IEEE Xplore® Digital Library

ISCAS 2020

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

CAS

A Process-Variation Robust RRAM-Compatible CMOS Neuron for Neuromorphic System-on-a-Chip

Vishal Saxena

ECE Department, University of Delaware

vsaxena@udel.edu

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

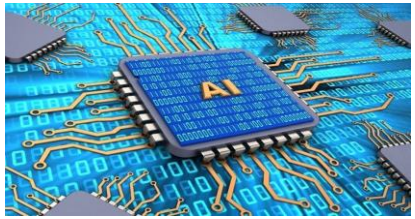
UNIVERSITY OF DELAWARE
ENGINEERING

Outline

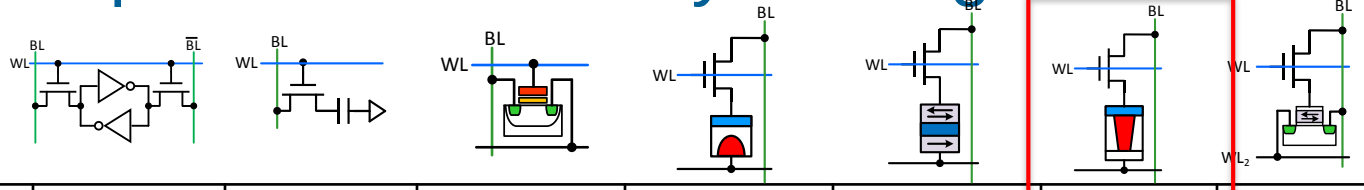
- Introduction
- RRAM-based Neuromorphic Circuits
- On-Chip and Transfer Learning
- RRAM-Compatible CMOS Neuron Design
- Simulation Results and Comparison
- Conclusion

Neuromorphic ICs and Edge-AI

- Deep learning AI is has shown unprecedented success with processing unstructured data
- Growing need for low-power Embedded-AI at the Edge
  - Reduce reliance upon Cloud and wireless infrastructure
  - Data privacy and personalized AI models
- Overarching goal is energy-efficiency of the brain
  - Eliminate von Neumann bottlenecks by computing inside memory
- Emerging non-volatile memory devices (NVM) for high in-memory compute density and low energy consumption

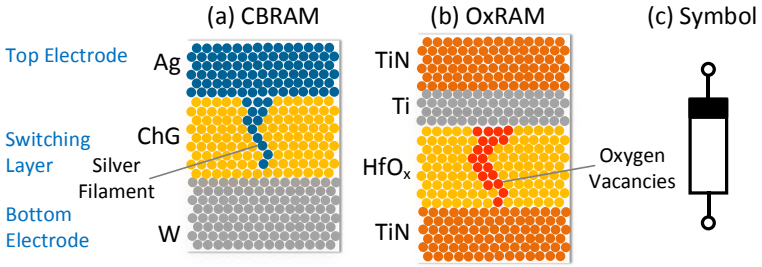


Comparison of Memory for Edge-AI

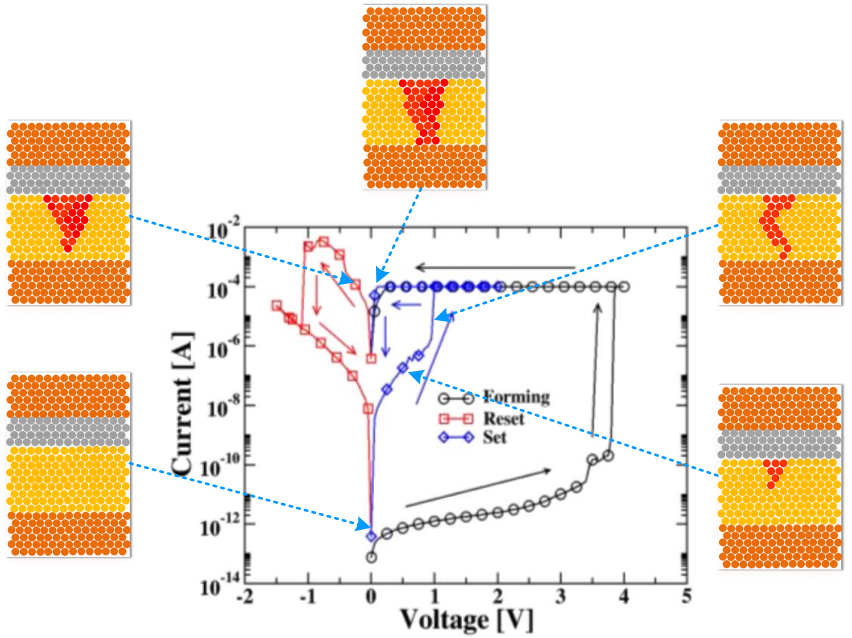


Parameters	SRAM	DRAM	NOR Flash	PCRAM	STTRAM	RRAM	FeFET
Cell Size	$100F^2$	$7F^2$	$10F^2$	$4F^2$	$12F^2$	$4\text{-}12F^2$	$24F^2$
Write Latency	1ns	5ns	10µs-1ms	100ns	2-25ns	10ns	3ns
Read Latency	1ns	20-80ns	50ns	10ns	2-25ns	1-10ns	2ns
Write Energy (pJ/bit)	<1	<1	100	2-25	0.1-2.5	0.1-3	0.1
Leakage Power	High	Medium	Low	Low	Low	Low	Low
Endurance (write cycles)	$>10^{16}$	$>10^{16}$	10^5	10^9	10^{15}	10^{10}	$>10^5$
MLC Capability	✗	✗	4-bit	4-bit	2-bit	4-bit	3-bit
MLC Retention	✗	✗	4-6 months	10^4 s	-	10^4 s	-
			Tunneling	R-drift, abrupt Reset	Tunneling	R-drift	Tunneling
3D Stacking	✗	✗	✗	✓	✗	✓	✓

RRAM Switching

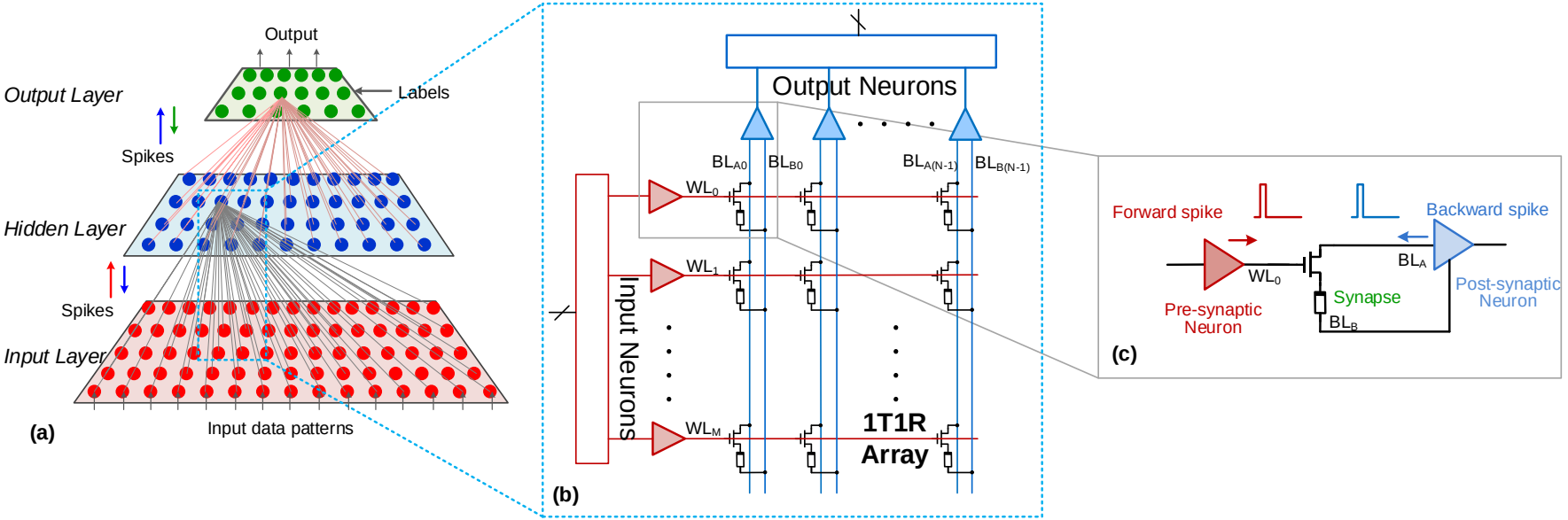


- Pristine devices need to undergo a forming step
- Program (Set) and Erase (Reset) threshold voltages, V_{tp} and V_{tm}.
- Compliance current is set by the semiconductor parameter analyzer
- Analog-like states due to formation of weak filaments



A. Grossi, IEDM 2016

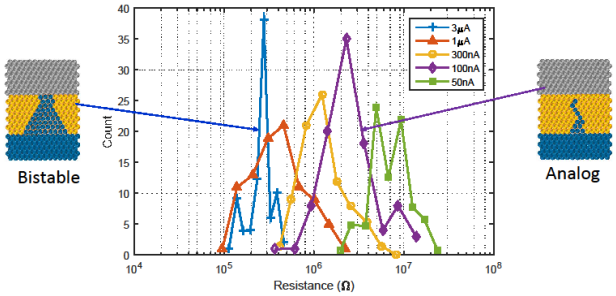
1T1R Array Architecture



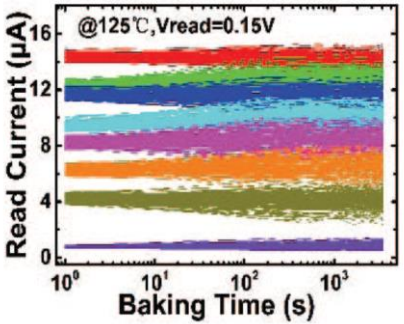
- Binary spikes are used for encoding, computation and communication
- Integrate-and-fire neurons are the array periphery
- Pre-neurons drive through the wordline (WL)
- Dual bitlines (BL_A and BL_B) that connect to the post-neuron

RRAM Device Challenges

- Variability
  - Device switching threshold voltages and resistances are variable across devices
- Resolution and Retention
  - RRAMs can realize multi-level cell (MLC) capability
  - Resistance drift over time and presents challenges in their use stable analog synapses
- Low Resistance
  - Typical RRAMs exhibit 10k low-resistance state (LRS) resistance which incurs static power consumption in driver circuits
- Endurance
  - Write endurance determines the continuous learning ability on chip
  - Reported best case OxRAM endurance is 10^{10} cycles



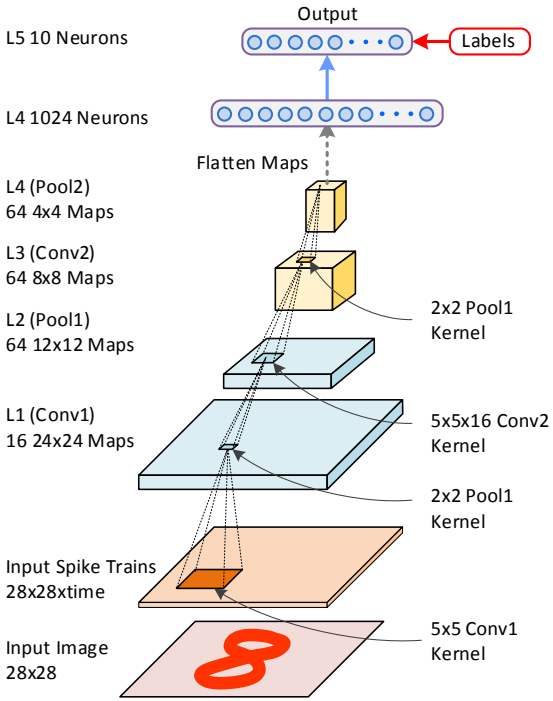
CBRAM Devices from Dr. Mitkova's group at BSU, 2016.



Zhao et al, "Investigation of Statistical Retention of Filamentary Analog RRAM for Neuromophic Computing," IEDM 2017.

SNN: On-Chip and Transfer Learning

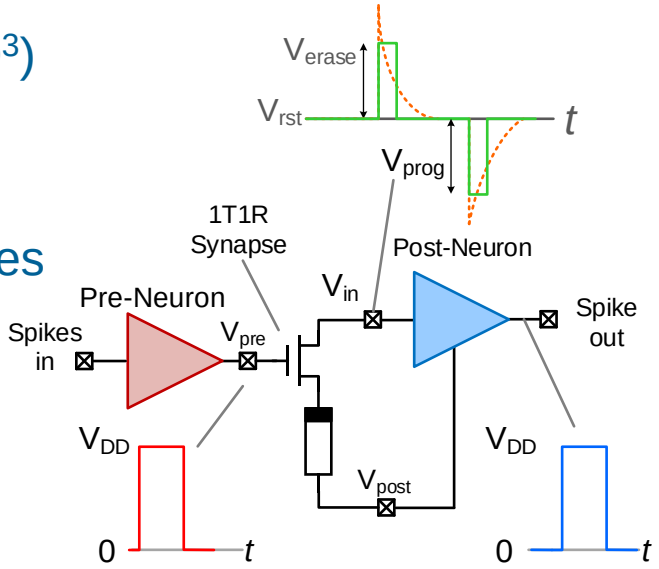
- On-chip Learning
  - Brain-inspired approaches where each layer learns in an unsupervised manner
    - Classification accuracy flattens out with two Conv layers
  - Adapt Backpropagation to spiking neural networks
    - Need to handle non-differentiable spiking neurons
- Transfer Learning[1]
  - Train deep ANN using standard neuron models (TensorFlow)
  - Convert ANN to equivalent SNNs
    - Rate coding of spikes and weight scaling



[1]Diehl et. al., "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," IJCNN 2015.

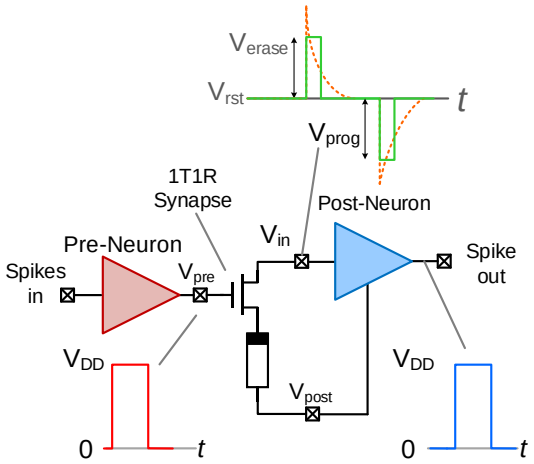
1T1R Synapse and Neuron

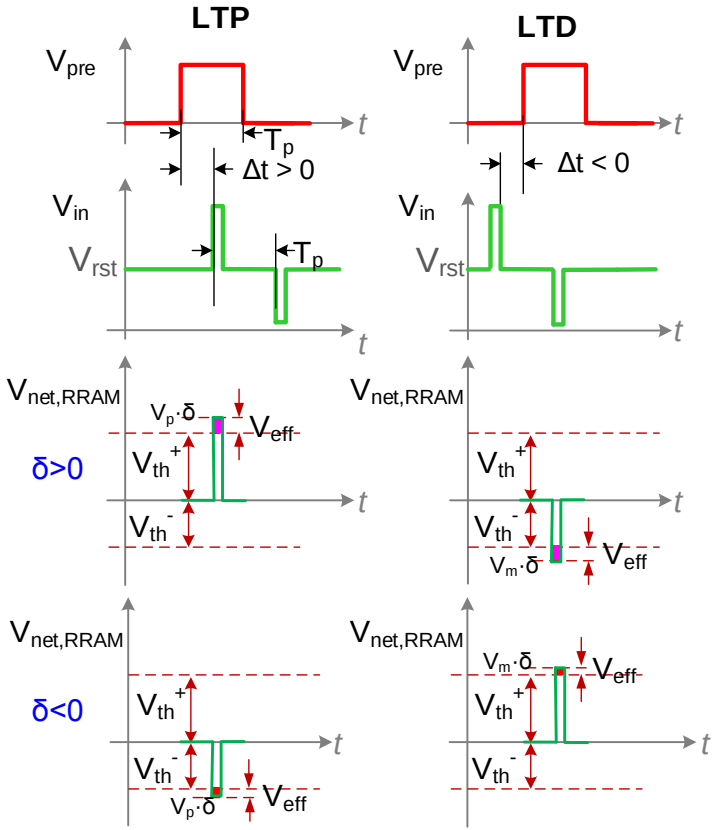
- CMOS-RRAM Neuromorphic architecture should support on-chip as well as transfer learning
- Neurons should be able to drive a large fan-out (>10^3)
- 1T1R synapse[1] was earlier proposed using discrete realization
- A novel CMOS neuron is proposed for 1T1R synapses
- Waveform engineering for weight update rules
- Neurons drive gate input capacitance in the forward path
- The RRAM resistance is driven in a sparse manner from node V_{in}

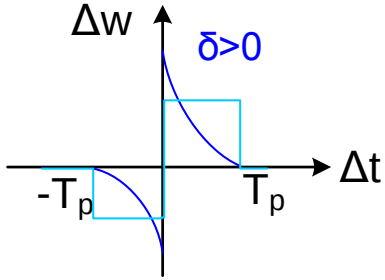


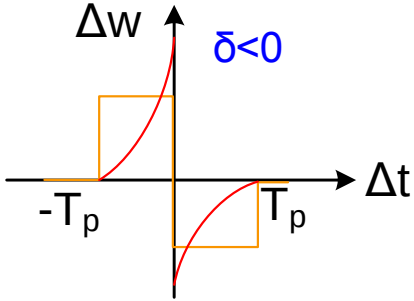
[1]Ambrogio et. al., "Novel RRAM-enabled 1T1R Synapse Capable Of Low-Power STDP Via Burst-Mode Communication And Realtime Unsupervised Machine Learning.," VLSI Tech., 2016.

STDP Waveforms



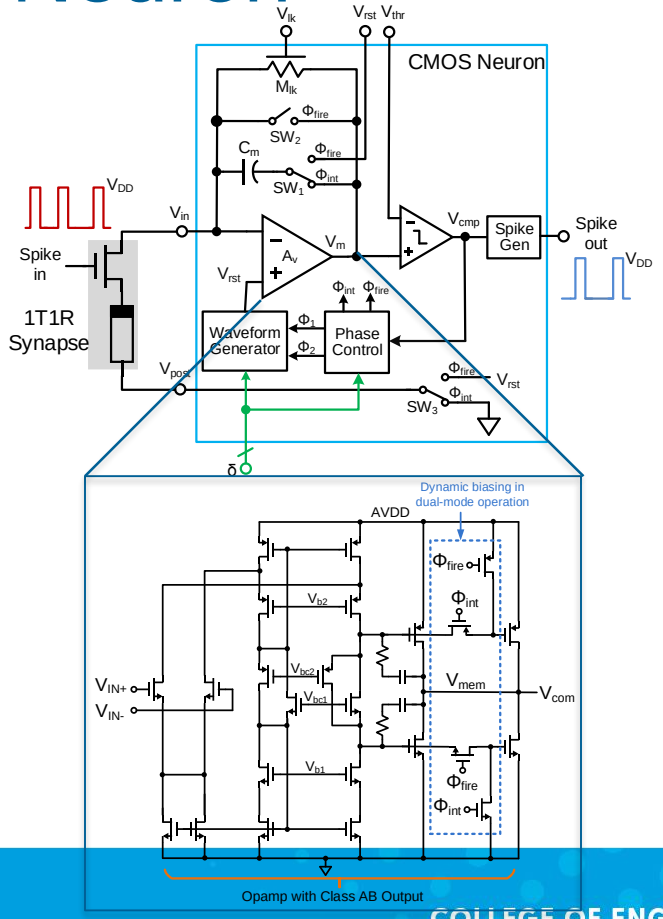




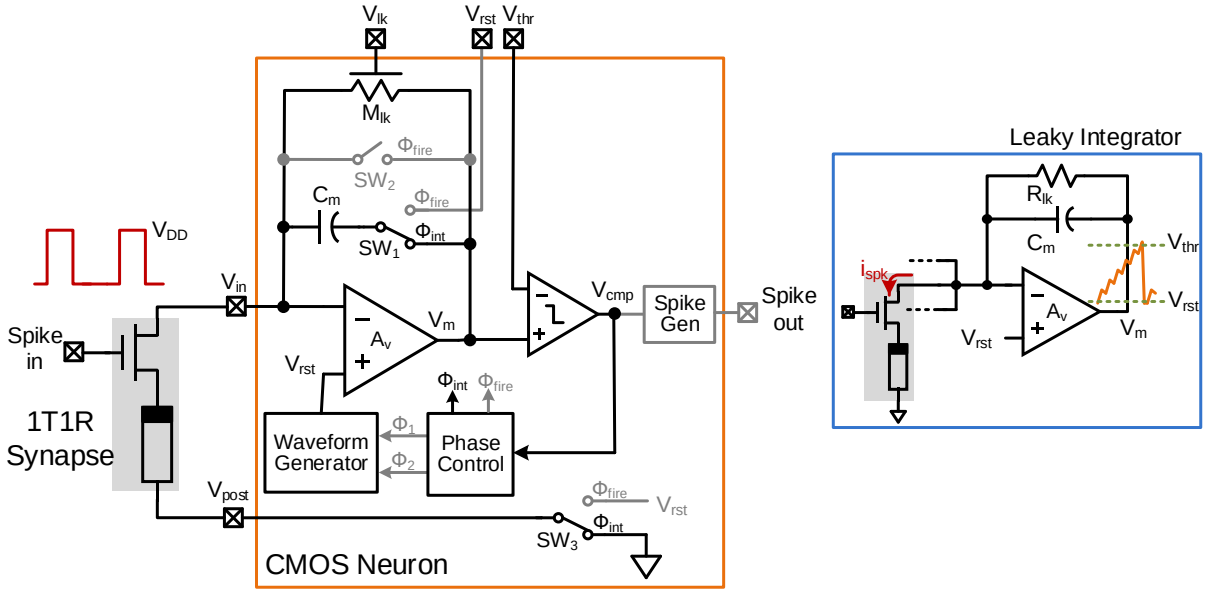


Event-Driven Neuron

- Event-driven switched-capacitor neuron
  - Integrate and fire modes
- Single-opamp architecture
- Asynchronous comparator
- Local configurable waveform generator for weight update
- Output spikes are full-CMOS levels

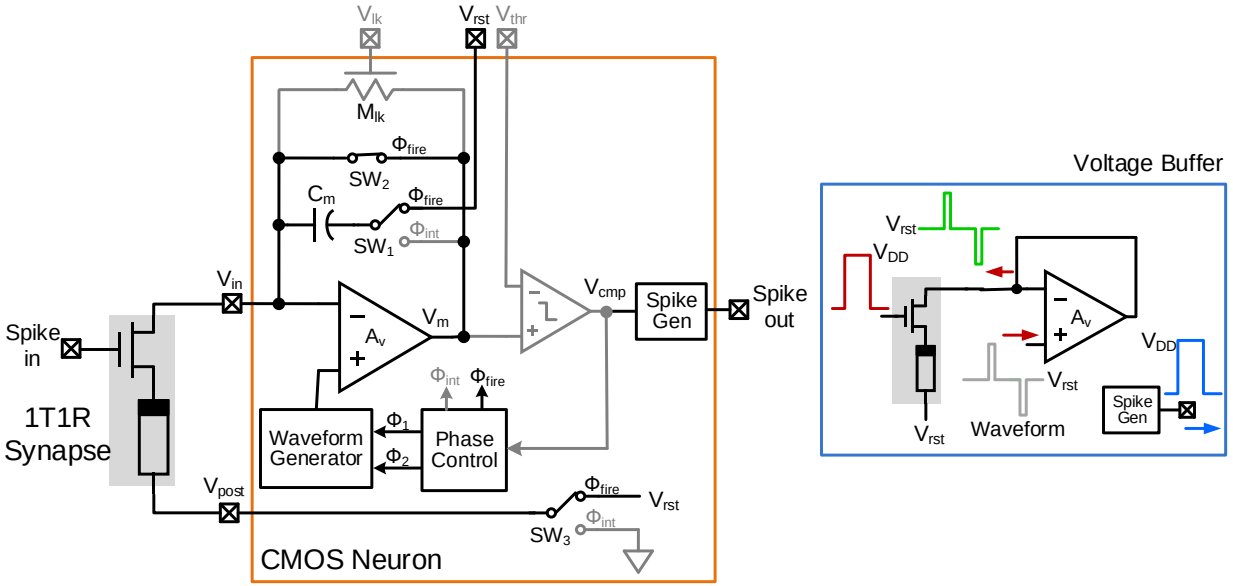


Event-Driven Neuron-Integration Mode



- Opamp configured as an integrator
  - Biased with a very low DC current
- Asynchronous comparator detects positive crossing of the threshold, V_{thr}

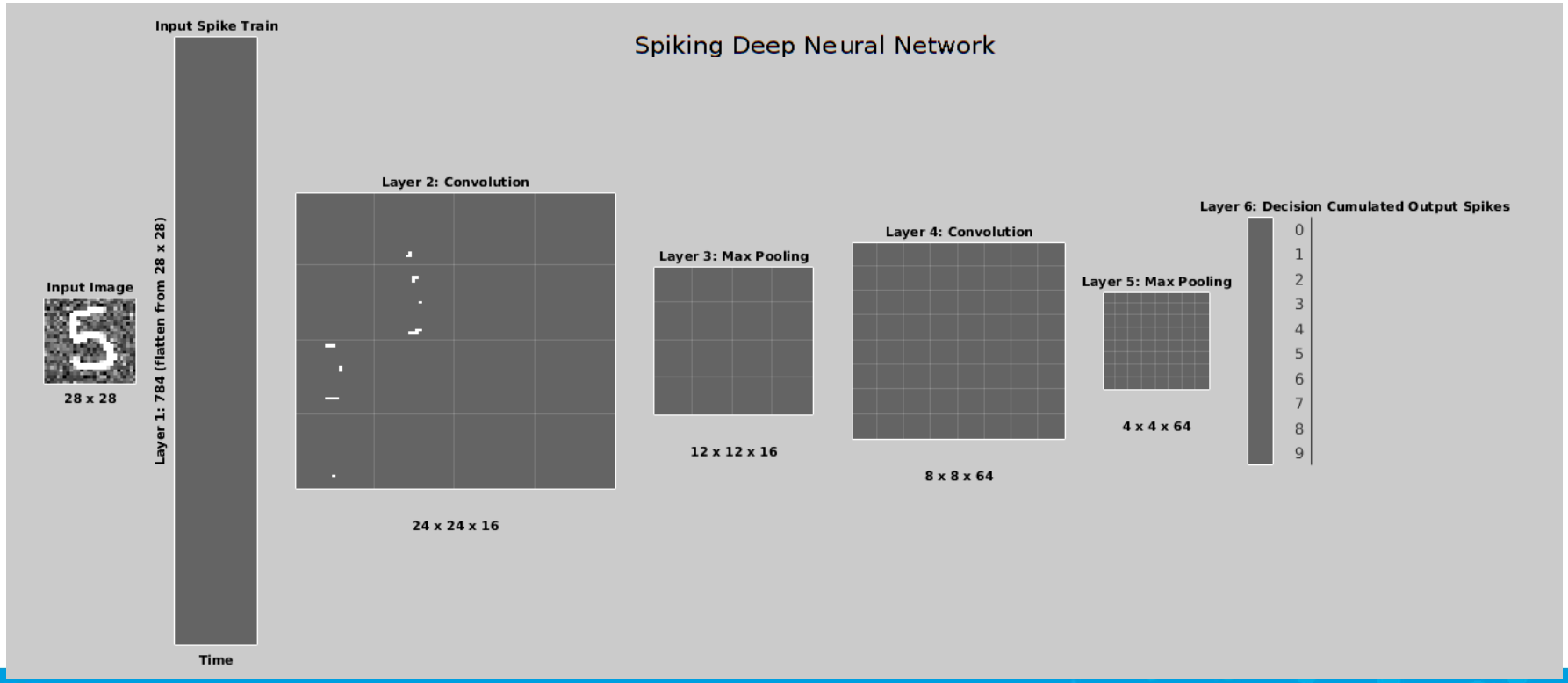
Event-Driven Neuron-Fire Mode



- During fire mode, the opamp is reconfigured as a voltage follower
  - Drives the RRAM in series with the select transistor (if On)
- Opamp is dynamically biased to drive a resistive load
- A binary output spike pulse is create and driven by a CMOS buffer

Spiking Conv Neural Network

ConvNet: 28x28-16c5-2s-64c5-2s-10o



Impact of Process Variations

- We analyze performance degradation in the context of transfer learning
- Opamp Offset
  - Changes summing node voltage
- Opamp Finite-gain
  - Changes summing node voltage and leak time-constant
- In 1R RRAM array, offset will appear as spurious synaptic current (V_{os}/R_M)
  - Affects on-chip SNN performance
- In 1T1R array, synaptic current is tolerant to the summing node voltage

Impact of Process Variations

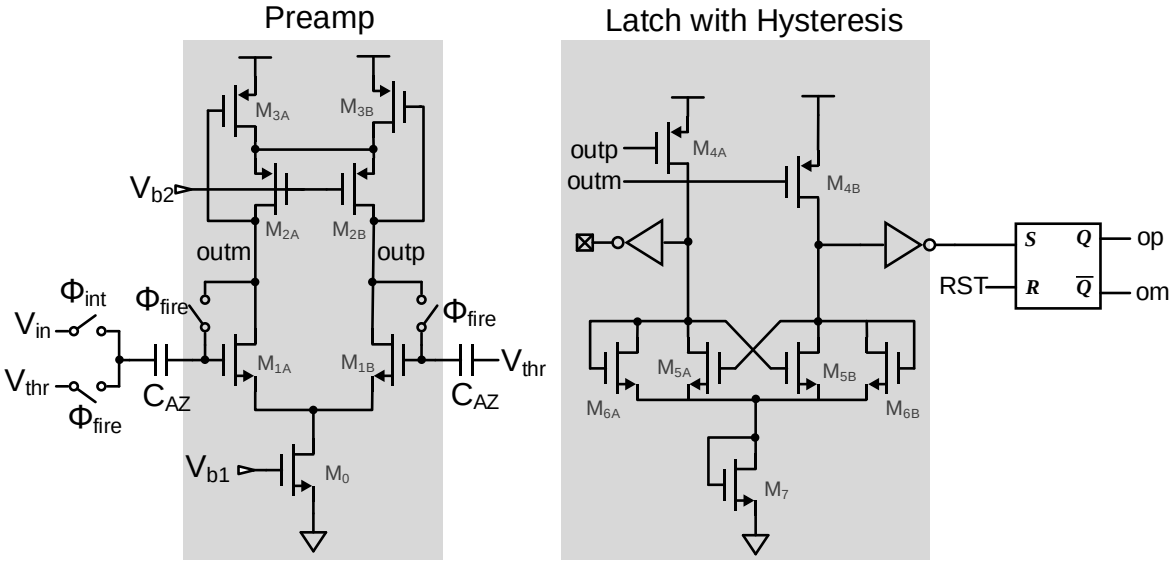
- Comparator Offset
  - Random offset changes the firing threshold voltage, $V_{thr,j}$
  - Different for each neuron, similar to changing bias in standard DNN
  - Affects classification performance of the SNN
- Process R and C variation:
  - ±20% variation from chip-to-chip
  - ±1% variation on the same chip
  - Changes the gain of the neuron's transfer function in integration mode
  - The trained model will differ from the on-chip neuron gains

SNN Performance Degradation for MNIST

Network	Fully-connected SNN	ConvNet SNN
Ideal Neuron	98.61%	99.10%
$\sigma_{os,c}$ = 100mV	95.24%	95.78%
$\sigma_{os,c}$ = 50mV	96.11%	97.07%
$\sigma_{os,c}$ = 5mV	98.53%	98.91%
C_m = -20%	98.49%	98.65%
C_m = +20%	98.32%	98.68%

- Impact of random comparator offset on transfer learning classification performance is the most significant
- Impact of RC variation is same across all the neurons

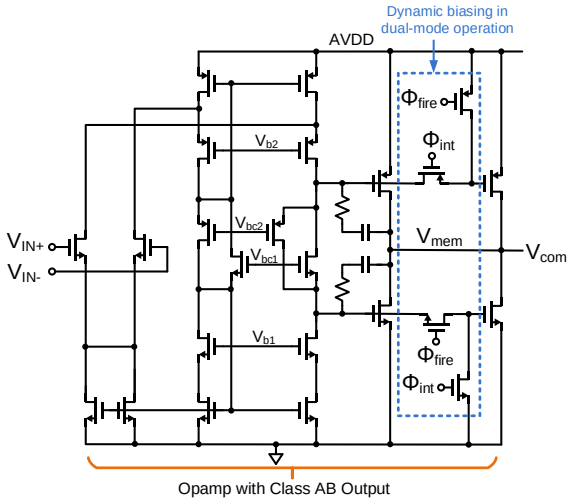
Asynchronous Comparator with AZ

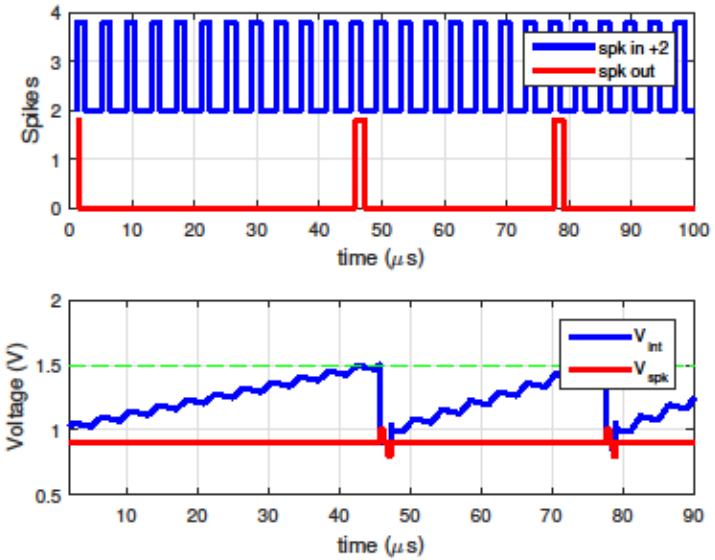


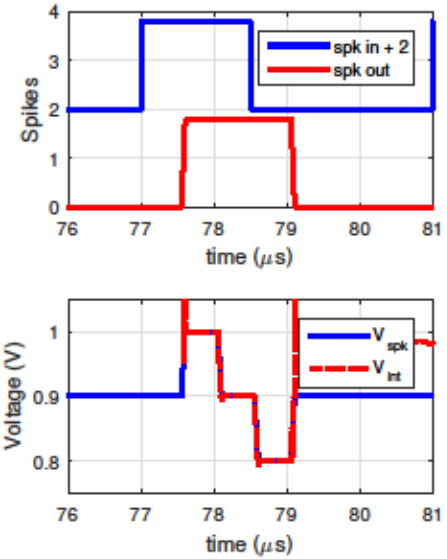
$$\sigma_{os,c} = \sqrt{\sigma_{os,PA}^2 + \frac{\sigma_{os,Latch}^2}{A_{PA}^2}}$$

- Autozeroing of latch offset using pre-amp gain
  - Offset stored in C_z during the fire mode
  - Stored offset canceled during the integrate mode
- Need for wait for a few spikes or implement AZ in the background

Event-Driven Neuron-Fire Mode







- Implemented in 180nm CMOS technology
- Folded cascode opamp with dynamically biased class-AB output stage
- Inference-only designs can be achieved with very low power
- For training, opamp unity-gain frequency (f_{un}) set by the STDP waveforms

Performance Comparison

- Only recent CMOS neurons that can interface with RRAMs are considered

Design	Type	Technology	Synapse Type	On-chip Learning	I_{VDD}	Energy-efficiency (fJ/spike/synapse)
Wu et al 2015	Opamp	180nm	1R	✓	13µA	140
Sahoo 2017	Ring VCO	65nm	None	✗	-	-
Larras et al 2017	Current-summing	65nm	Digital	✗	-	7
Sourikopolous et al 2017	Subthreshold	65nm	None	✗	-	4
This work (inference)	Opamp	180nm	1T1R		9µA[1]	8.1[1]
This work (training)	Opamp	180nm	1T1R	✓	9µA	40

[1] Can be further optimized for inference-only realization

Conclusion

- Impact of circuit non-idealities on performance due to process variations has been investigated in the context of transfer learning
- Random comparator offsets cause the largest degradation in classification accuracy and must be compensated on chip
- Synaptic resistances are driven only during intermittent STDP update events
- The neuron allows digital-like drive during inference with 8fJ/synapse/spike energy consumption.

Questions?

UNIVERSITY OF DELAWARE
ENGINEERING