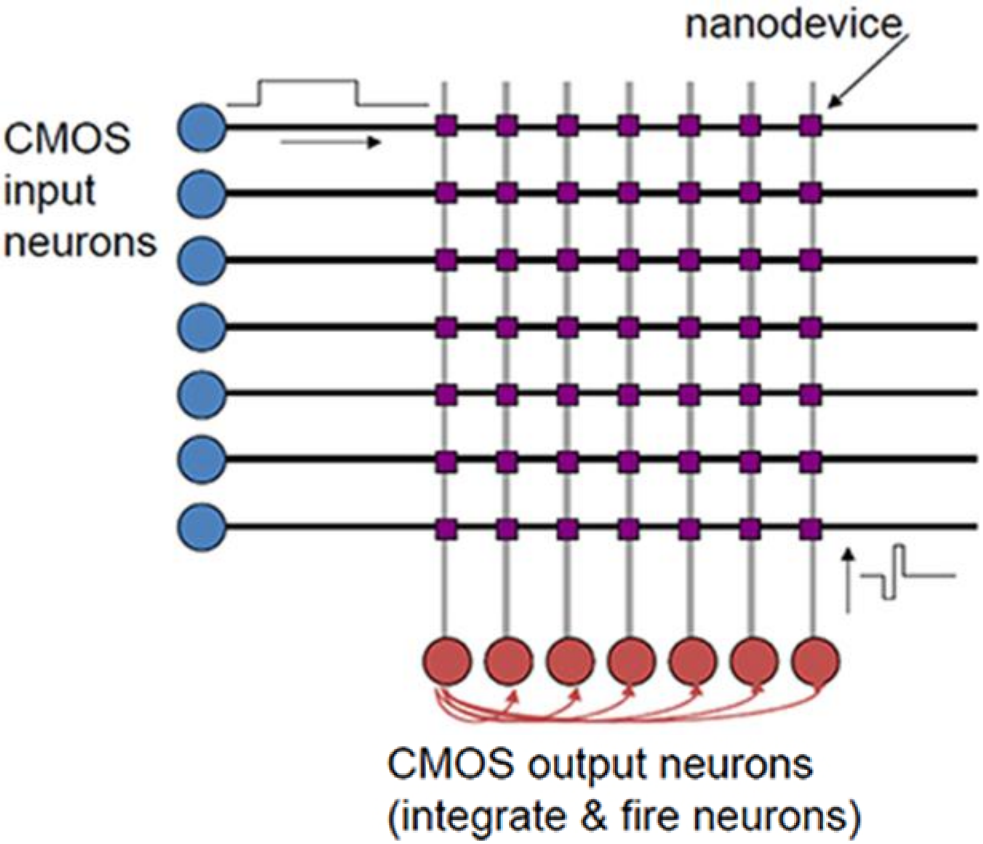


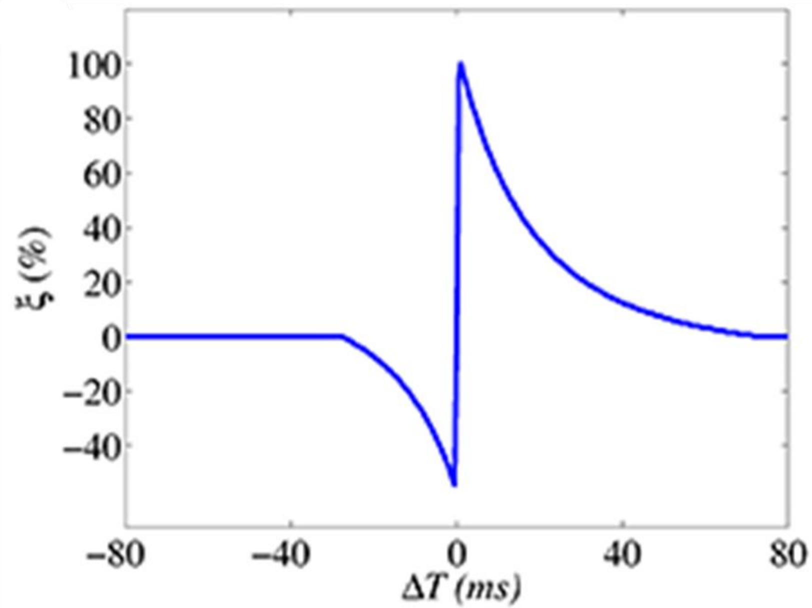
Neural Networks with Memristors

- Memristor conductance ~ synaptic weights
- Voltage/current on memristors adapts weights
- Many memristive Spike-Timing-Dependent Plasticity (STDP) papers



Spike-Timing-Dependent Plasticity (STDP)

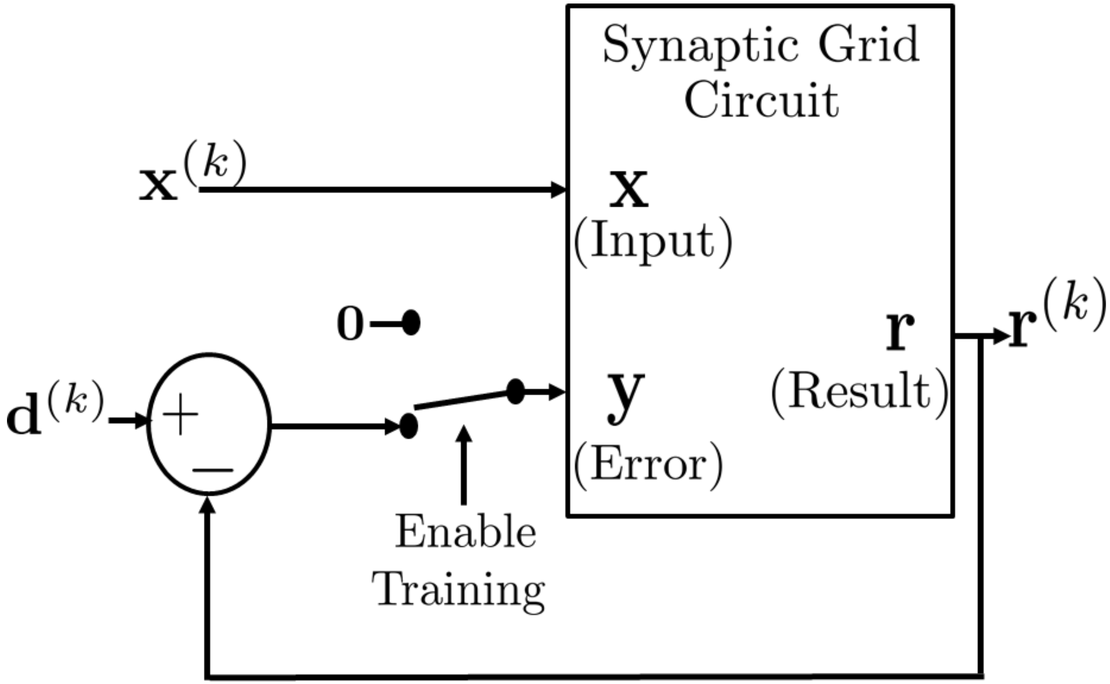


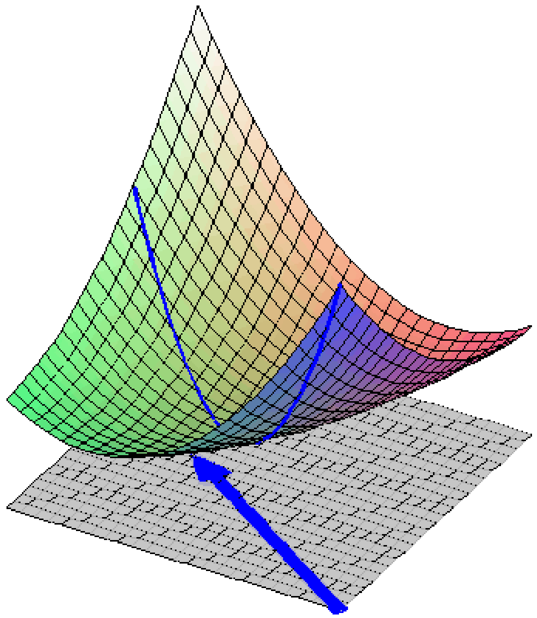


- Biological motivation
- Not useful for machine learning

S. Saighi et al., "Plasticity in Memristive Devices for Spiking Neural Networks," Frontiers in Neuroscience, March 2015

Gradient Descent Learning

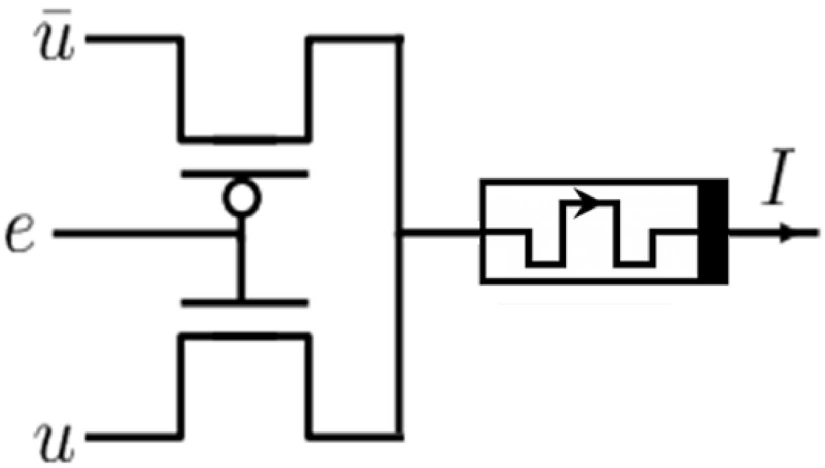


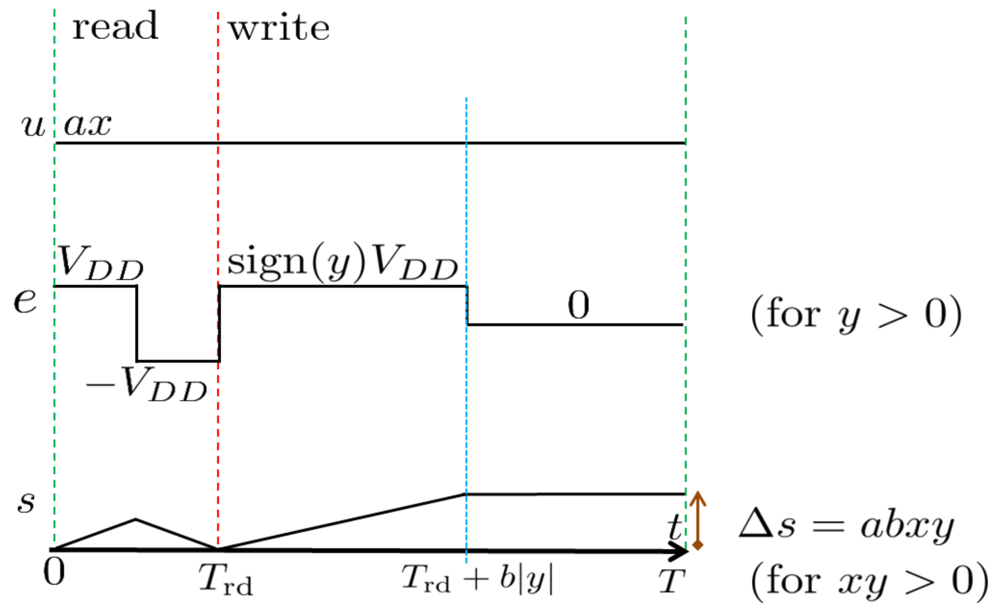


Update rule – a multiplication

$$\Delta \mathrm{W}_{\mathrm{nm}}{}^{(k)} = \eta \mathrm{x}_{m}{}^{(k)} \cdot y_{n}{}^{(k)}$$

Online Memristive Gradient Descent Training



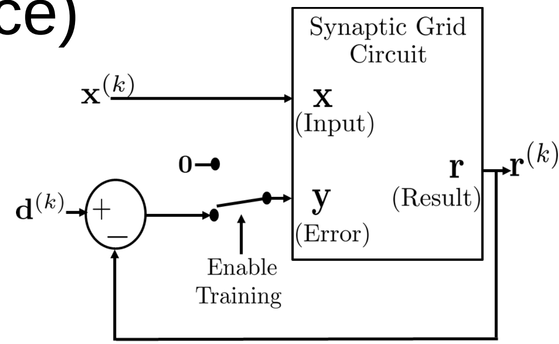


s – Memristor state variable (e.g., resistance)

Moving from voltage to time and voltage

x --> u (voltage)

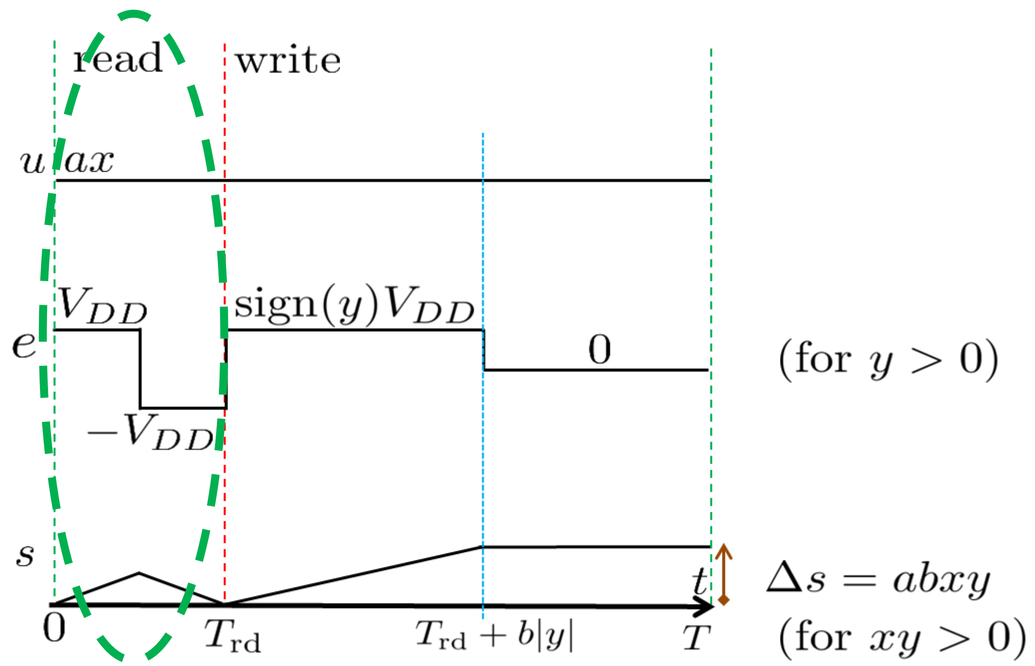
y --> e (voltage and duration)



D. Soudry, D. Di Castro, A. Gal, A. Kolodny, and S. Kvatinsky, "Memristor-based Multilayer Neural Networks with Online Gradient Descent Training," IEEE Transactions on Neural Networks and Learning Systems, October 2015

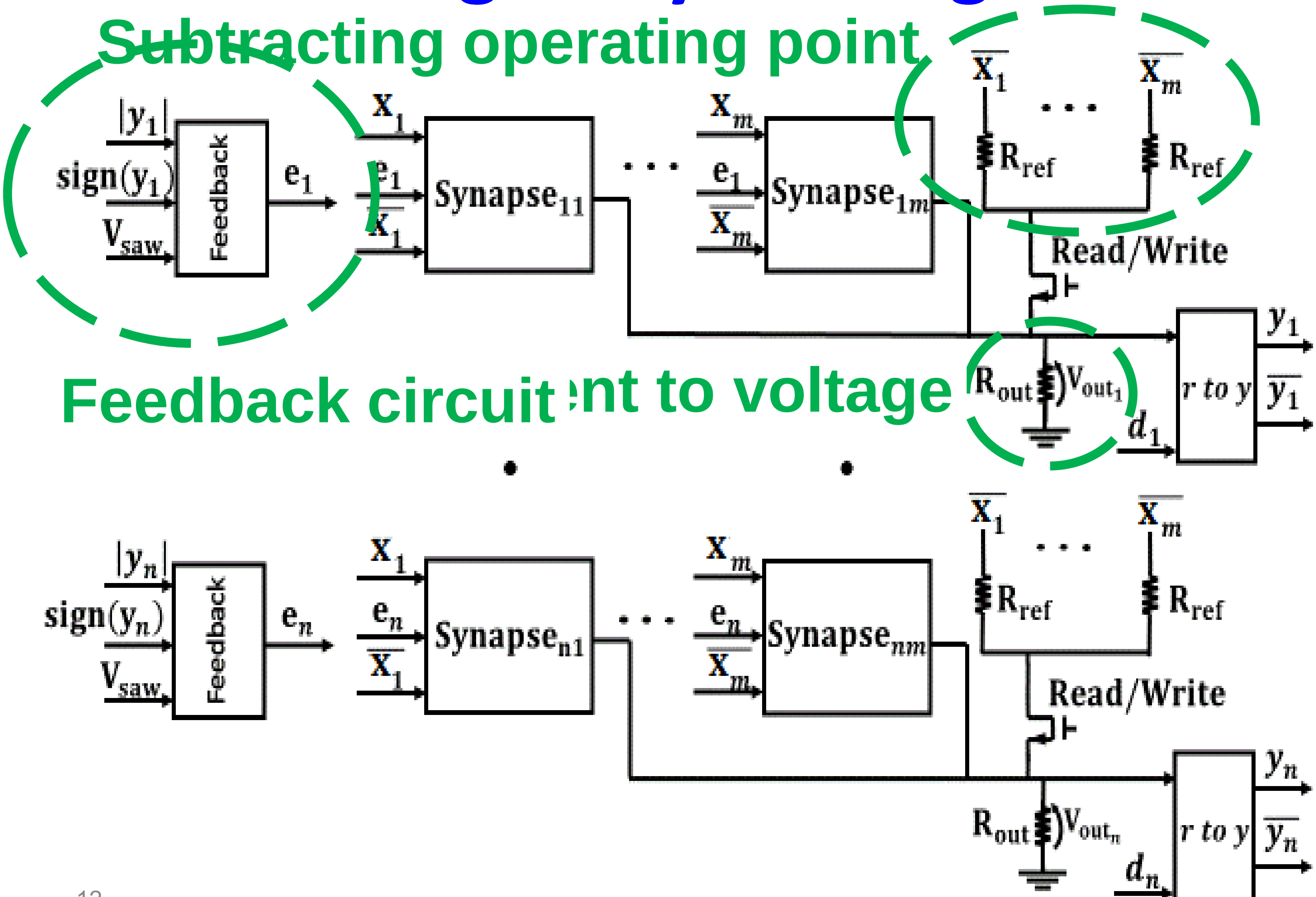
Synapse with TEAM Model

- TEAM is nonlinear – single step read
- Increasing *s* increases resistance, 0<*s*<1
- *s* = 0.5 equivalent to *w* = 0 (negative weights)



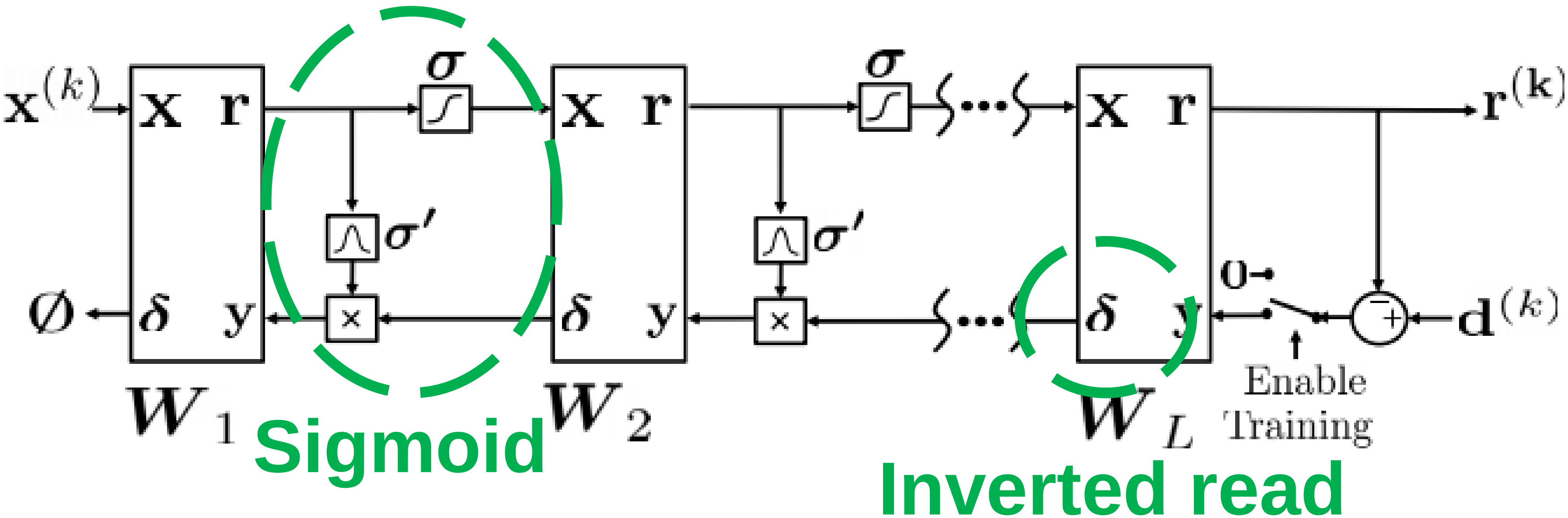
S. Kvatinsky, E. G. Friedman, A. Kolodny, and U. C. Weiser, “TEAM: ThrEshold Adaptive Memristor Model,” IEEE Trans. Circuits and Systems I, 2013

Single Layer Design



E. Rosenthal, S. Greshnikov, D. Soudry, and S. Kvatinsky, “A Fully Analog Memristor-Based Multilayer Neural Network with Online Backpropagation Training,” IEEE Conference on Circuits and Systems, May 2016

Multi-Layer Design



E. Rosenthal, S. Greshnikov, D. Soudry, and S. Kvatinsky, "A Fully Analog Memristor-Based Multilayer Neural Network with Online Backpropagation Training," IEEE Conference on Circuits and Systems, May 2016

Results – Single Layer

Dataset	Unique Training Samples	Unique Test Samples	No. of Inputs	No. of Outputs	NN Size
Wisconsin Diagnostic Breast Cancer	300	120	30	2	30x2
Wine	96	48	13	3	13x3
Iris	90	60	4	3	4x3

Similar accuracy as software

10X faster than software

		Simulation Type - Error %			Runtime		
			Noisy Analog Model	Matlab	Analog	Analog	Matlab
Wine	1200	3.75% ± 0.52%	2.5% ± 0.52%	2.29% ± 1.09%	18ms	~35 *min*	278.5ms
Breast Cancer	1200	3% ± 0.5%	4.67% ± 0.67%	3.1% ± 1.83%	18ms	~30 *min*	210ms
Iris	1080	15.67% ± 0.79%	16.5% ± 0.67%	15.33% ± 0.03%	16.2*ms*	~20 *min*	95.3*ms*

E. Rosenthal, S. Greshnikov, D. Soudry, and S. Kvatinsky, "A Fully Analog Memristor-Based Multilayer Neural Network with Online Backpropagation Training," IEEE Conference on Circuits and Systems, May 2016

Results – Multi Layer

Dataset	Unique Training Samples	Unique Test Samples	No. of Inputs	No. of Outputs	NN Size
Iris	90	60	4	3	$4x4x3$

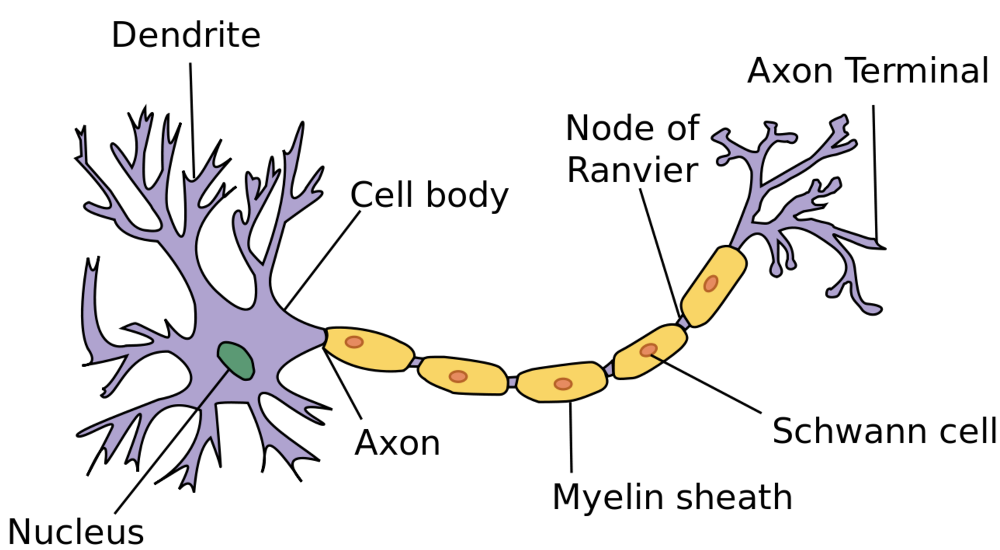
	Simulation Type - Error %			Runtime		
Total Training Samples	Analog Model	Noisy Analog Model	Matlab Model	Analog Model	Analog Model Wall Clock	Matlab Model
2160	$8.16 \pm 1.47\%$	$9.83\% \pm 1.06\%$	$4.5\% \pm 1.93\%$	$43.2ms$	$\sim 10\ hours$	$16.6s$

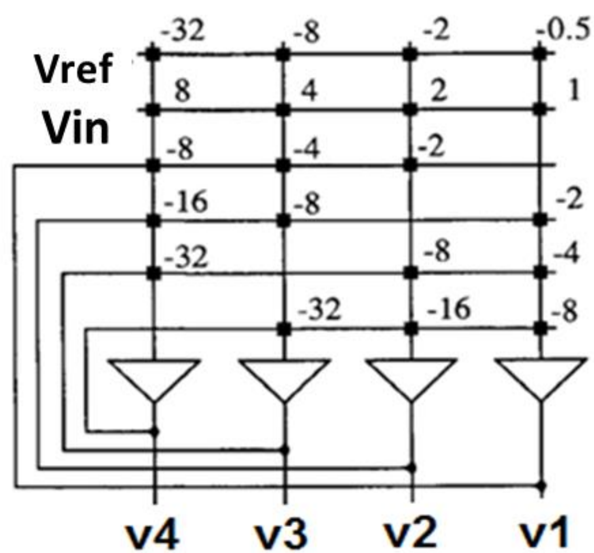
2X more accurate 400X faster than software

Worse than software (why?)

Ongoing Directions

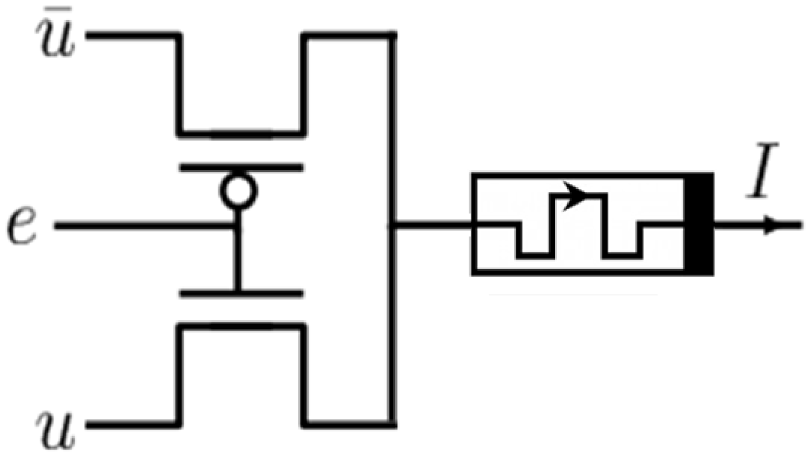
- Machine learning accelerators
- Reconfigurable adaptive hardware (ADC, DAC, etc.)
- Memristors for excitation

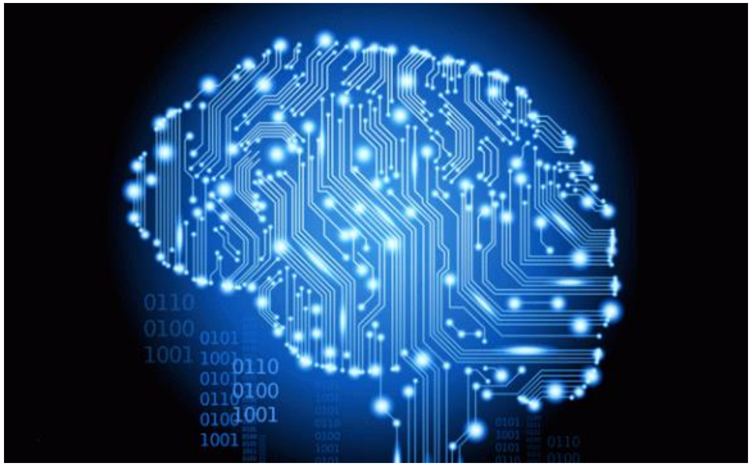




Conclusions

- Neuromorphic accelerators have **huge potential** for machine learning
  - Fast (400X for small network)
  - Accurate (with noise and variations)
  - Dense (2T1M synapse)





Thanks!

ASIC²
ARCHITECTURES
SYSTEMS
INTELLIGENT COMPUTING
INTEGRATED CIRCUITS

TCE
technion computer engineering center

משרד הכלכלה
Ministry of Economy

Advanced Circuit Research Center ACRC

ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

BSF
United States – Israel
Binational Science Foundation