Landscape of Synaptic Weight Memories

Prof. Shimeng Yu

Georgia Institute of Technology

Email: shimeng.yu@ece.gatech.edu

Web: https://shimeng.ece.gatech.edu/

Outline

- Background and Motivation
- Synaptic Devices: State-of-the-Art
- Variability and Reliability Characterization at Array-level
- Benchmark of Synaptic Devices for Inference and Training
- Chip-level Demonstrations: State-of-the-art

Artificial Intelligence (AI) Applications

Waymo is first to put fully self-driving cars on US roads without a safety driver

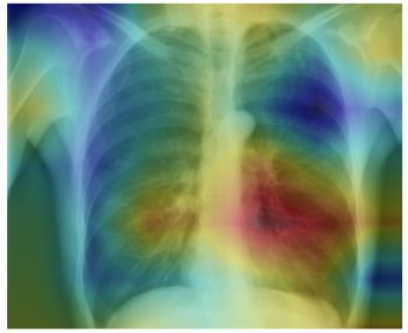
Going Level 4 in Arizona

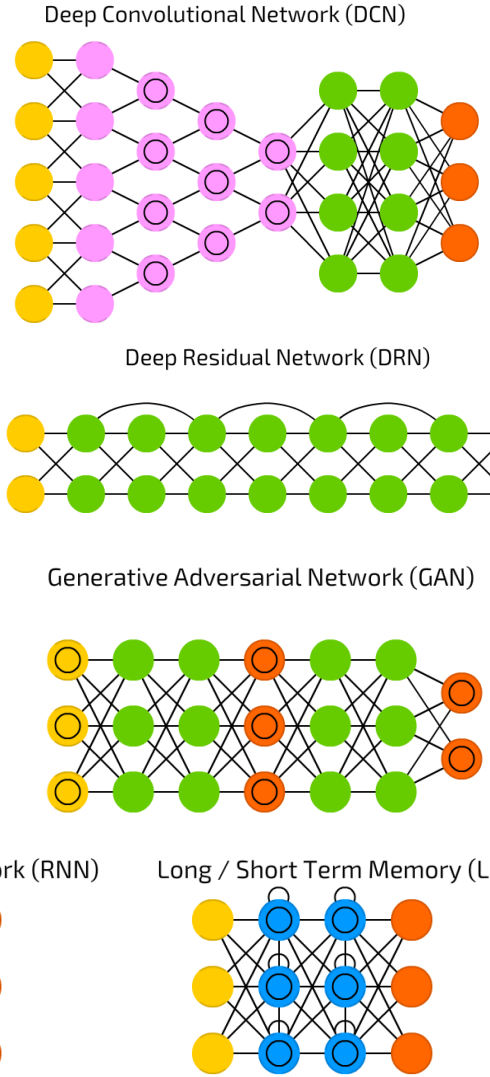
THE VERGE

CNBC

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng





AI today is widely used in computer vision (i.e. image classification), natural language processing (i.e. language translation), etc.

Deep neural network (DNN) topologies

http://www.asimovinstitute.org/neural-network-zoo/

Typical DNN Models for Image Classification

CIFAR-10			IMAGENET		
Network	**Parameters (MB)**	**Total operations(G)**	**Network**	**Parameters (MB)**	**Total operations(G)**
VGG-8	13	0.60	AlexNet	61	1.44
ResNet-20	0.27	0.04	VGG-16	138	31
ResNet-32	0.46	0.07	VGG-19	144	39
ResNet-44	0.66	0.11	ResNet-18	11	3.7
ResNet-110	1.7	0.26	ResNet-34	23	7.2
DenseNet-40	1.0	0.28	ResNet-152	60	23
DenseNet-100	7.0	0.94	DenseNet-121	7.9	5.9

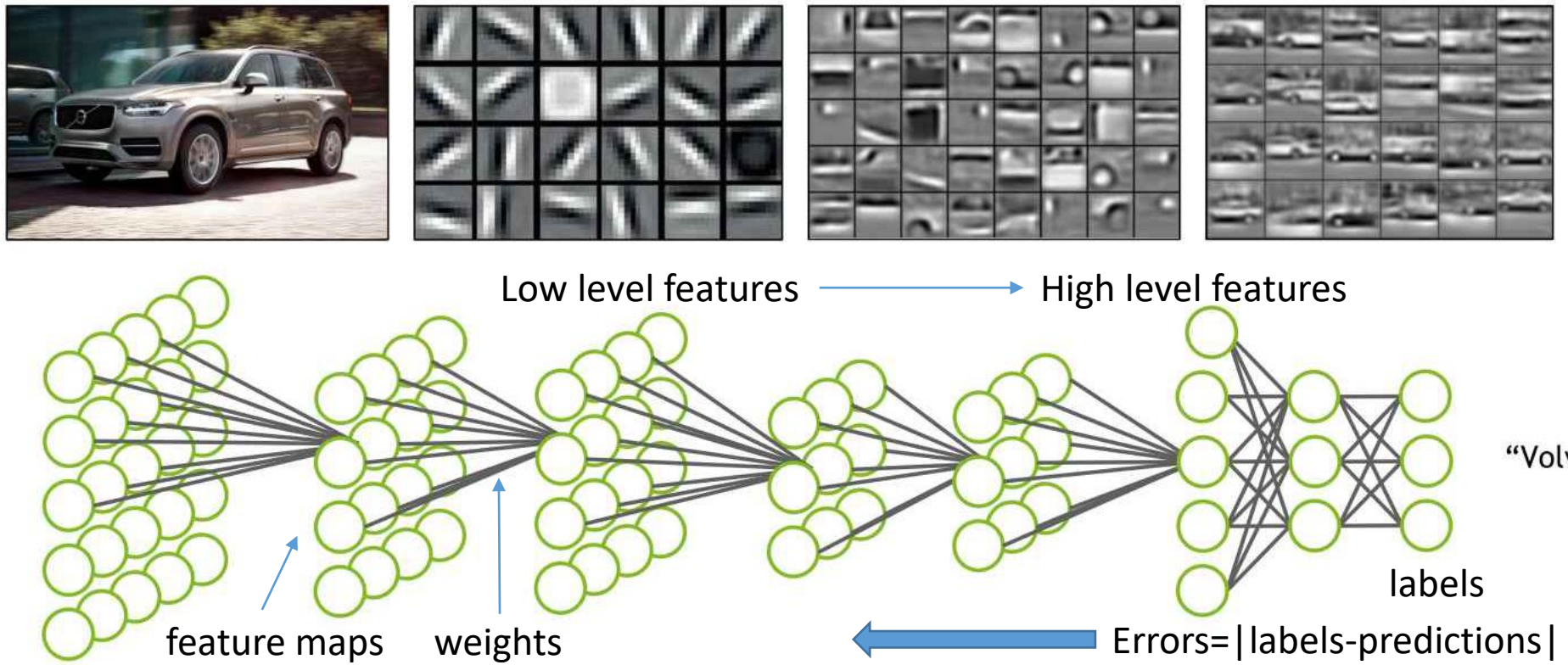
For image classification, model size tens of MB

For language translation, model size can be up to 10 GB

→ Require 10MB to 10 GB on-chip memories

→ Thus requires multi-bit and 3D integration

CONVOLUTIONAL NEURAL NETWORKS



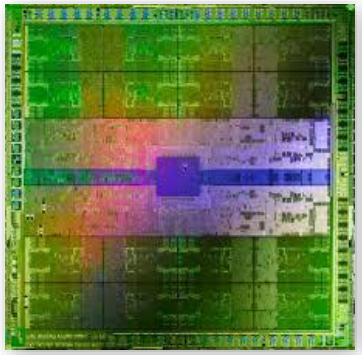
Training: to learn weights iteratively with back-propagation of errors from the output labeled data→ "write" intensive to synaptic weight memories

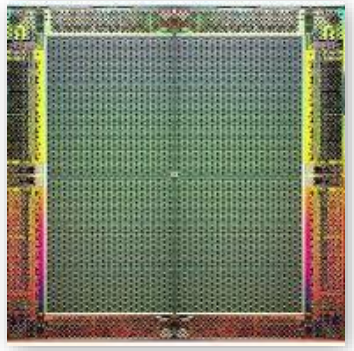
Inference: after training is done, feedforward propagation for prediction only→ → "read" intensive to synaptic weight memories

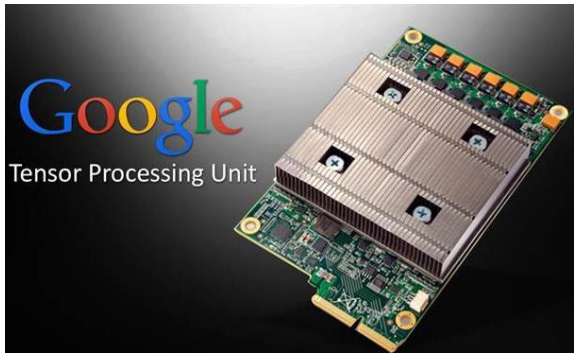
Most intensive computation: vector-matrix-multiplication (to be accelerated by hardware)

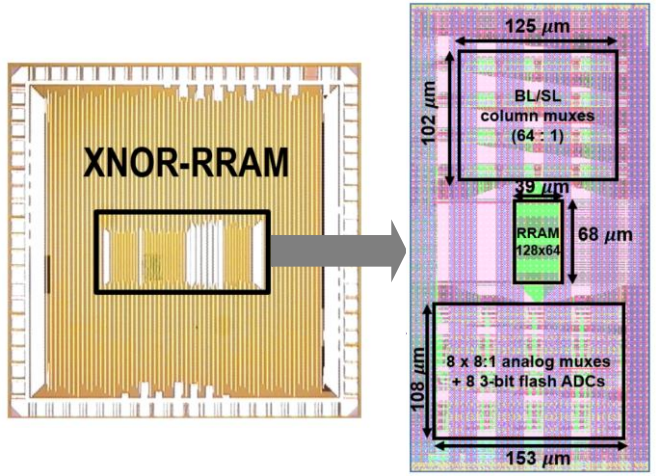
Hardware Accelerators for AI

- GPU still dominates the training in cloud, FPGA is good for inference for fast prototyping
- TPU (or similar digital ASIC) is ramping up in cloud as well as edge









GPU | **FPGA** | **TPU** | **Compute-in-memory (CIM)**

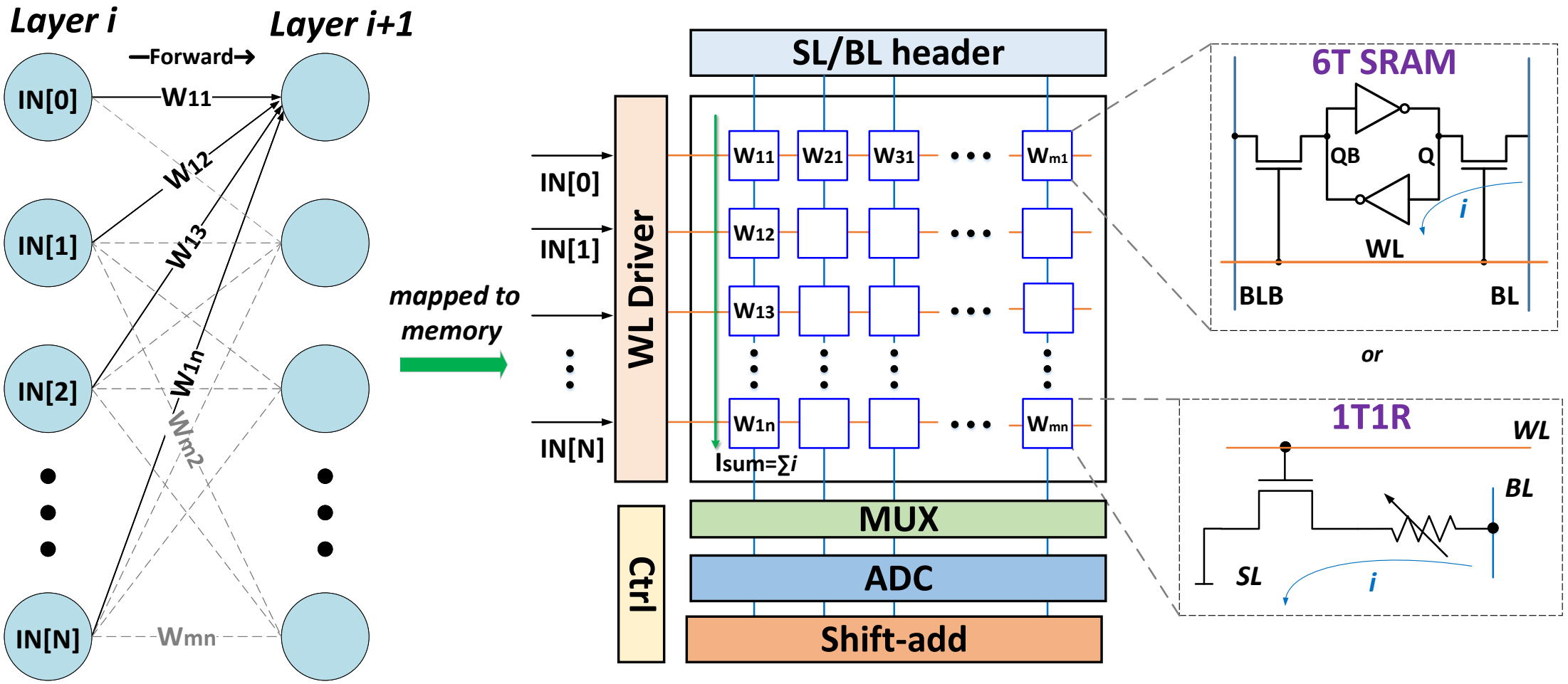
Conventional computing platforms
~ 0.1 TOPS/W
Floating-point

Digital CMOS ASICs
~ 1-10 TOPS/W
Fixed-point

Analog CMOS (or eNVMs)
~ 10-100 TOPS/W
Low-precision→ accuracy?

- To further improve energy efficiency (TOPS/W), analog CIM (possibly with eNVMs) is promising especially in the edge inference where the model is pre-trained.
- CIM chip could also support incremental learning with continuous (possibly unlabeled) new data (e.g. with reinforcement learning) when deployed to the field.

CIM Basics: Mixed-Signal Compute



8-bit weight may need 8 SRAM cells, and shift-add

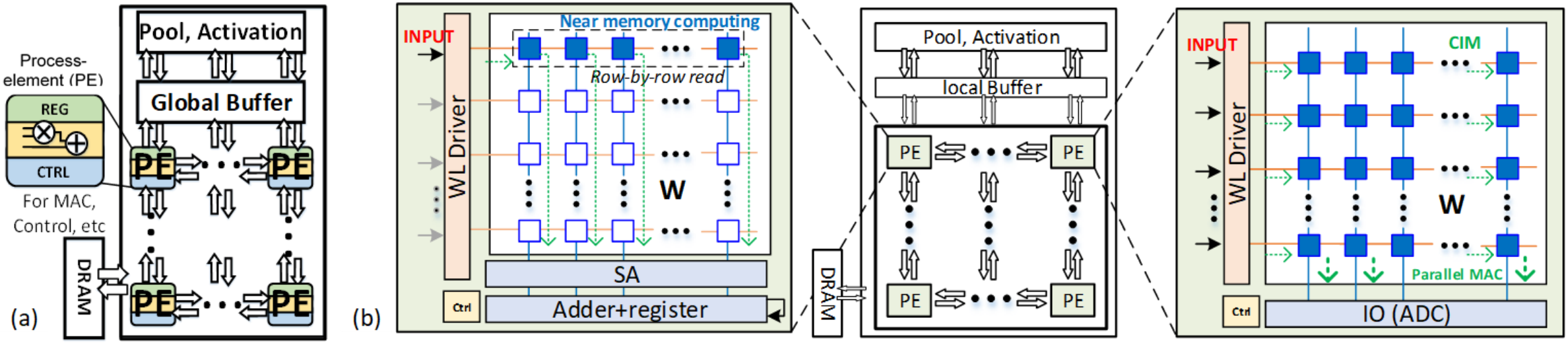
8-bit weight may need 2 1T1R cells (if each cell is 4b/cell), and shift-add

Digital vs. Near-Memory vs. CIM Accelerator

TPU-like digital accelerator

Near-memory-compute accelerator (row-by-row)

In-memory-compute accelerator (parallel)



TPU-like digital accelerator: PE only has MAC units such as multiplier and adders, while the data (both activation and weights) are accessed by shared global buffer (e.g. SRAM cache) → Single row access, slow and inefficient

Weights are stored in memory array, while the activations are loaded in as input to WLs

Near-memory compute: Row by row access with digital adders at periphery

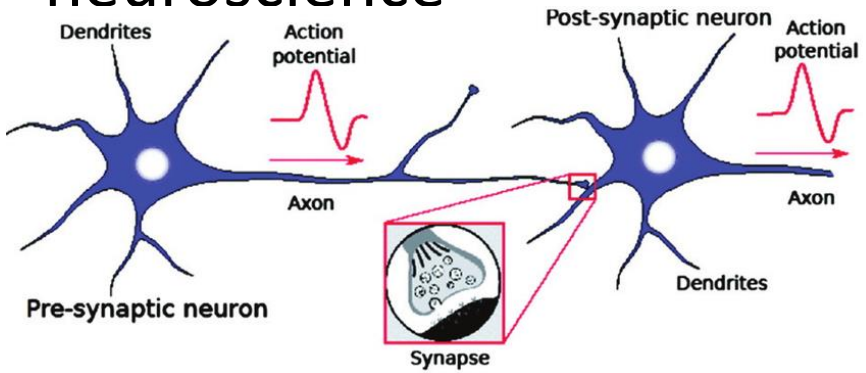
In-memory compute (CIM): Parallel access and ADC for partial sum quantization

Outline

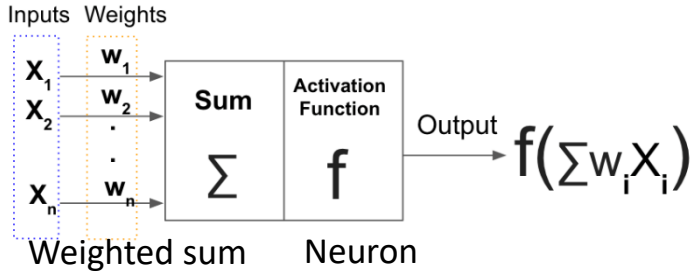
- Background and Motivation
- Synaptic Devices: State-of-the-Art
- Variability and Reliability Characterization at Array-level
- Benchmark of Synaptic Devices for Inference and Training
- Chip-level Demonstrations: State-of-the-art

Electronic Synapses and Neurons

- Inspirations from biology and neuroscience



- Mathematical formulation in machine learning



Structure of artificial neuron

Abstractions for device engineers:

Synapses: local memories that carry weights

→ Multi-bit memories

1) Two-terminal resistor

2) Three-terminal transistor

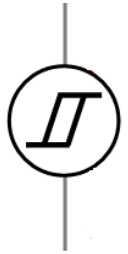
(biased at linear region)

Neurons: simple thresholding compute units

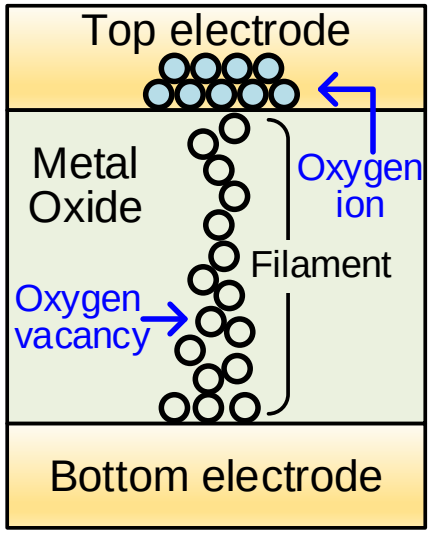
→ Threshold switches

1) Abrupt switching in I-V

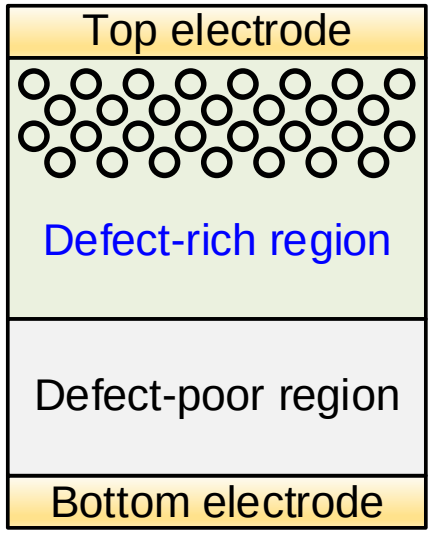
2) Returns to off-state at zero voltage (not memory)



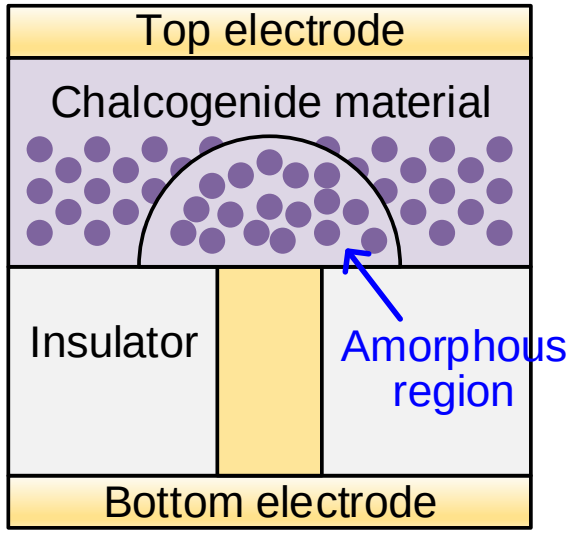
Landscape of Analog Multi-bit Memories



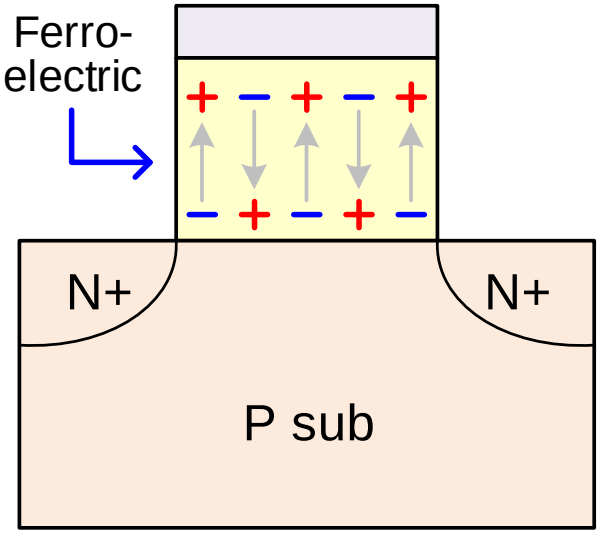
(a) Filamentary RRAM



(b) Non-filamentary RRAM



(c) Phase Change Memory



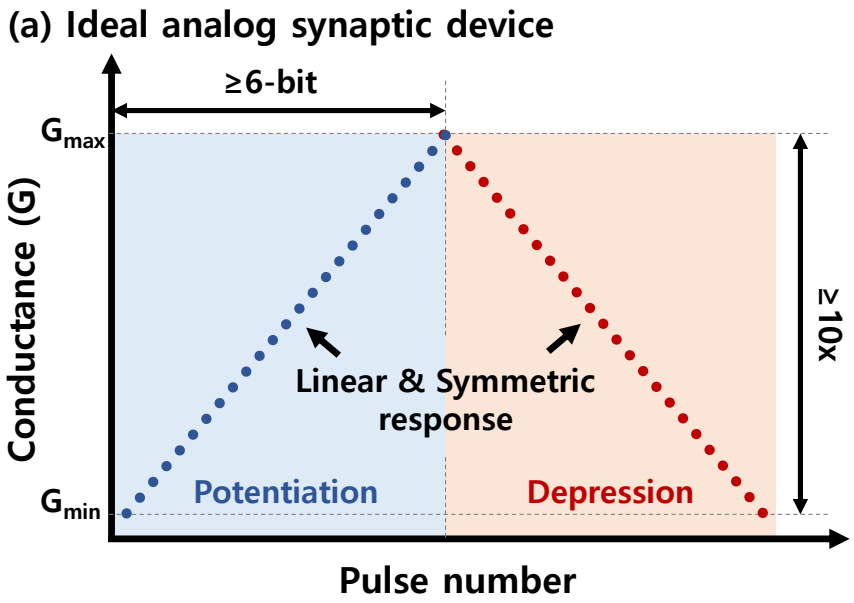
(d) Ferroelectric FET

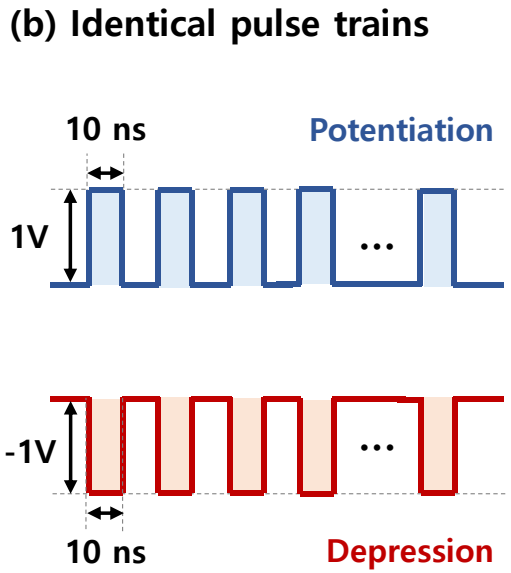
Partial switching in these materials leads to analog multi-bit memories as synaptic weights
RRAM and PCM are more current driven, and FeFET is electric field driven (less energy!)

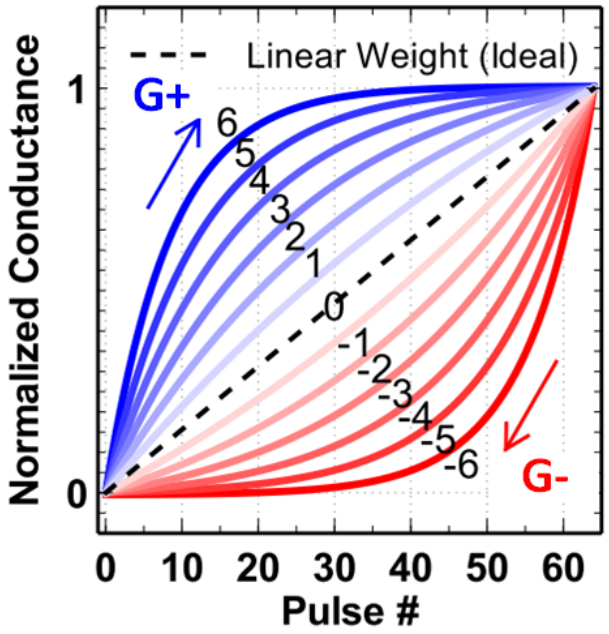
STT-MRAM/SOT-MRAM can be used as binary synapse in principle, electrochemical random access memory (ECRAM) is premature. Therefore, we will not discuss about these candidates in this short course.

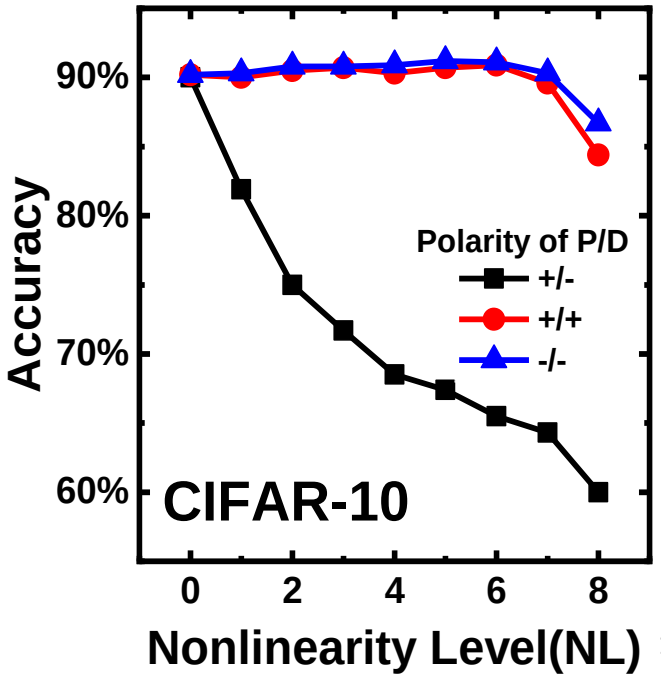
Key Device Properties for Training

- Symmetry and linearity in weight update







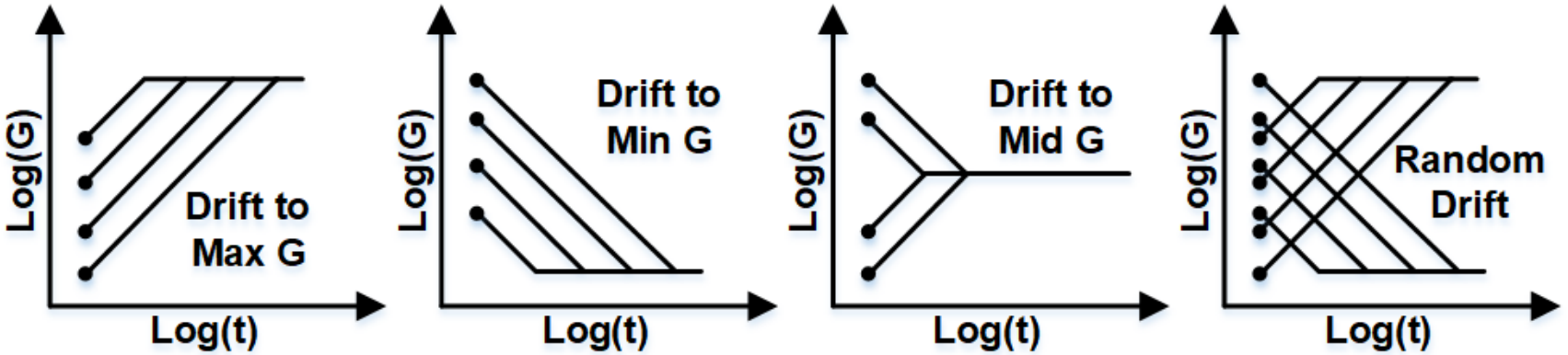


- **Asymmetry (w/nonlinearity) is the primary cause of the in-situ training accuracy degradation.**
- **Algorithmic techniques such as momentum [1] has been introduced to compensate for the accuracy loss.**

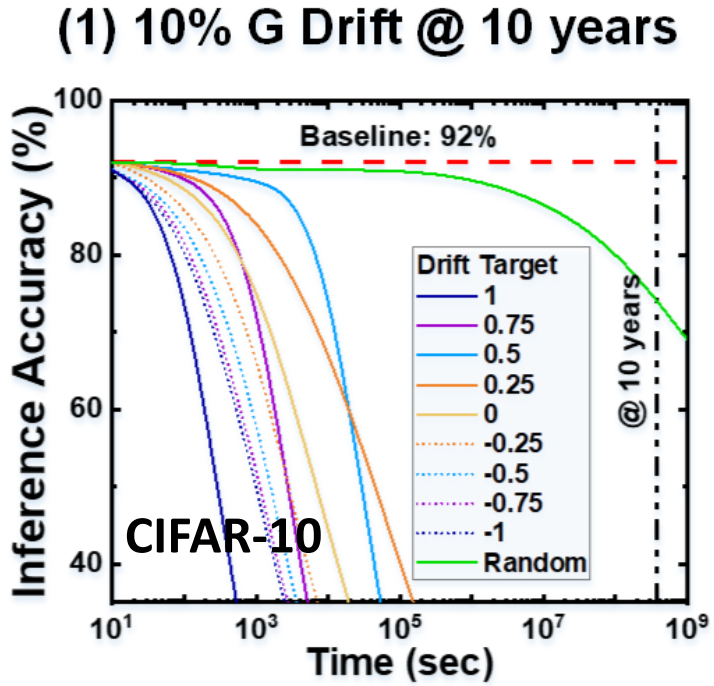
[1] S. Huang, et al. DATE 2020

Key Device Properties for Inference

After training, the weights should be stable over time for inference (read only)



Four possible data drift schemes

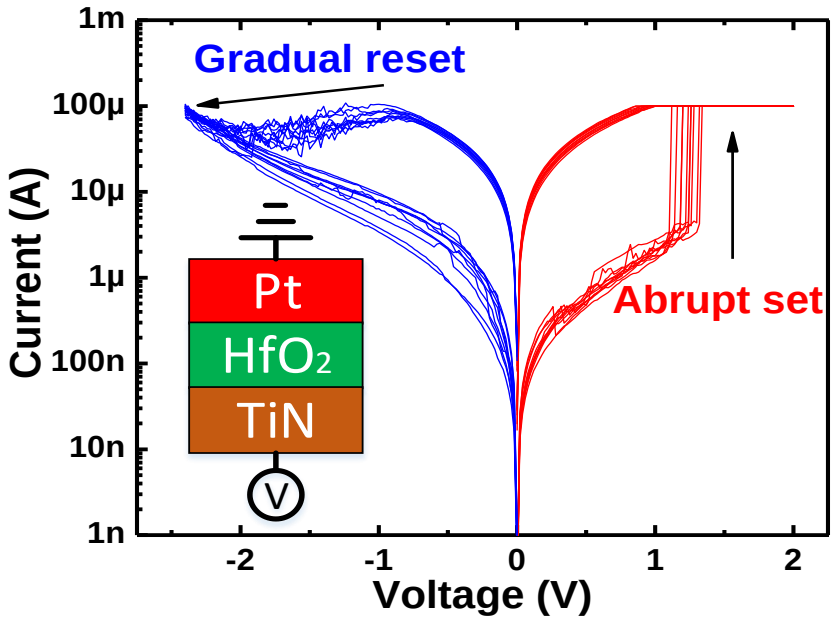


X. Peng, et al. IEDM 2019

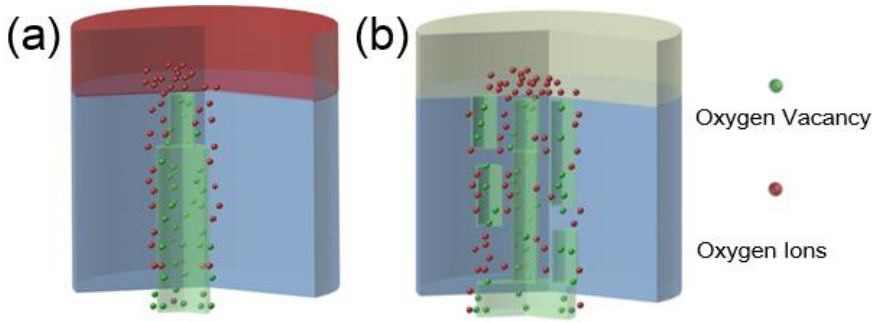
Device effects that affect reliability

- **Relaxation (after programming)**
- **Read stress or disturb**
- **Retention at high temperature**
- **Intermediate state stability is the key concern**

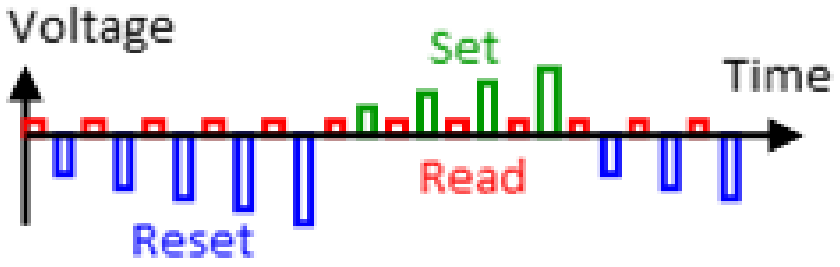
Multi-bit RRAM

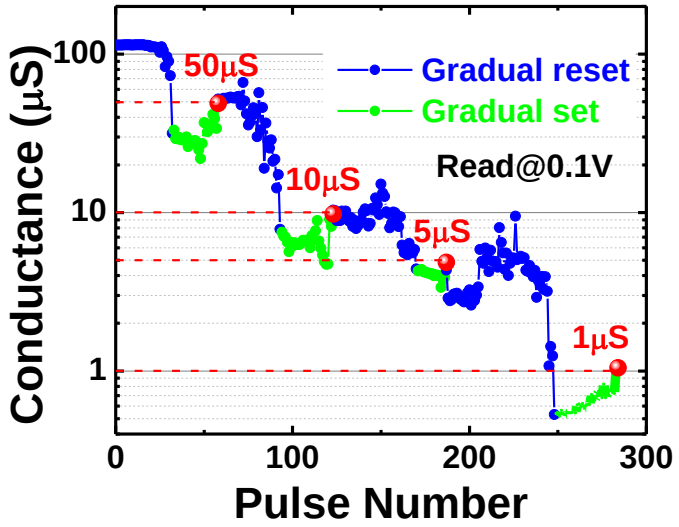


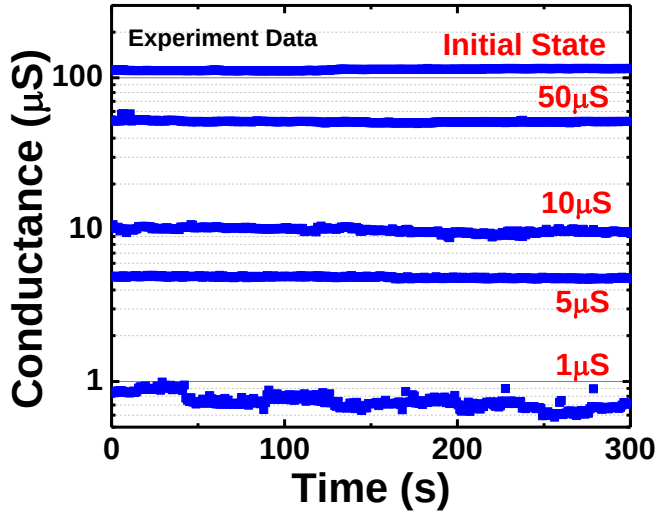
Ultimate Goal: Engineer for multiple weak filaments instead of a single strong filament



Varying-pulse amplitude scheme in the gradual reset regime to converge the target conductance into arbitrary analog levels within the dynamic range.

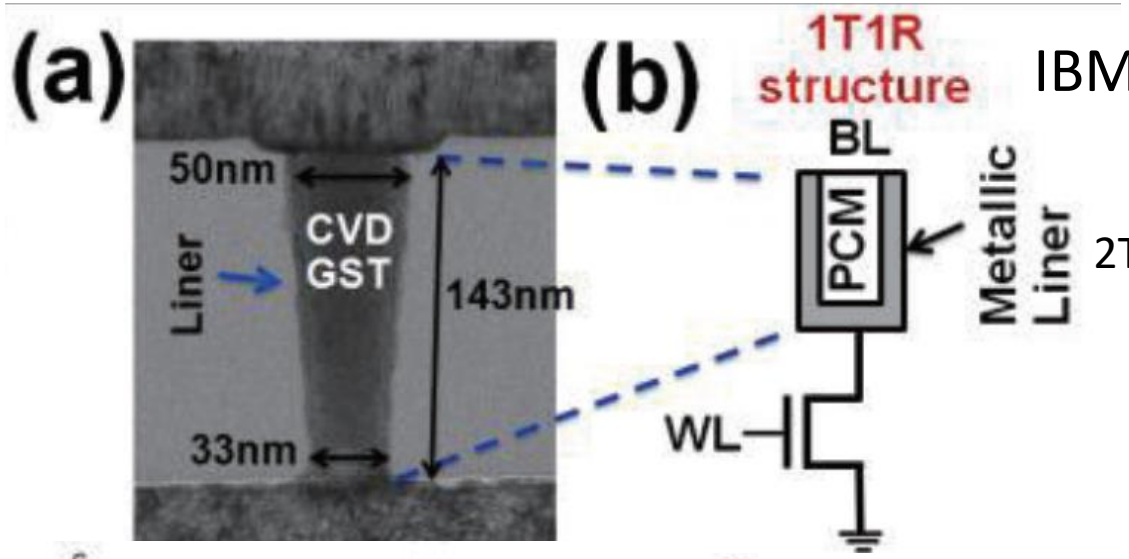




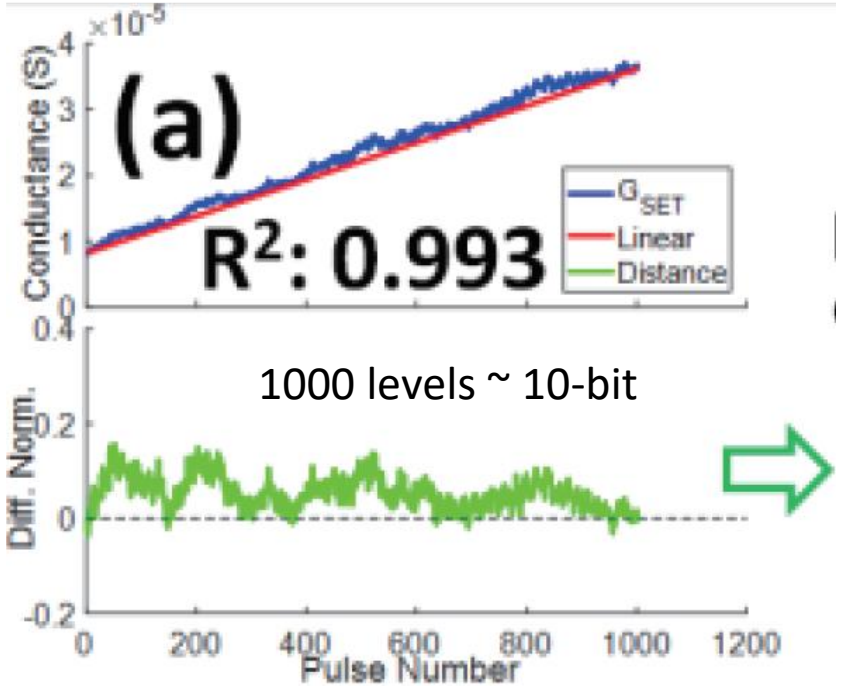


L. Gao, et al. IEEE EDL vol. 36, no. 11, pp. 1157–1159, 2015.

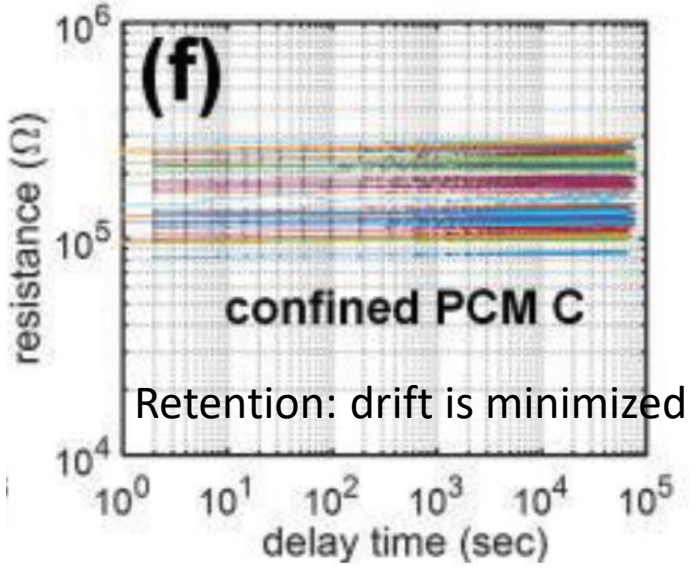
Multi-bit PCM



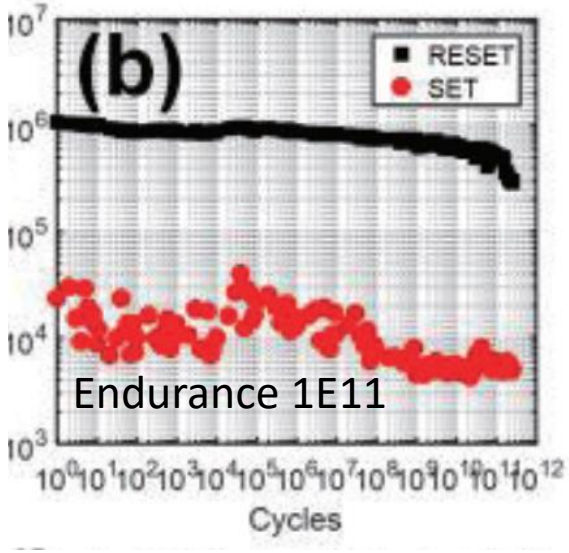
IBM's confined PCM
Metallic liner
Only gradual SET
2T2R differential cell (for bidirectional tuning)



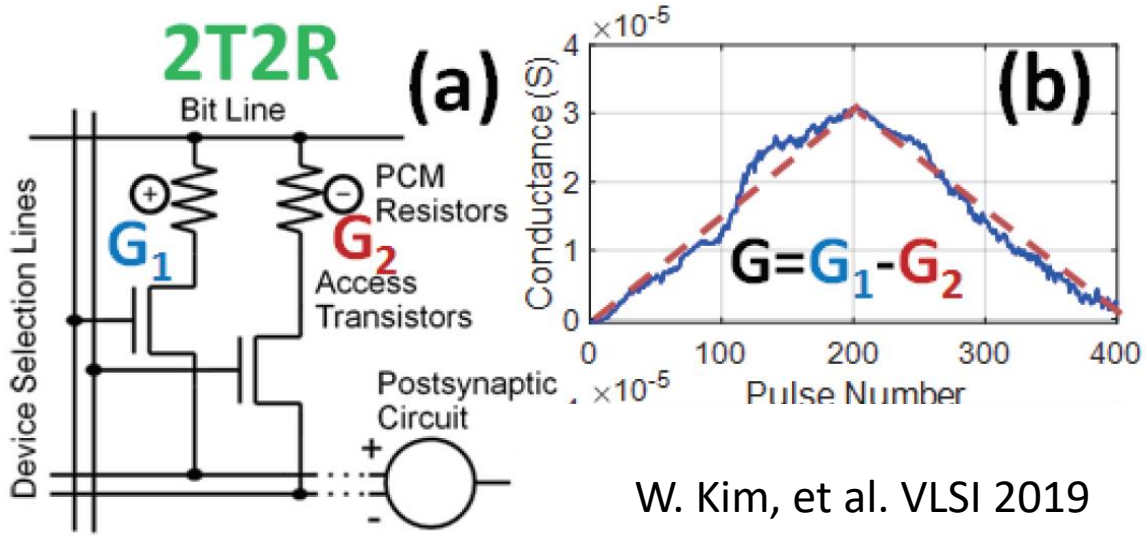
1000 levels ~ 10-bit



Retention: drift is minimized



Endurance 1E11

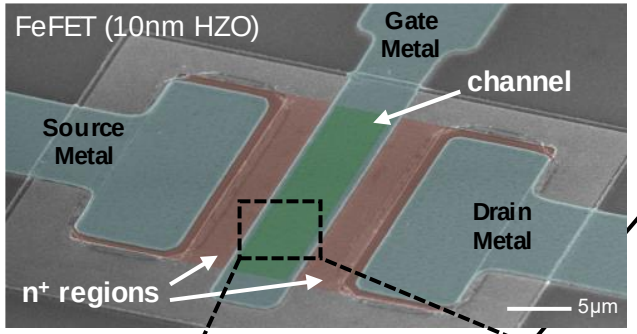


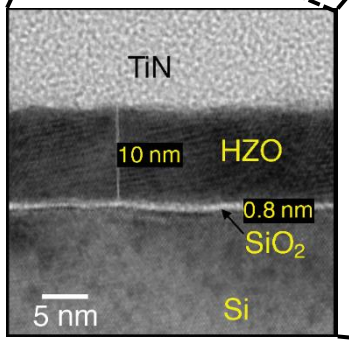
W. Kim, et al. VLSI 2019

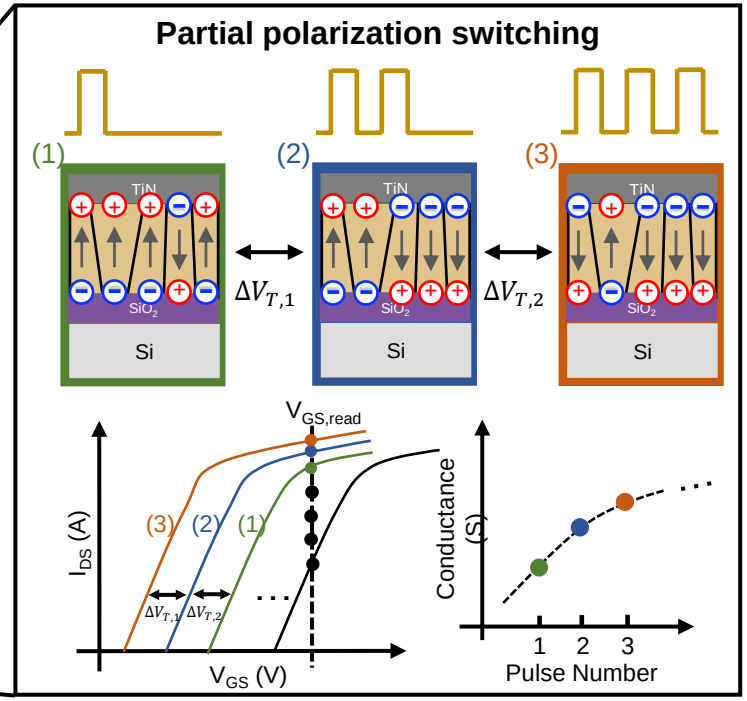
FeFET

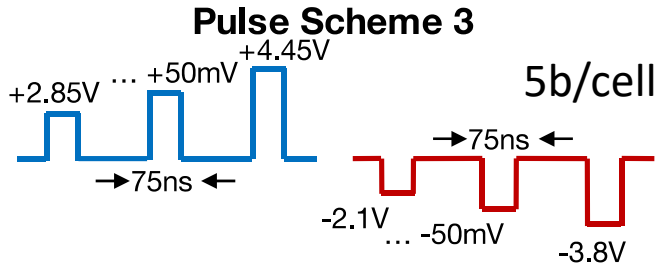
Notre Dame's FeFET based on $Hf_{0.5}Zr_{0.5}O_2$ (HZO) gate stack

M. Jerry, et al. IEDM 2017 & K. Ni, et al. IEDM 2018

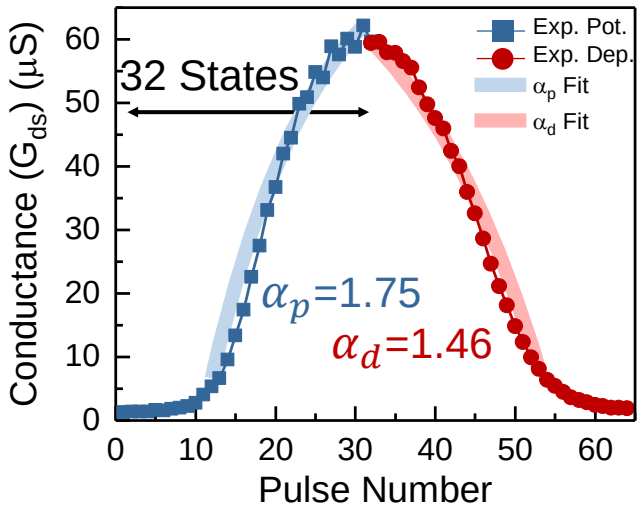


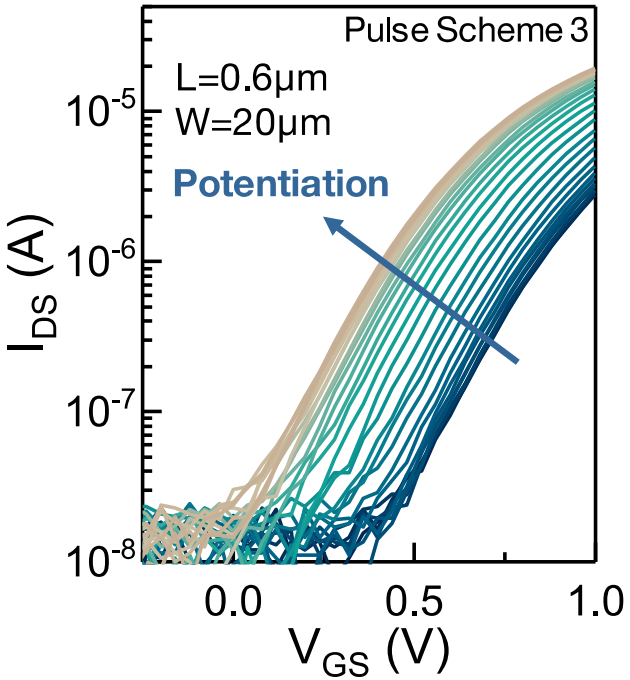


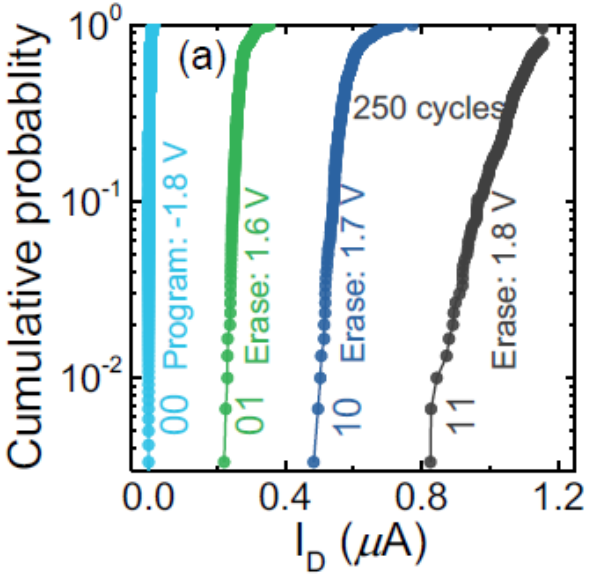




5b/cell for training





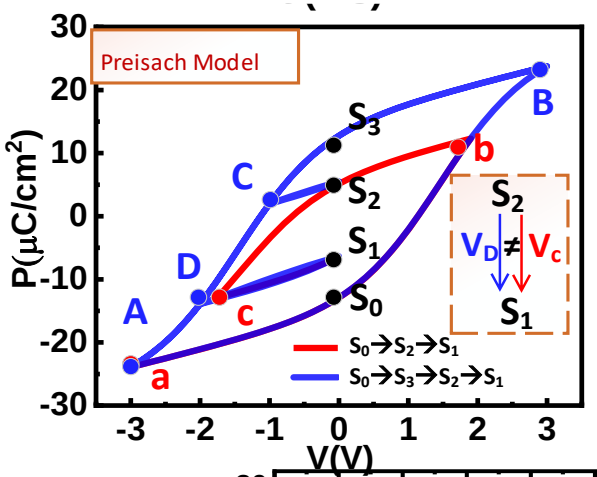


2b/cell for inference

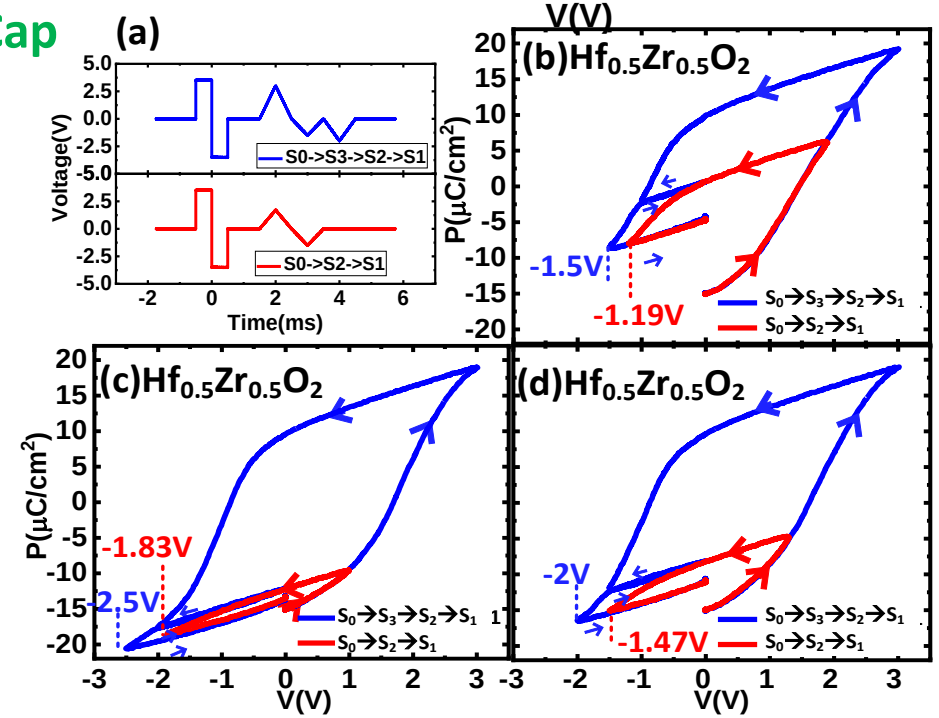
Program voltage down to 1.8V if using FeMFET structure

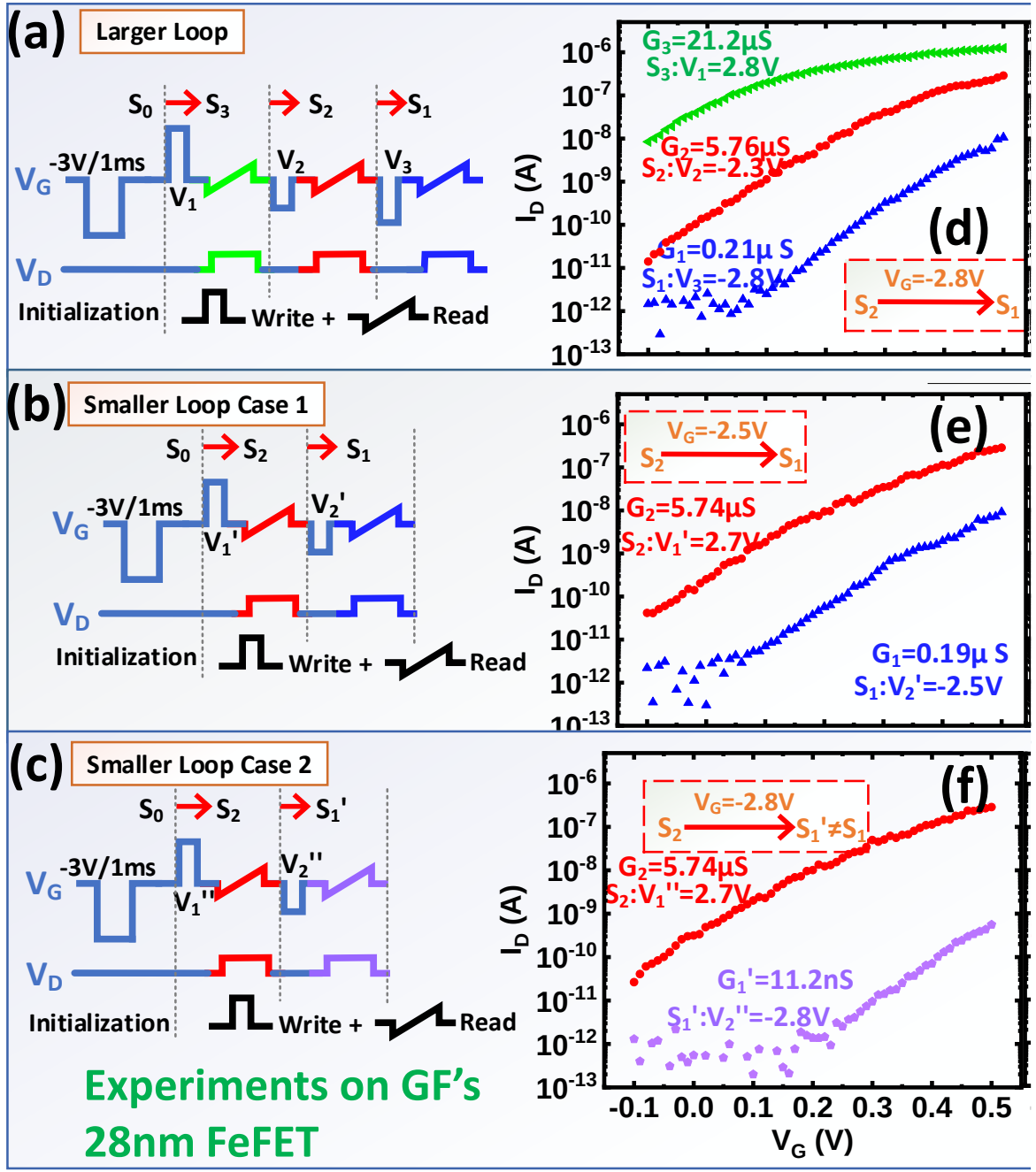
FeFET (History Effect)

Simulation



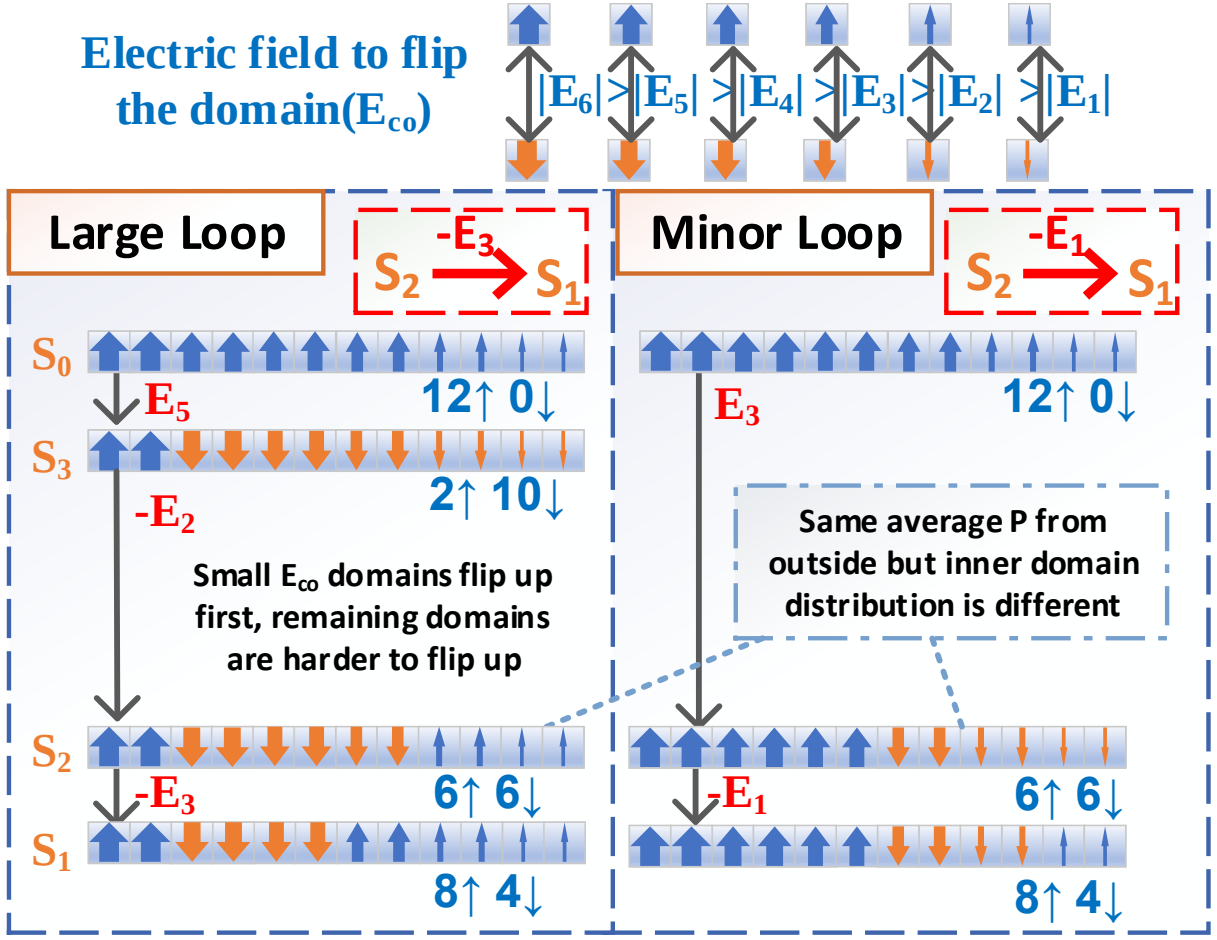
Experiments on FeCap





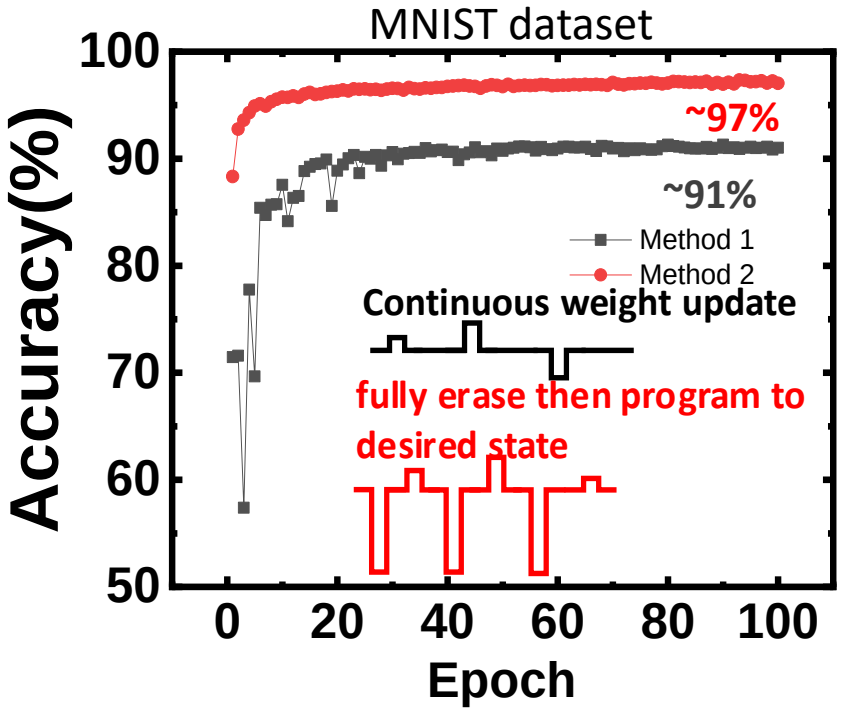
P. Wang, et al. TED 2020

FeFET (History Effect Physics and Mitigation)



Multi-domains have variations in coercive field (Eco), S2 state has more harder domains in large loop, thus needs higher field (E3) than (E1) to flip from S2 to S1 compared to minor loop

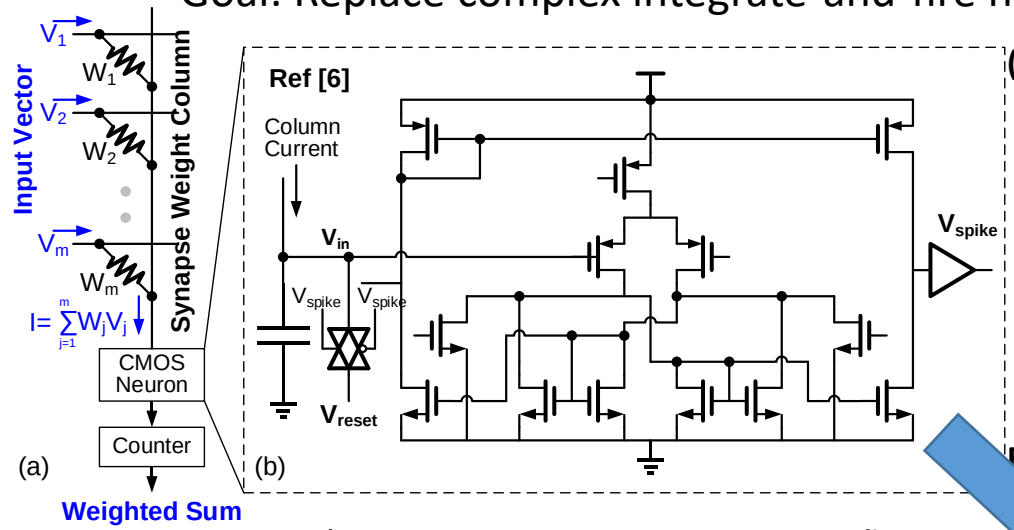
Mitigation: Always erase (to ground state) before program (to intermediate state), to ensure operate on saturation loop

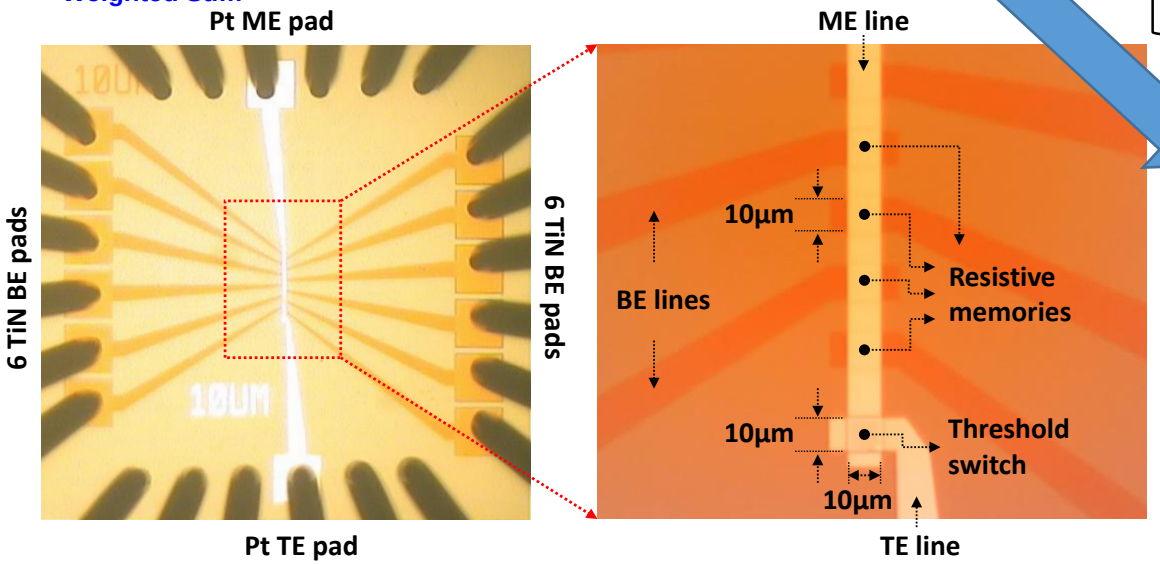


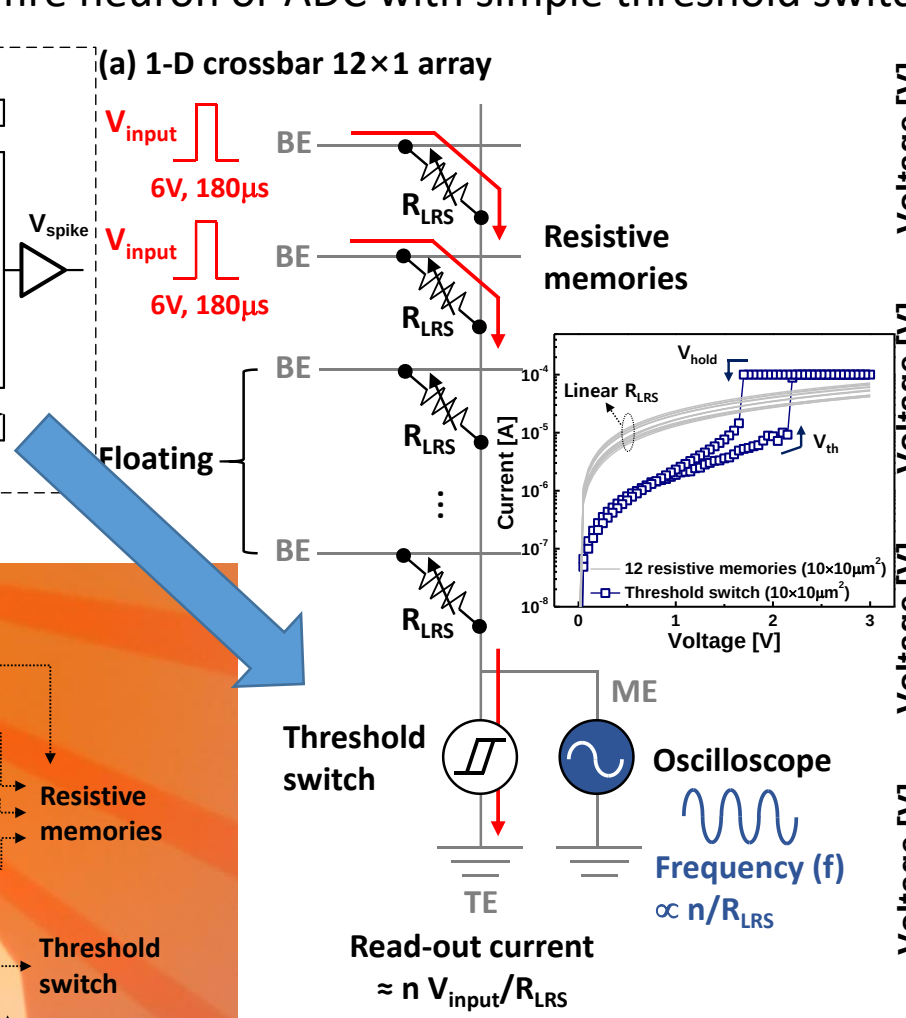
P. Wang, et al. TED 2020

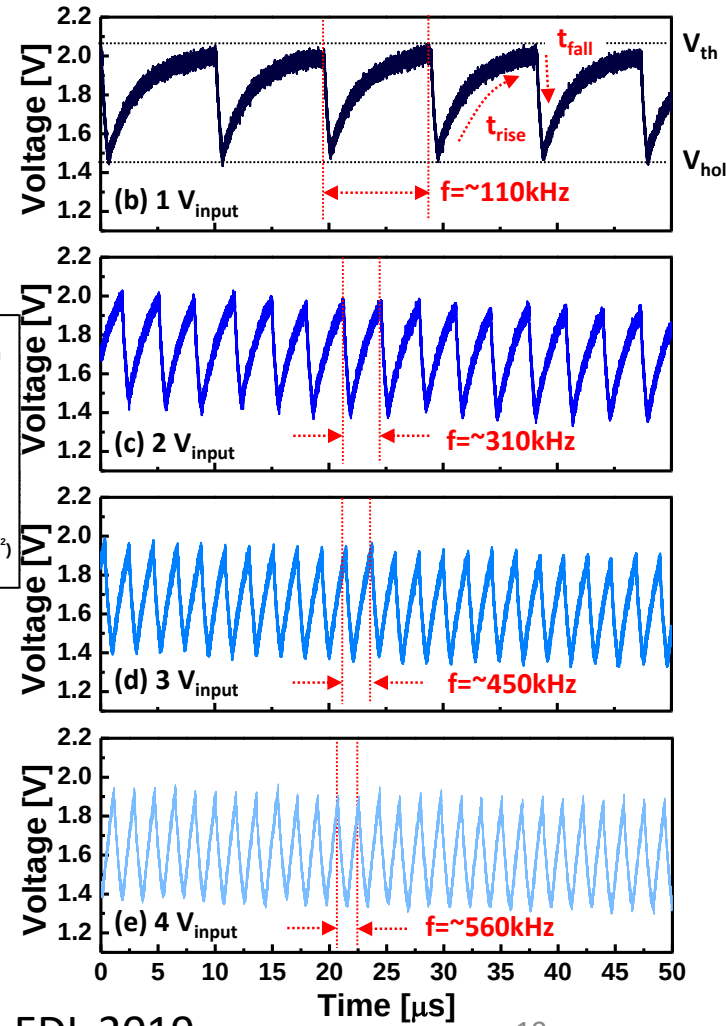
Oscillation Neuron based on Threshold Switch

Goal: Replace complex integrate-and-fire neuron or ADC with simple threshold switch (e.g. NbO_2)







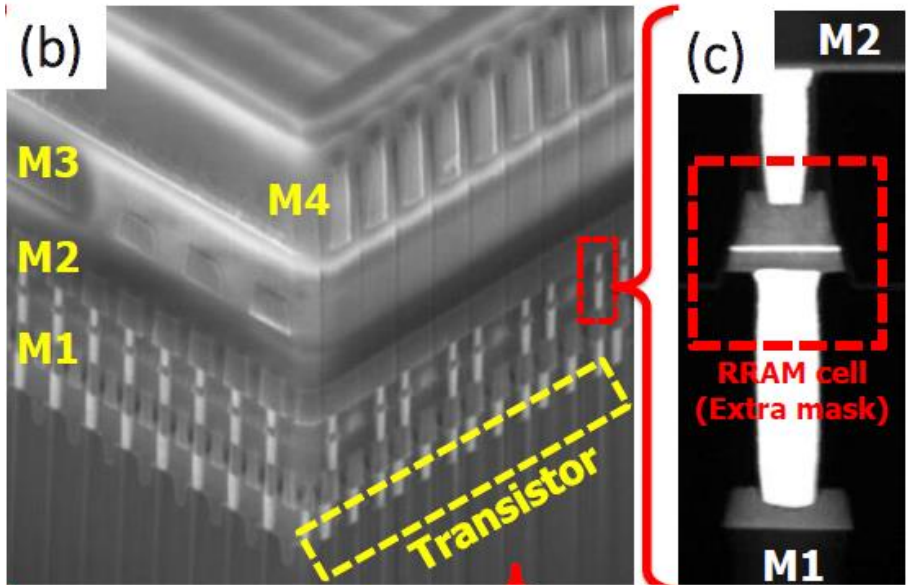


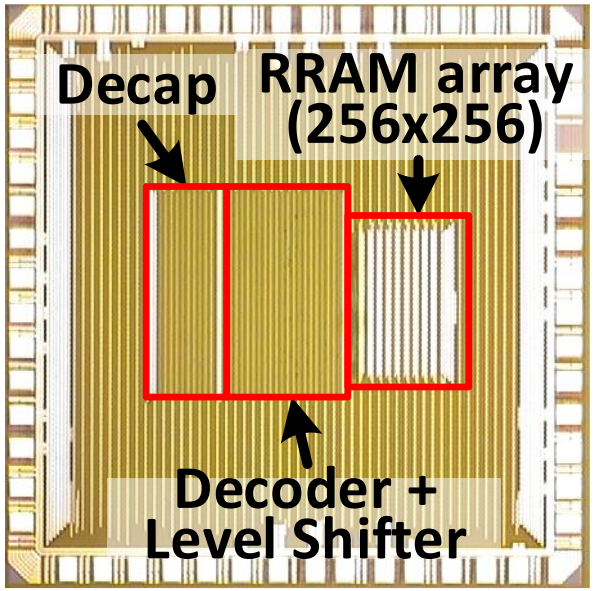
J. Woo, et al. EDL 2019

Outline

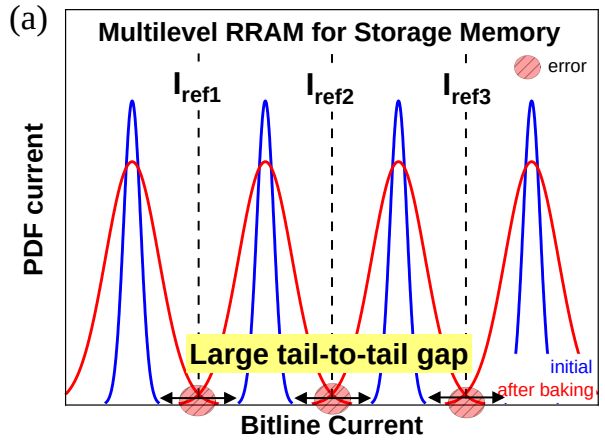
- Background and Motivation
- Synaptic Devices: State-of-the-Art
- Variability and Reliability Characterization at Array-level
- Benchmark of Synaptic Devices for Inference and Training
- Chip-level Demonstrations: State-of-the-art

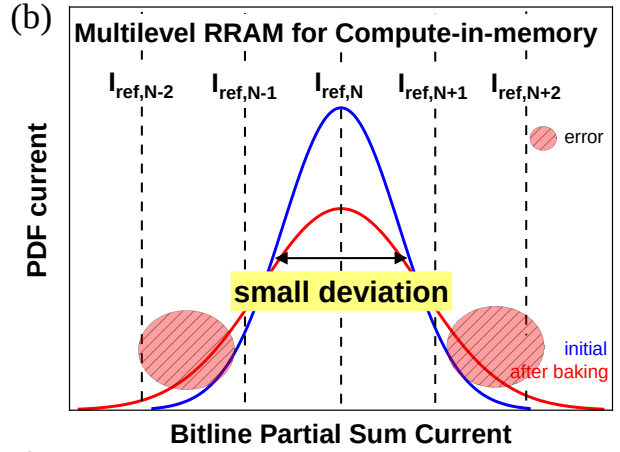
RRAM Test Vehicle for Multilevel Characterization





(Gatech)

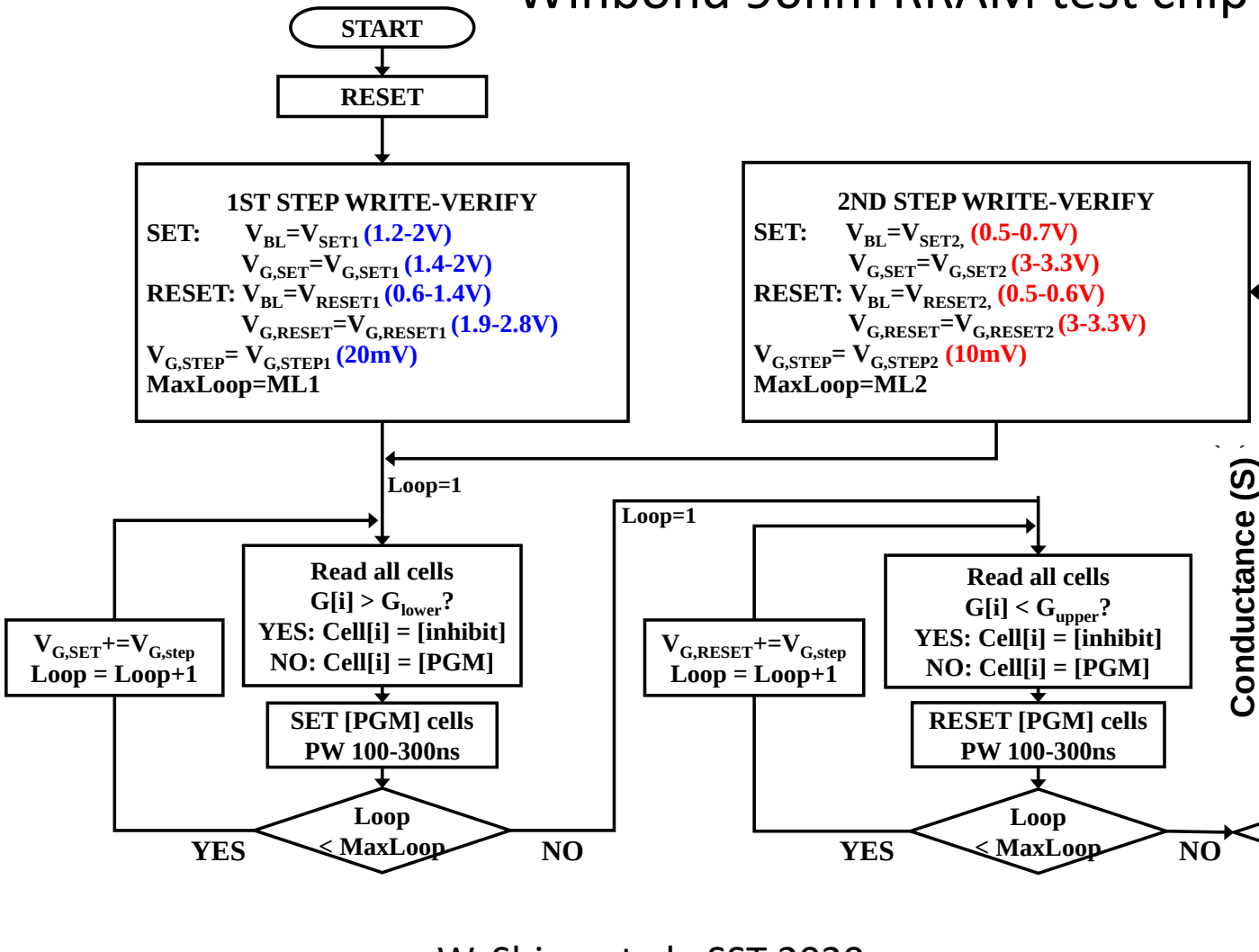


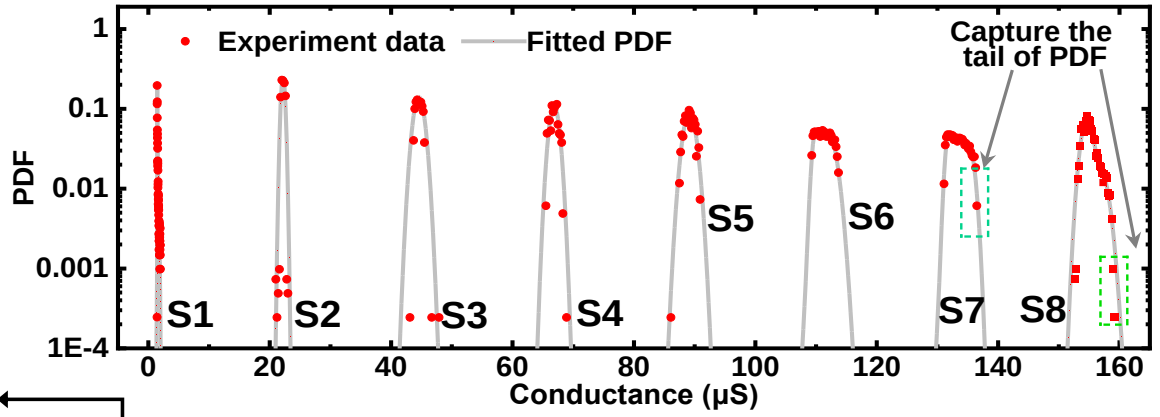


- Winbond HfOx RRAM at 90nm (C. Ho, et al. IEDM 2017)
- RRAM is integrated between M1 and M2
- Originally developed for binary cell operation, now explored for multilevel operations
- Variability and reliability are characterized on 256x256 test vehicle with CMOS decoder
- For MLC storage, tail-to-tail gap is important; for compute-in-memory, the small deviation around center of each state is important. Therefore, the requirement is more stringent for analog synapse

Write-Verify Protocol to Tighten RRAM States

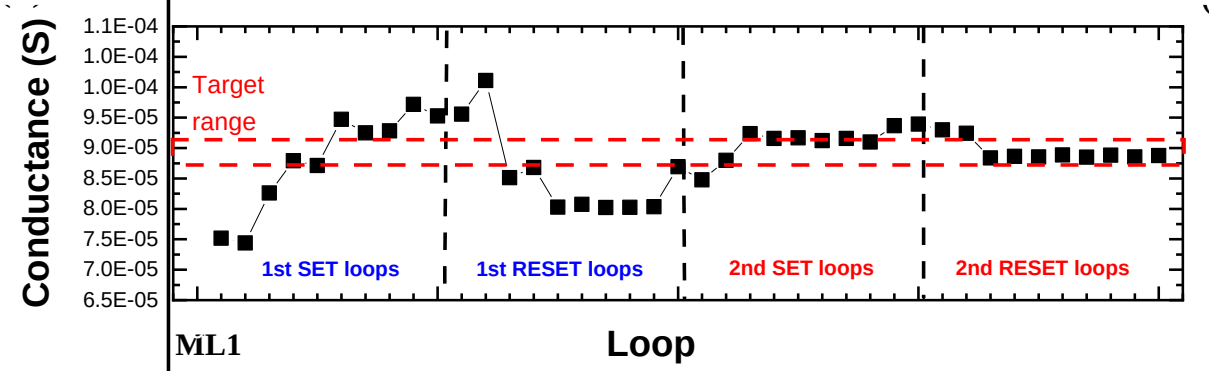
Winbond 90nm RRAM test chip





Demonstrated 8 levels (3-bit)

Sigma <1.5% achieved for each level

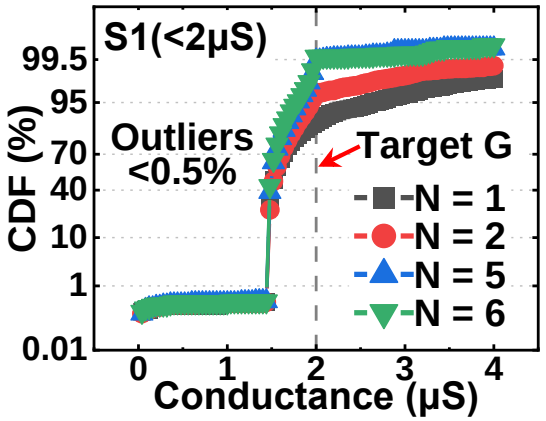


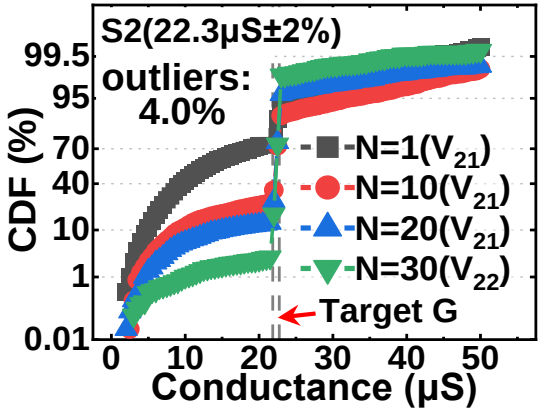
Fine-tune RRAM state

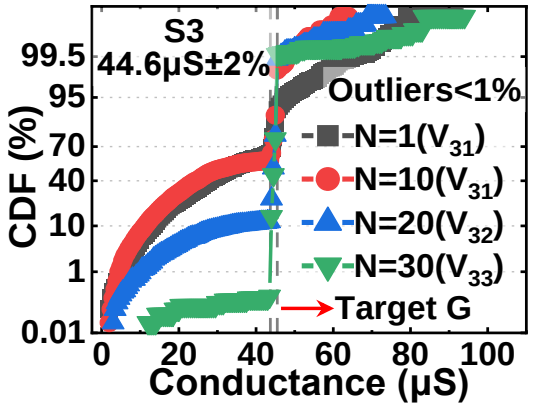
W. Shim, et al., SST 2020

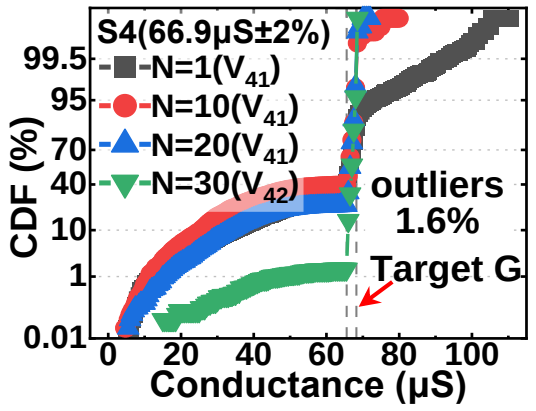
3-bit Weight Programming on RRAM Array

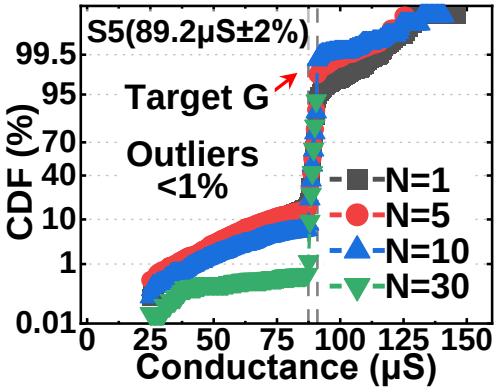
- 4kb cells tested for each state
- Write-verify loop number N

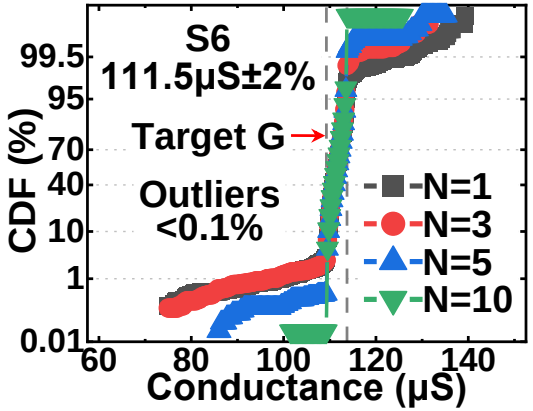


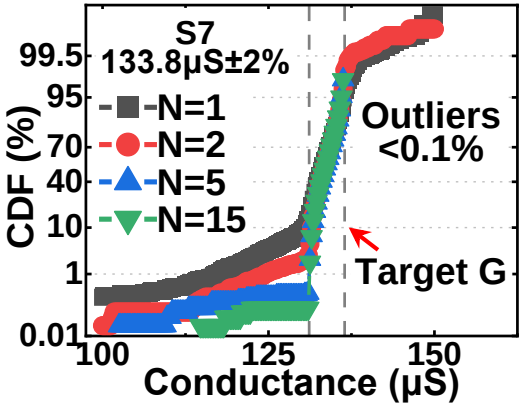


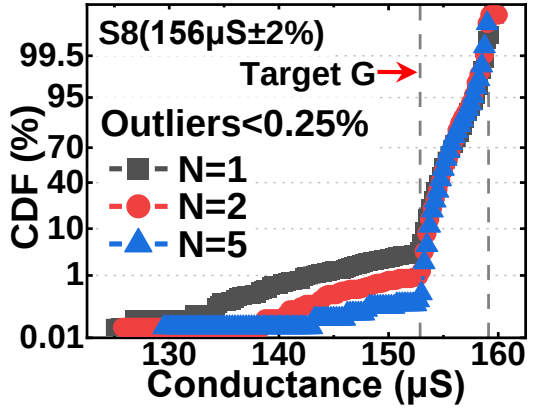






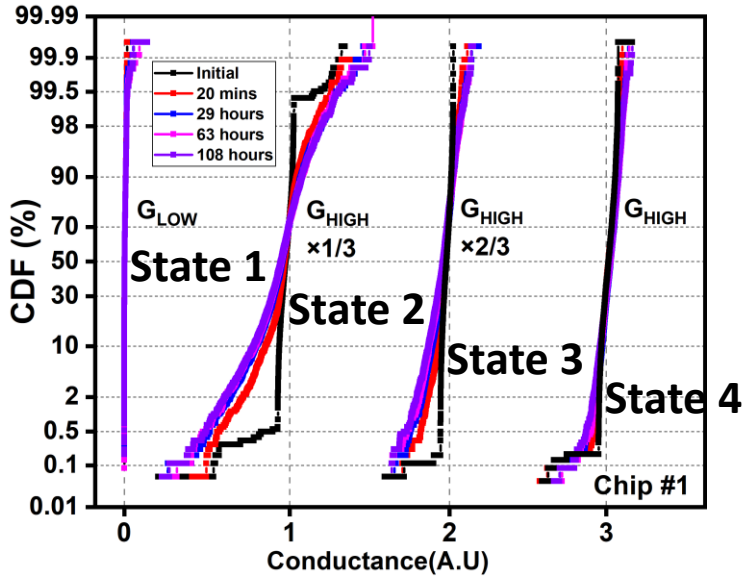






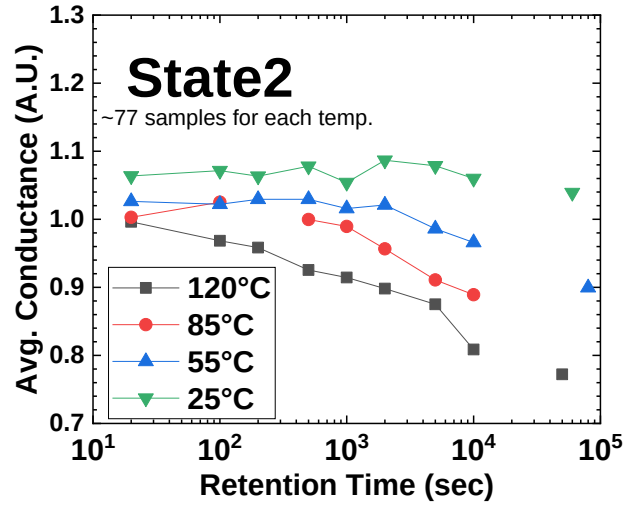
Y. Luo, et al., TED 2020

Multilevel RRAM Stability (for Inference)



Relaxation (after programming)

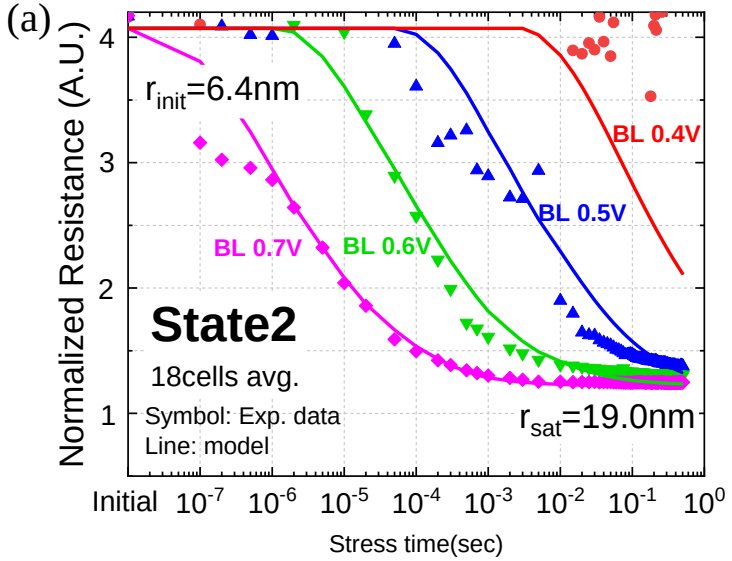
State 2 is the vulnerable state where it has a weak filament

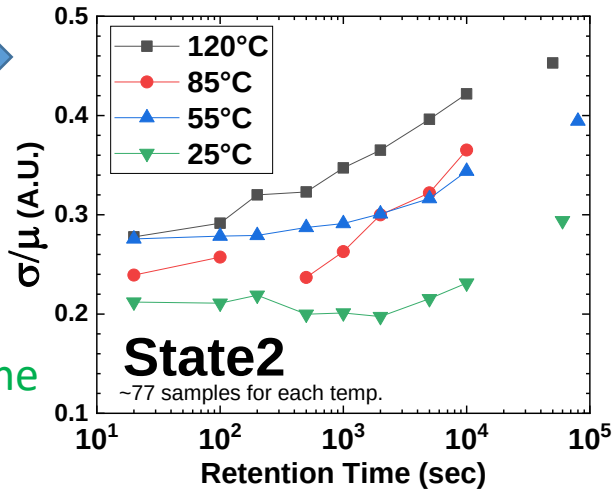


Avg. conductance tends to decrease

High temperature baking

Read disturb or stress





Sigma conductance tends to increase

Algorithm-level mitigation:

- Training with regularization to reduce the weights in intermediate states
- Retraining with batch norm parameters

W. Shim, et al. IRPS 2020 and TED 2020.

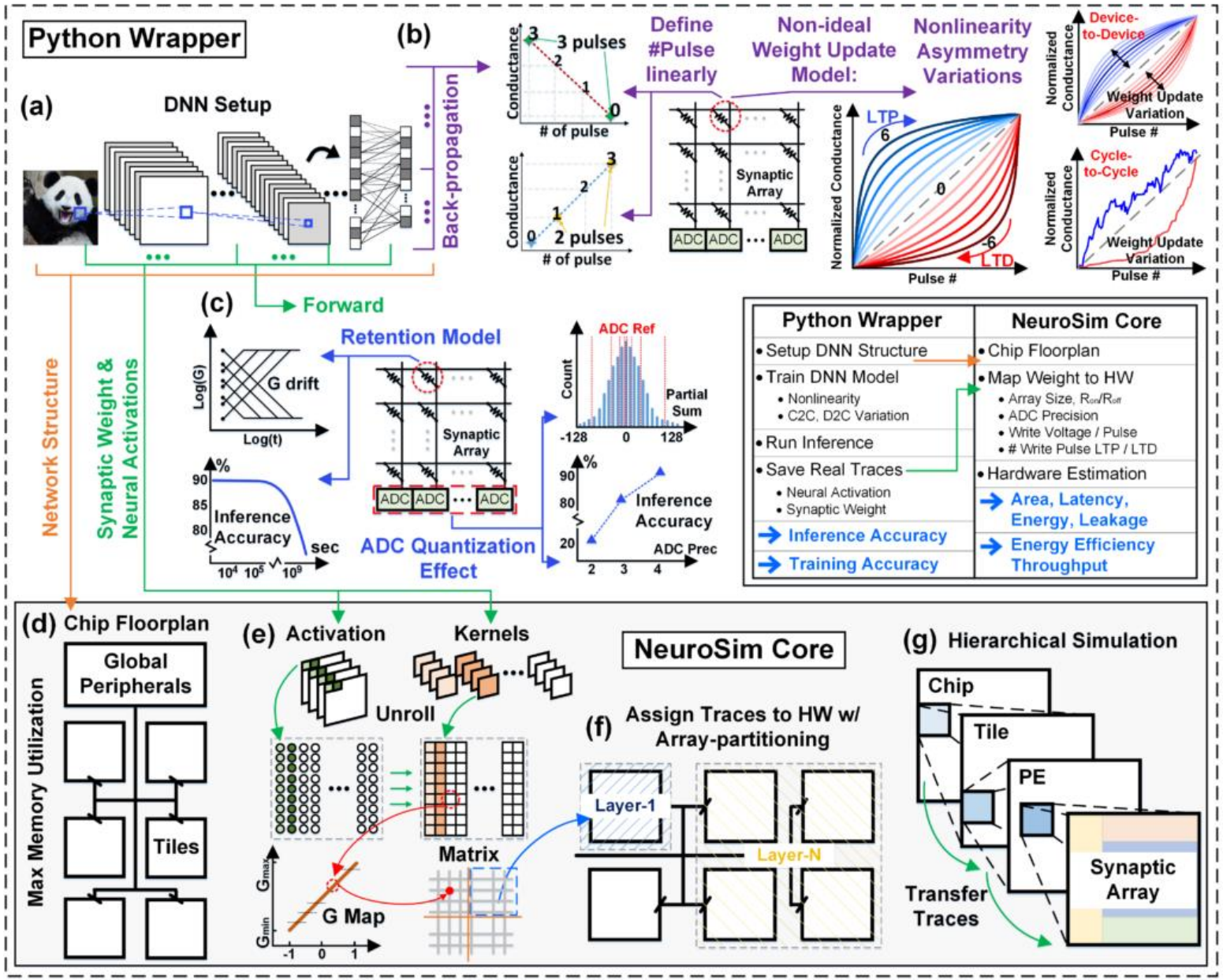
Outline

- Background and Motivation
- Synaptic Devices: State-of-the-Art
- Variability and Reliability Characterization at Array-level
- Benchmark of Synaptic Devices for Inference and Training
- Chip-level Demonstrations: State-of-the-art

DNN+NeuroSim Framework Overview

- Integration of NeuroSim with Pytorch and Tensorflow
  - An end-to-end framework to benchmark configurable CIM-based hardware accelerators
- NeuroSim Core
  - Built upon a hierarchy of chip/tile/PE/subarray with all the necessary peripheral circuitry
  - Technology parameters calibrated with PTM model from 130nm to 7nm
  - Reports energy efficiency, throughput, area and memory utilization
- Python Wrapper
  - Defines arbitrary deep neural network and reports inference/training accuracy
  - Introduced device retention model and ADC quantization effects for inference
  - Introduced device nonlinearity/asymmetry and variation effects for training
- DNN+NeuroSim V1.3 for inference
  - Github: https://github.com/neurosim/DNN_NeuroSim_V1.3
- DNN+NeuroSim V2.1 for training
  - Github: https://github.com/neurosim/DNN_NeuroSim_V2.1
- Community: more than 300 users including Intel, TSMC, Samsung, and SK Hynix

DNN+NeuroSim Key Features



X. Peng, et al. IEDM 2019

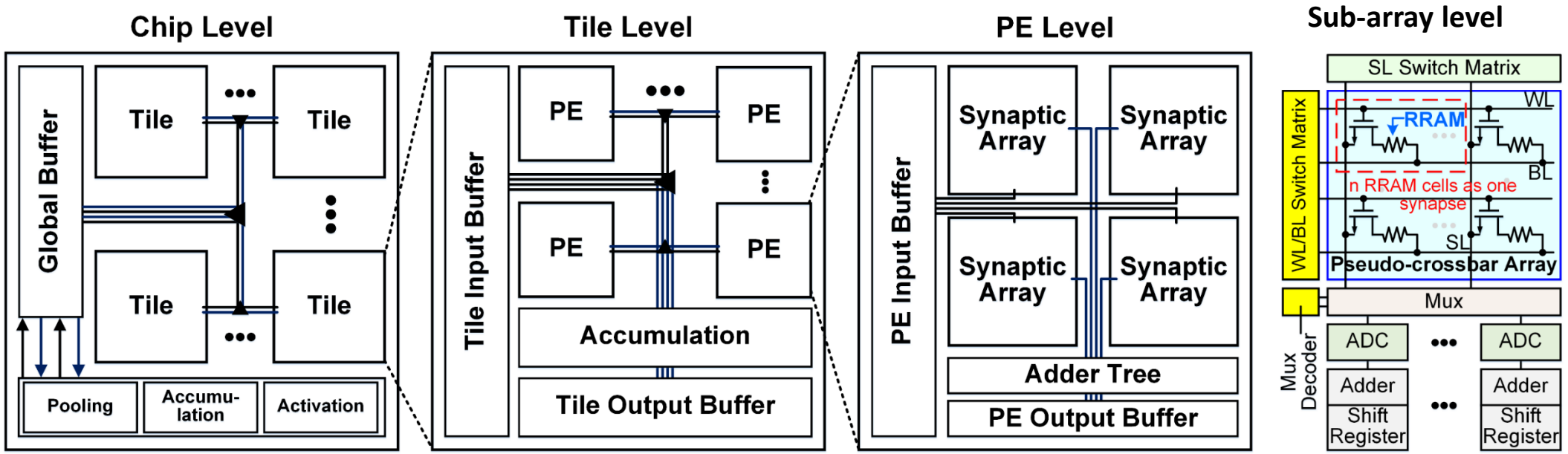
DNN+NeuroSim Methodologies

Algorithm accuracy estimation based on WAGE method

- Hardware-aware quantization for weight, activation, gradient, error, as well as partial sum quantization based on ADC precision.
- Support various network models for CIFAR-10/-100 and ImageNet

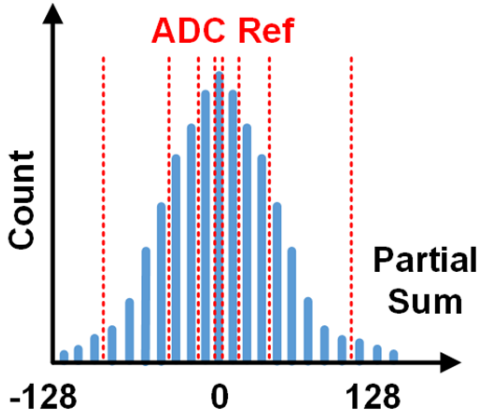
Hardware metrics estimation based on analytic models that are calibrated with SPICE at module-level.

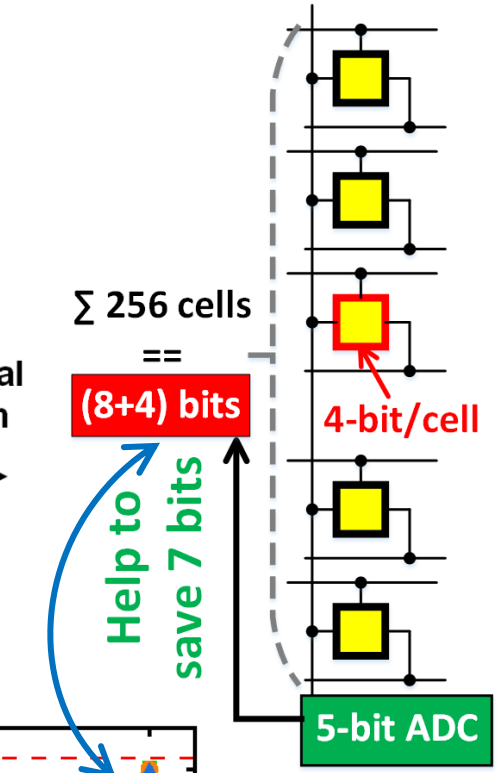
- Analog modules (e.g. ADC) calibrated with Cadence custom simulation;
- Digital modules estimated with standard cell area and logic gate delay/dynamic power/leakage power;
- Interconnect modules (e.g. H-tree) estimated with parasitic RC delay and power;

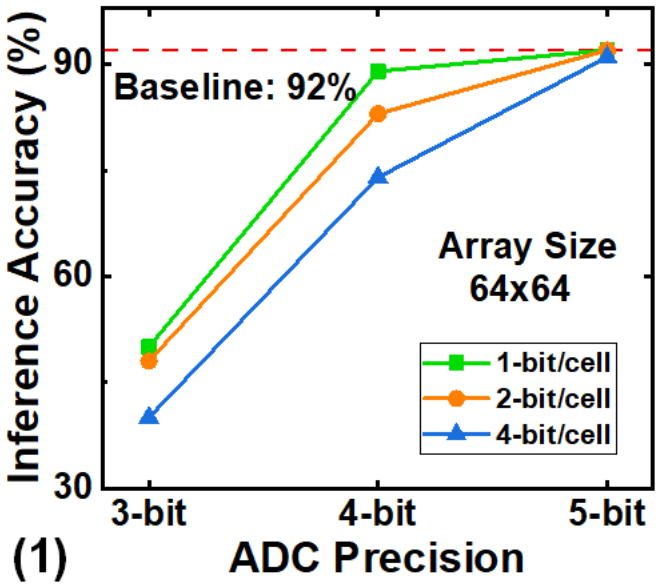


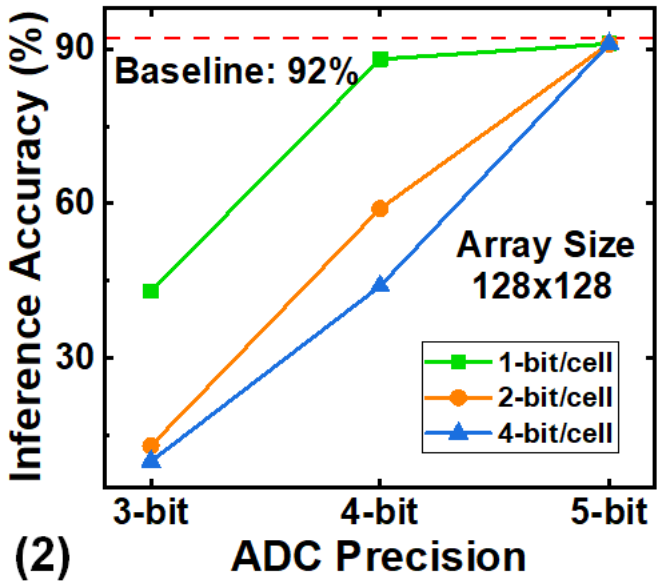
Analysis on ADC Precision

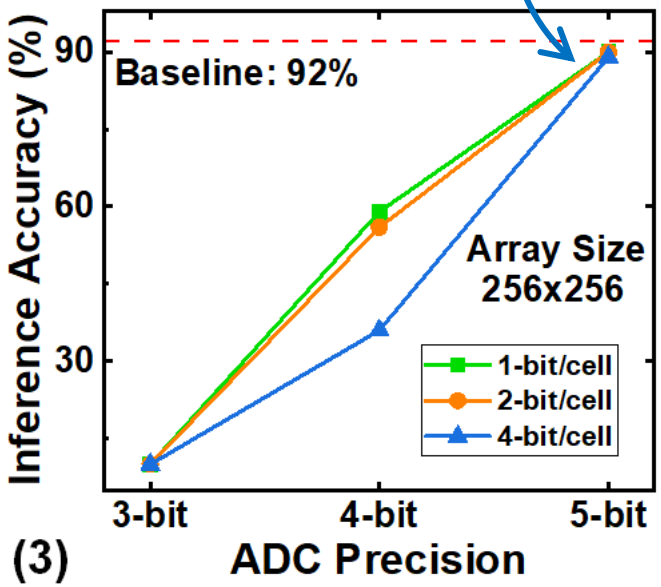
- Inference Accuracy of VGG-8 (8-bit weight) on CIFAR-10
  - Sweep device precision & synaptic array size
  - Sweep ADC precision (non-linear quantization)











X. Peng, et al. IEDM 2019

Benchmark for Compute-in-Memory (Inference)

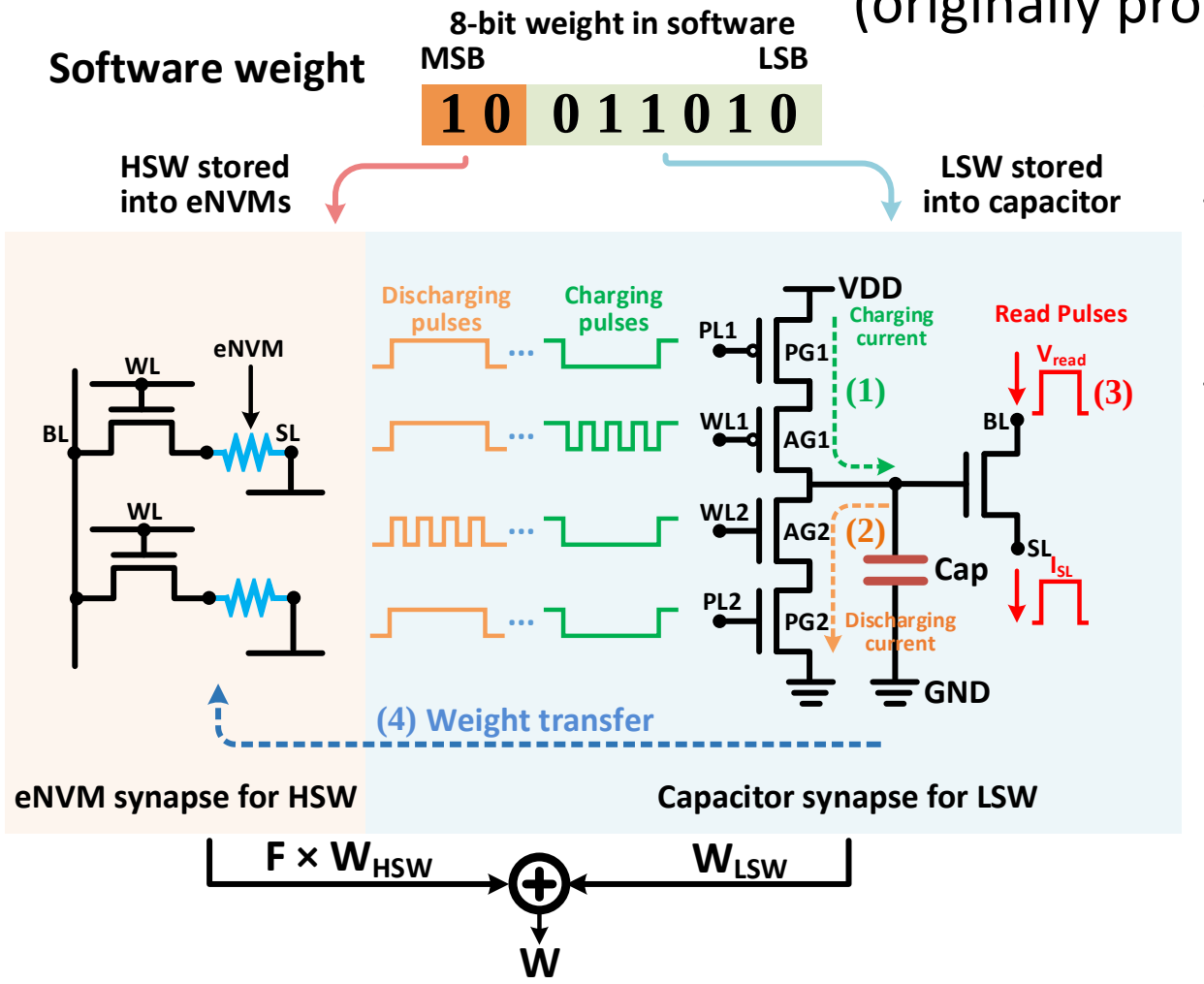
VGG-8 (8-bit activation; 8-bit weight) on CIFAR10, with Novel Weight Mapping and Dataflow								
Technology node (LSTP)	**7nm**		**22nm**					
Device	**8T-SRAM**	**TPU-like (Google)**	**8T-SRAM**	**RRAM (Winbond)**	**RRAM (Tsinghua)**	**PCM (IBM)**	**Si:HfO2 FeFET (GF)**	**TPU-like (Google)**
MLSA-ADC precision **Memory Cell Precision**	4-bit 1-bit	8-bit digital MAC	4-bit 1-bit	5-bit 2-bit	5-bit 4-bit	5-bit 4-bit	5-bit 4-bit	8-bit digital MAC
Ron (Ω)	\			6k	**100k**	**40k**	**240k**	\
On/Off Ratio	\			150	10	12.5	100	\
Inference Accuracy (%)	92%			91%				92%
Area (mm^2)	**13.61**	15.71	59.05	60.78	**33.26**	**33.39**	**32.69**	**107.05**
Memory Utilization (%)		39.36%	98.73%	96.86%	93.47%	93.47%	93.47%	39.36%
L-by-L Latency (ms)	0.63	0.53	0.76	0.79	0.61	0.61	0.61	0.75
L-by-L DynamicEnergy (uJ)	22.86	526.21	56.56	33.14	17.24	17.85	16.28	687.25
L-by-L Leakage power (mW)	1.47	75.96	1.11	0.17	0.09	0.09	0.09	5.27
Energy Efficiency (TOPS/W)	**51.100**	**1.110**	21.360	36.980	**71.170**	**68.730**	**75.360**	**0.690**
Compute Efficiency (TOPS/mm2)	**0.144**	**0.075**	0.027	0.026	**0.061**	**0.060**	**0.060**	**0.004**

- Emerging NVMs outperform SRAM at the same tech node (e.g. at 22nm)
- Increasing on-state resistance (Ron) to >100kΩ is critical to improve the energy efficiency (TOPS/W)
- FeFET is promising due to high R_{on} that is modulated by the gate voltage bias
- 7nm SRAM (if compute-in-memory) still achieves the best compute efficiency with area scaling advantage
- Compared to IEDM 2019 results, here we added the level shifter module for NVMs that need high write voltage

X. Peng, et al. IEDM 2019

Hybrid NVM+Capacitor for Training

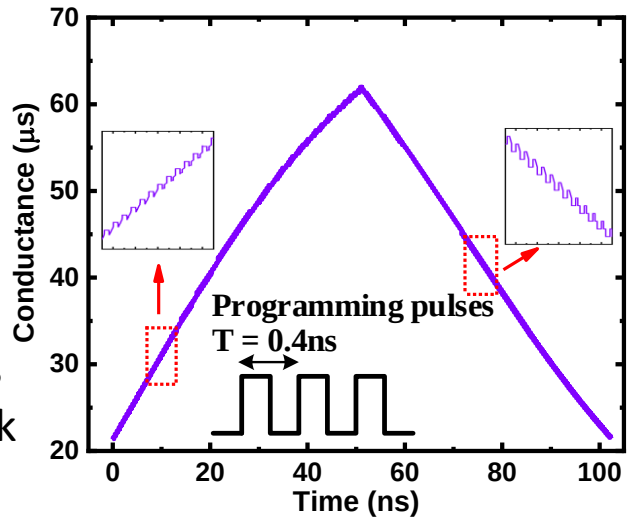
(originally proposed by IBM)

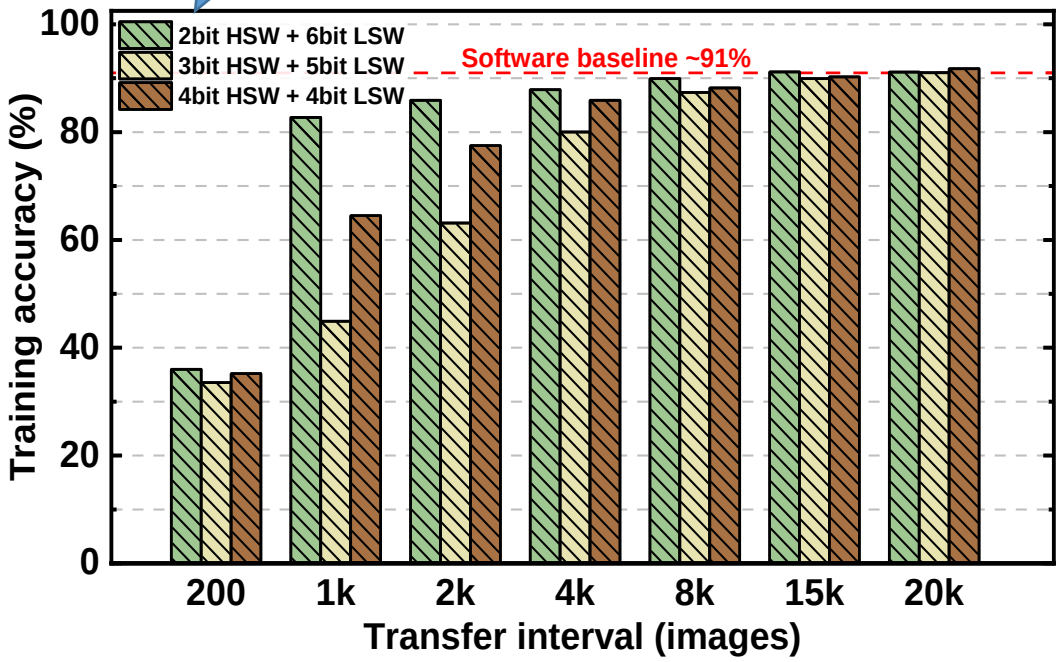


Weight representation with hybrid precision synapse

Symmetric and fast charging/discharging from capacitor

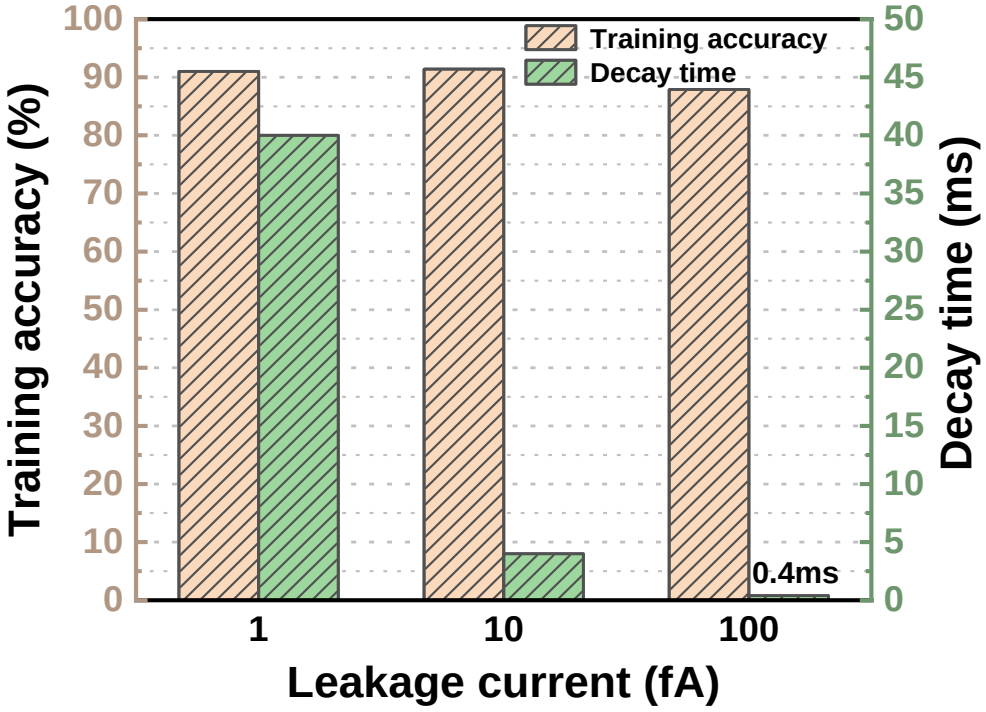
Leakage to the cap limits the transfer interval to 8k





Y. Luo, et al. T-computer 2020.

Cap Leakage and Endurance Requirement



- 10fA or below is needed for maintaining the retention time above ms and ensure no training accuracy loss
- Oxide channel transistor may be preferred with low leakage and large drive voltage to program NVM.

Dataset	Number of write	
	HPS	Pure NVM
CIFAR-10	750	37,500
ImageNet	20,000	6,250,000

Transfer interval = 10k images as example

For HPS, # write = # images per epoch * # epochs / transfer interval
CIFAR-10: 50k * 150 epoch / 10k = 750
ImageNet: 1M * 200 epoch / 10k = 20, 000

For pure NVM, # write = # images per epoch * # epochs / batch size
CIFAR-10: 50k * 150 epoch / 200 = 37,500
ImageNet: 1M * 200 epoch / 32 = 6,250,000

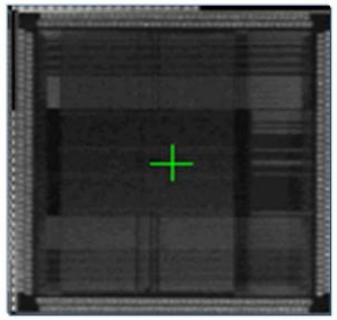
Outline

- Background and Motivation
- Synaptic Devices: State-of-the-Art
- Variability and Reliability Characterization at Array-level
- Benchmark of Synaptic Devices for Inference and Training
- Chip-level Demonstrations: State-of-the-art

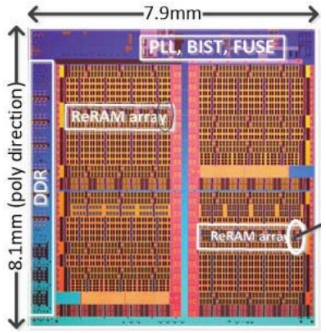
State-of-the-Art Industrial Emerging NVMs

- A survey of the industrial platforms (developed for embedded memories, not necessarily tailored for synaptic weights)

RRAM:

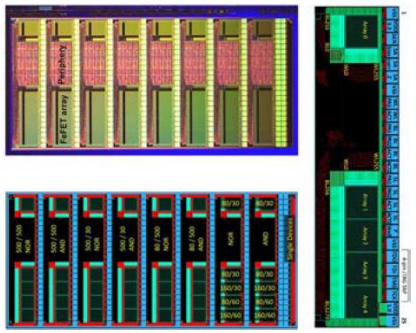


TSMC 40nm RRAM (ISSCC 18)



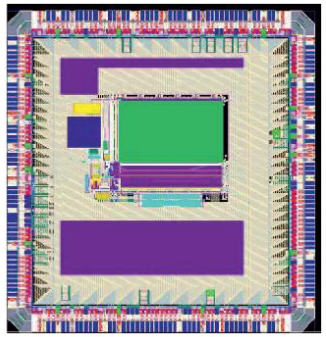
Intel 22nm RRAM (ISSCC 19)

FeFET:



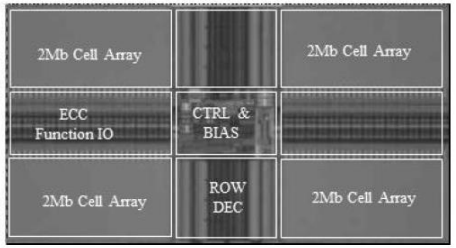
GF 28nm and 22nm FeFET (IEDM 16 & 17)

PCM:

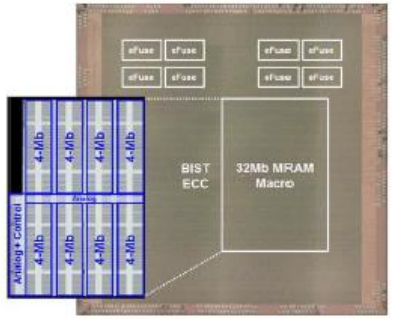


TSMC 40nm PCM (IEDM 18)

STT-MRAM:

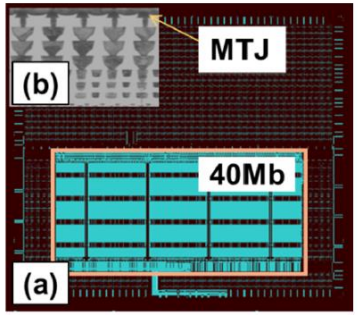


Samsung 28nm STT (IEDM 18)



TSMC 22nm STT (ISSCC 20)

Intel 22nm STT (ISSCC 19)



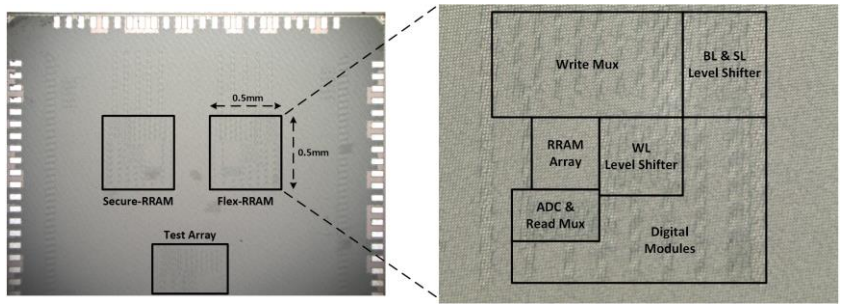
GF 22nm STT (IEDM 19)

Summary of RRAM-based CIM Macros

	ISSCC' 18 NTHU	ISSCC' 19 NTHU	ISSCC' 20 NTHU	TED' 20 ASU/GT	SSCL' 20 ASU/GT
Technology (nm)	65	55	22	90	90
No. of bit per cell	1	1	1	1	2
Subarray size	512×256	256×512	512×512	128×64	128×64
Capacity	1Mb	1Mb	2Mb	8Kb	8Kb
Precision(I,W,O)	1,1,3	2,3,4	4,4,11	1,1,3	1,2,1
Column sensing	3b ADC	4b ADC	6b ADC	3b ADC	1b SA
# of rows turned on	9	9	16	64	64
Supported algorithm	CNN	CNN	CNN	CNN	CNN
Energy efficiency	0.6 TOPS/W	2.05 TOPS/W	3.79 TOPS/W	0.38 TOPS/W	1.61 TOPS/W
Accuracy	98% (MNIST)	88.5% (CIFAR10)	90.18% (CIFAR10)	83.5% (CIFAR10)	87.1% (CIFAR10)

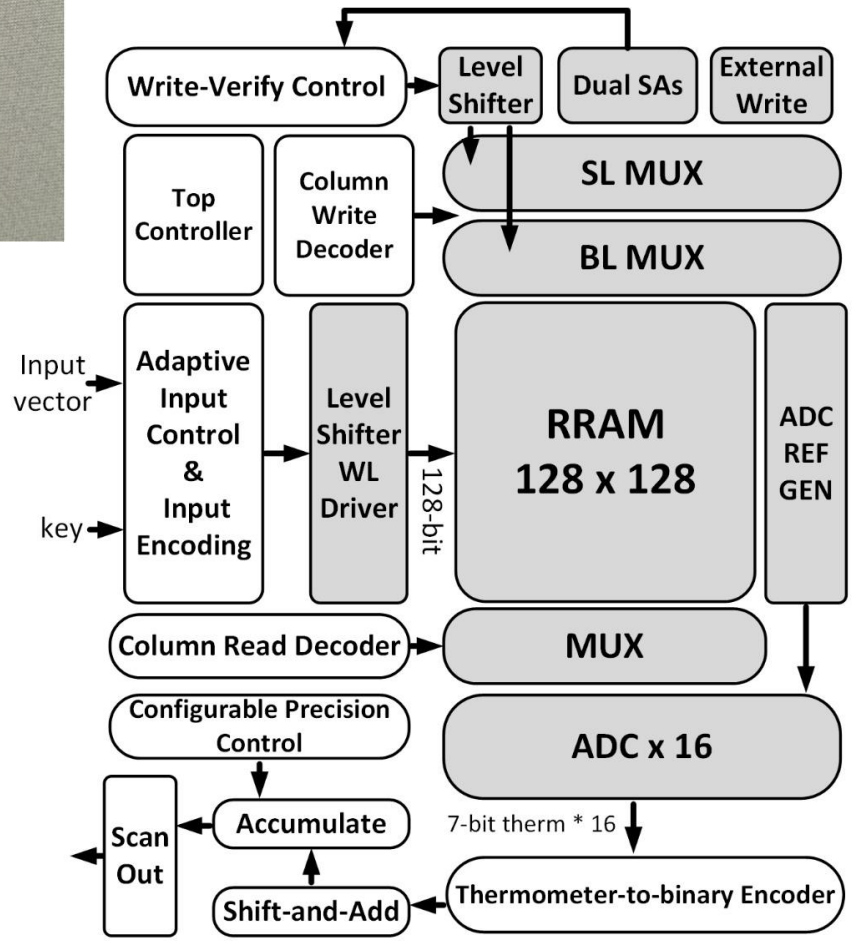
Note: TOPS/W is normalized to 8bit by 8 bit MAC (1b MAC = 2 ops)
TOSP/W is less than NeuroSim prediction, due to 1) older tech node, 2) partially # of rows turned-on

Secure-RRAM CIM Prototype Chip (TSMC 40nm)



New Features

- Adaptive input sparsity control
- Reconfigurable weight precision
- Integrated digital compute units
- Input-aware on-chip ADC reference
- On-chip write-verify controller
- Input encoding for embedded security



W. Li, et al. CICC 2021 and ESSCIRC 2021

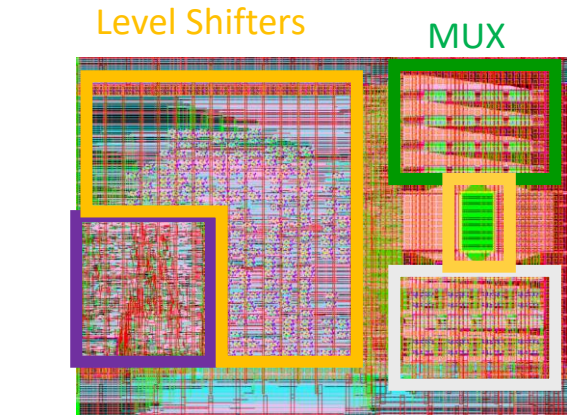
Technology	TSMC 40nm w/ RRAM	
Array size	128 x 128b	
Weight precision (bits)	1, 2, 4, or 8	
Rows turned on simultaneously	7	
Operating voltage	0.9V	
Clock frequency	100MHz	
	0% Input Sparsity	95% Input Sparsity
Compute efficiency (GOPS/mm^2)	36.01 (1x1b MAC) 4.50 (1x8b MAC)	100.80 (1x1b MAC) 12.60 (1x8b MAC)
Energy efficiency (TOPS/W)	8.48 (1x1b MAC) 1.06 (1x8b MAC)	56.10 (1x1b MAC) 7.01 (1x8b MAC)

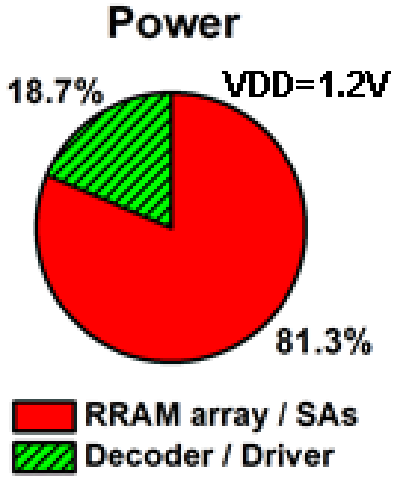
Performance on VGG-8	Sparsity Control Enabled	Sparsity Control Disabled
CIFAR-10 accuracy	90.4%	91.9%
Compute efficiency (GOPS/mm^2)	83.50 (1x1b MAC)	36.01 (1x1b MAC)
Energy efficiency (TOPS/W)	36.39 (1x1b MAC)	8.48 (1x1b MAC)

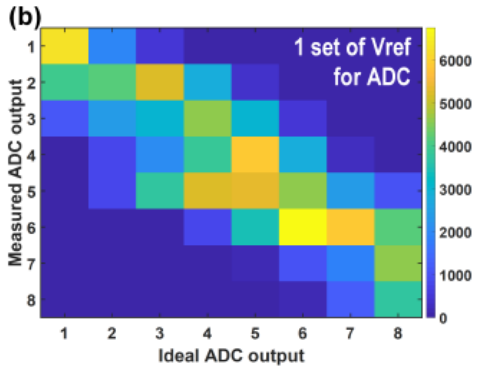
2nd-gen RRAM CIM chip taped-out (May 2021)

Challenges for RRAM-CIM Chip Design

- Low R_{on} → Large column current → Analog MUX at end of the column size up → Poor area efficiency
- High Vw → Large transistor needed for 1T1R cell → Bit cell size may be >30F^2
- High Vw → Significant area on the level shifters
- ADC area/power bottleneck → Multiple columns share one ADC → Time multiplexing required → Reduced throughput
- Process variation → ADC offset → Inaccurate partial sum computation → Inference accuracy degradation







Process variation

S. Yu, et al. CICC 2020

Summary and Outlook

- NVM (RRAM, PCM, and FeFET) can be tuned to multilevel (possibly by iterative write-verify), and the read-intensive inference is most suitable application with advantages over SRAM (e.g. low leakage and non-volatility) for edge intelligence.
- FeFET is the most promising candidate with features like improved on-state resistance (>100kΩ) with gate biasing, and low write energy (~fJ/bit) due to field-driven switching, fast read/write speed (~10ns), and 2-5 bit/cell potential. Need to build array-level test vehicles (e.g. GF' 28nm) for characterizing statistics.
- NVM based inference engine still faces challenges such as high write voltage and low on-state resistance, ADC overhead, intermediate state stability, process variation caused inference accuracy degradation, etc.
- DNN+NeuroSim is an integrated framework for benchmarking different CIM technologies that is open source to the research community.

Acknowledgement

Students/Postdoc: Xiaochen Peng, Yandong Luo, Wonbo Shim, Panni Wang, Hongwu Jiang, Xiaoyu Sun, etc.