Arizona State University

Binary Neural Network and Its Implementation with 16 Mb RRAM Macro Chip

Shimeng Yu
Assistant Professor of Electrical Engineering and Computer Engineering
shimengy@asu.edu

http://faculty.engineering.asu.edu/shimengyu/

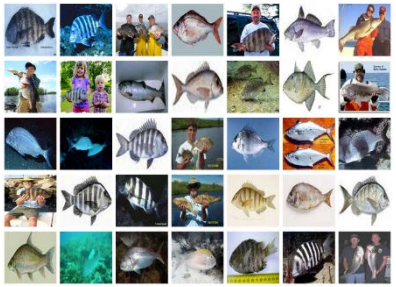
Outline

- **Challenges of Analog Synapses and Why We Need Binarize the Neural Network**
- **Binary Neural Network and Its Implementations on Tsinghua's 16 Mb RRAM Macro Chip**
- **Benchmark of Binary and Analog Synapses**
- **Summary**

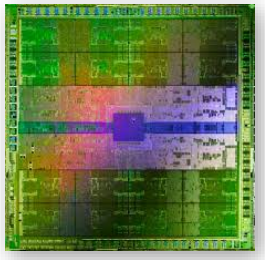
Demands for Neuromorphic Hardware

- Deep learning in Cloud: huge training **"labeled"** dataset, **high precision** training, power-hungry, etc.

Google Cat:
16,000 CPU cores

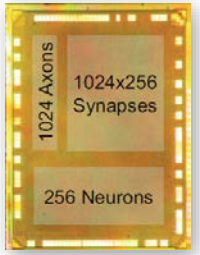


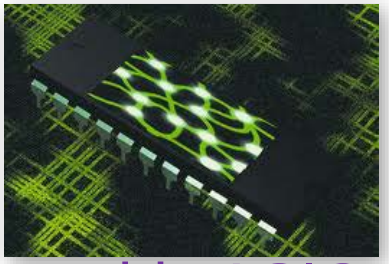
MS Residual-CNN: 8 GPUs



GPU
FPGA

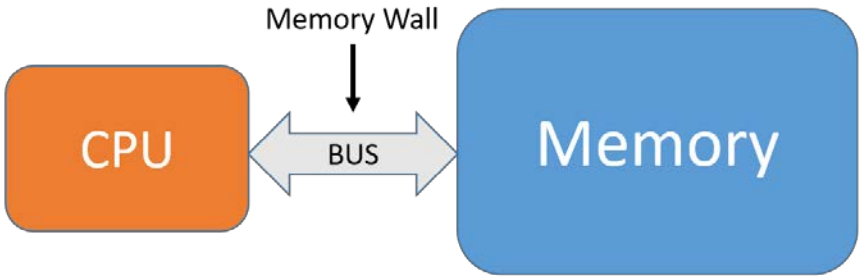
- Edge (IoT) computing needs novel hardware / algorithms
  - **Local** to the sensor, **real-time** inference, small area and low-power
  - Adaptive **on-line** learning with continuous (possibly unlabeled) data



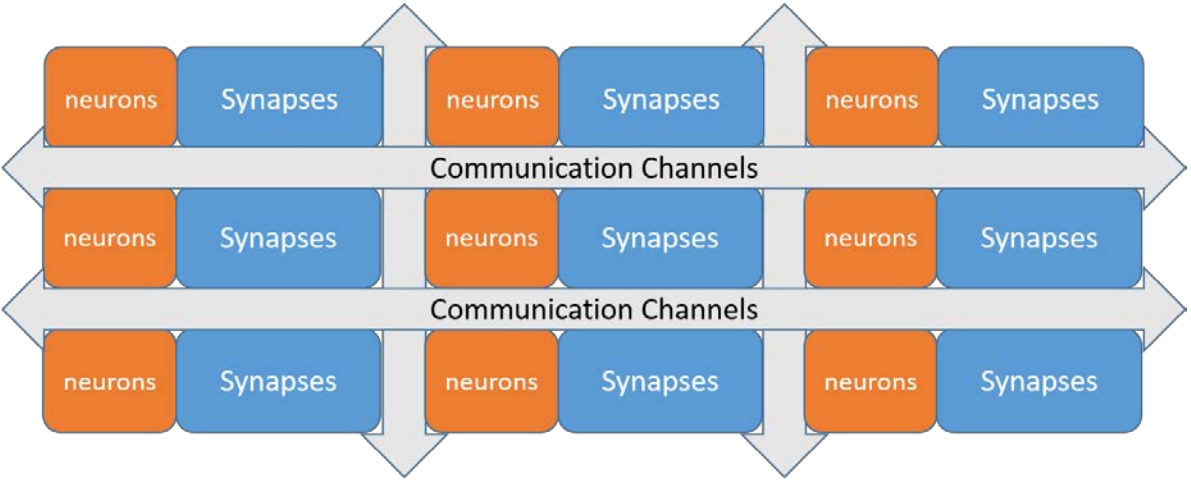


Neuromorphic ASIC

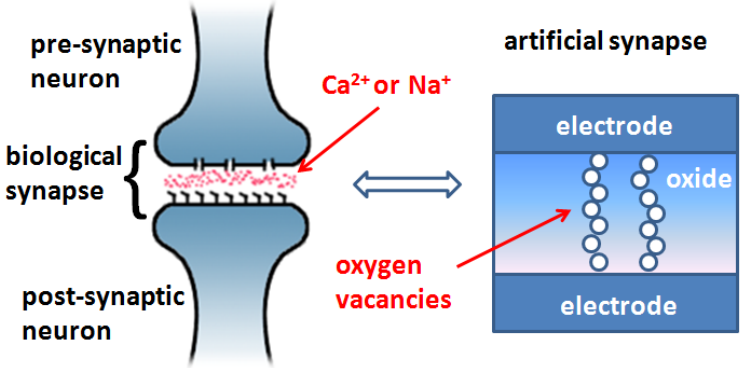
A Shift in Computing Paradigm towards Neuro-inspired



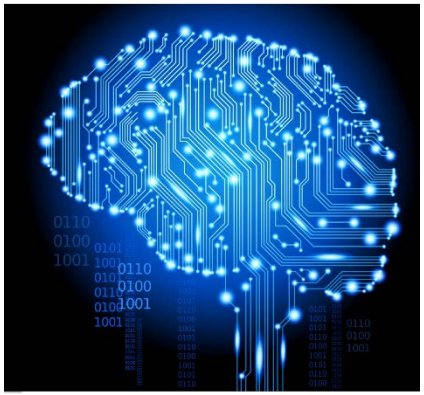
(a) von-Neumann architecture



(b) Neuro-inspired Architecture



Resistive synaptic device



Long-term vision: Brain-like computer

Current Status of eNVM based Neuromorphic Research

- **Mostly focused on device-level engineering…**

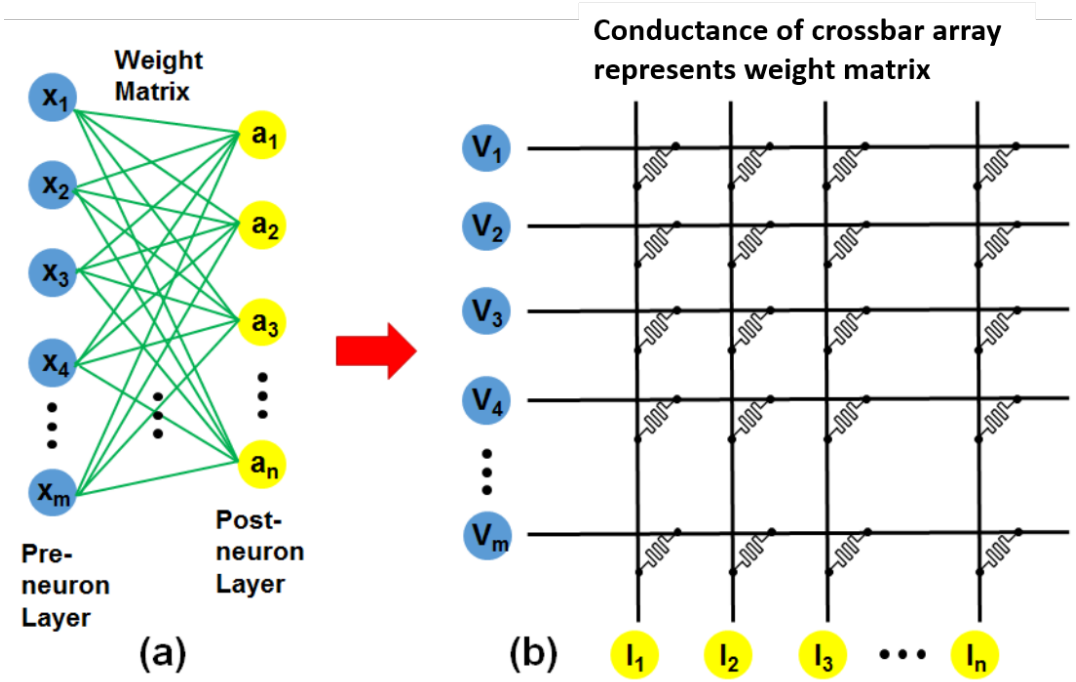
Performance metrics	Desired Targets
Device dimension	< 10 nm
Multilevel states number	>100* with a linear symmetric update
Energy consumption	<10 fJ/programming pulse
Dynamic range	>100*
Retention	>10 years*
Endurance	>10^9 updates*

Note: * these numbers are application-dependent

- **A few array-level demo with simple pattern classification, such as:**
  - UCSB's 12*12 crossbar array with memristors (Nature 2015)
  - IBM's 256*256 1T1R array with PCM (IEDM 2015)
  - ASU's 12*12 crossbar array with multilevel RRAM (EDL 2016)
  - ASU-Tsinghua's 400*400 1T1R array with binary RRAM (IEDM 2016)

Cross-point Architecture for Accelerating Weighted Sum and Weight Update

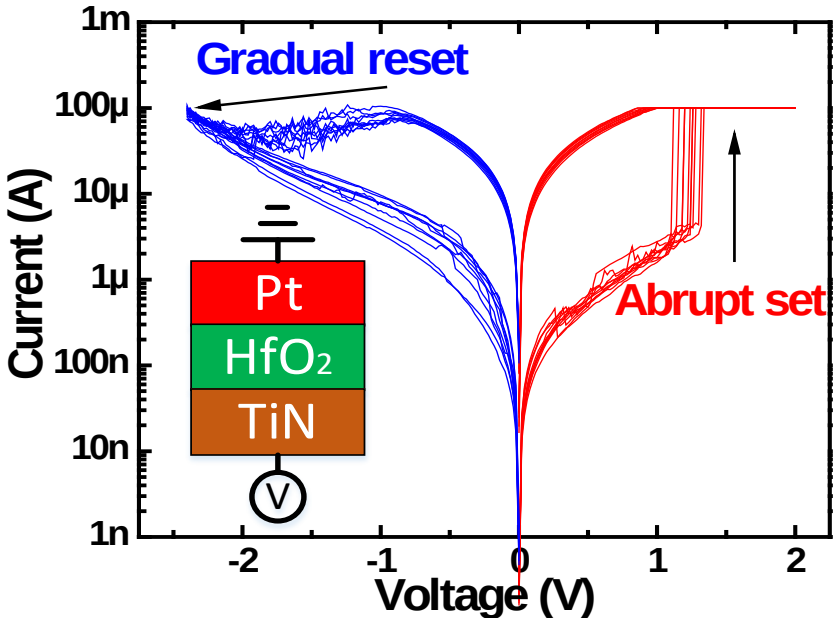
- **Weighted sum (inference): all cells are activated in parallel, summing up column current→perform vector-matrix multiplication**
- **Weight update (training): cell's conductance could be updated by applying programming voltage from row/column at the same time.**

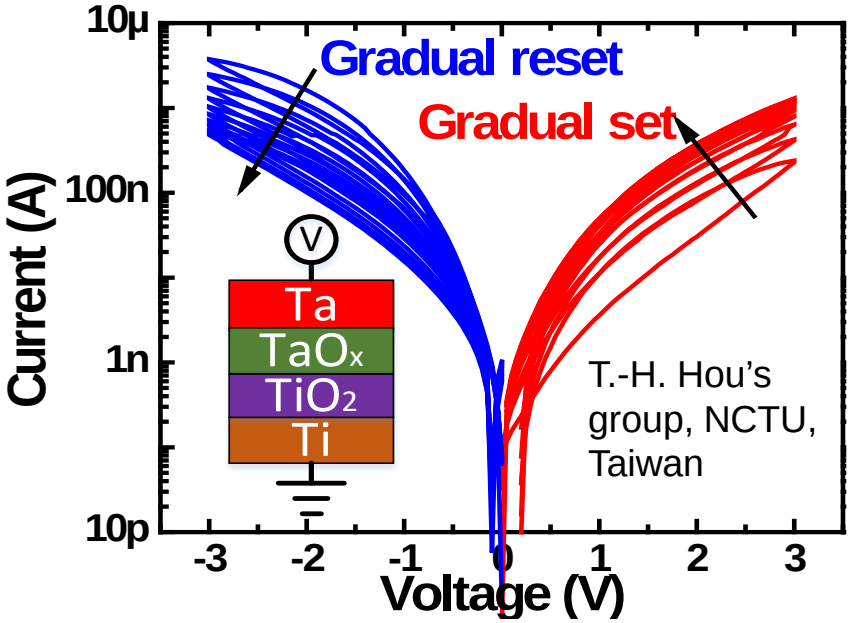


Task	Operations
$\boldsymbol{W} \cdot \boldsymbol{X}$	$I_i = \sum_j G_{ij} \cdot V_j$
$\boldsymbol{W}$ update	$\Delta G_{ij} = \eta \cdot V_i \cdot V_j$

(analog computation inside the array, may need ADC at edge of array)

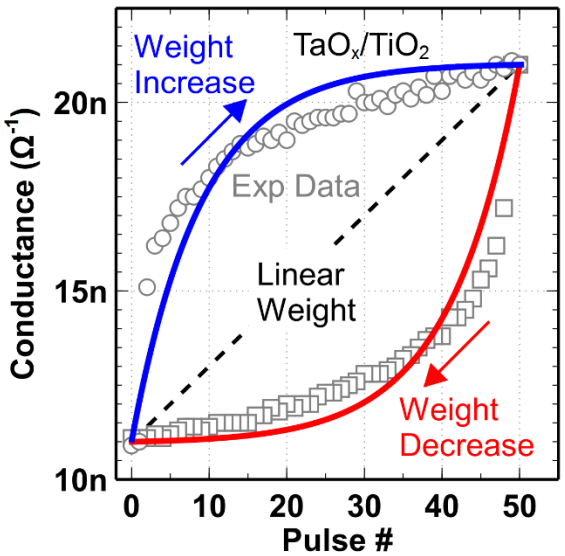
Binary RRAM and Analog RRAM Synaptic Devices

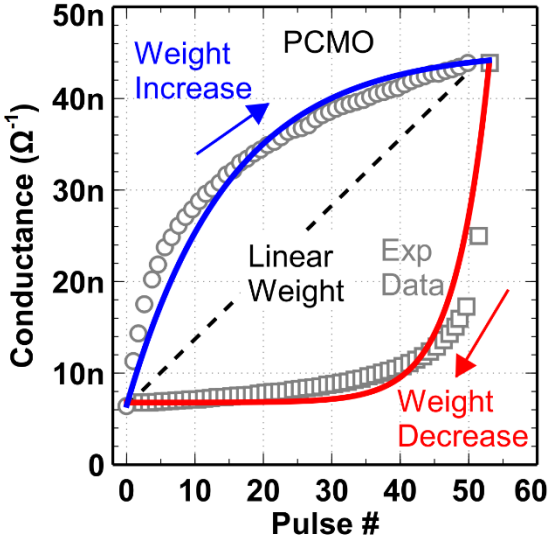


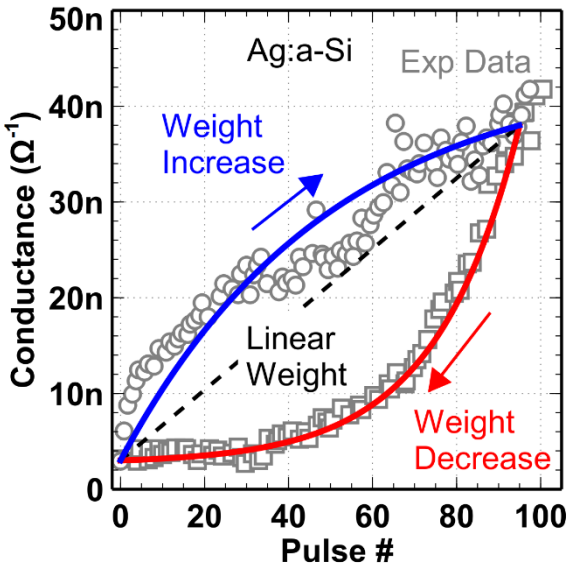


- **Binary Synapses:** Conventional *filamentary switching* RRAM with <u>abrupt set and gradual reset</u>, multilevel states achievable in the reset, could be used for offline training.
- **Analog Synapses:** Special *interfacial switching* RRAM with <u>both smooth set and reset</u>, attractive for online training.

Realistic Analog Device's Weight Update Behaviors



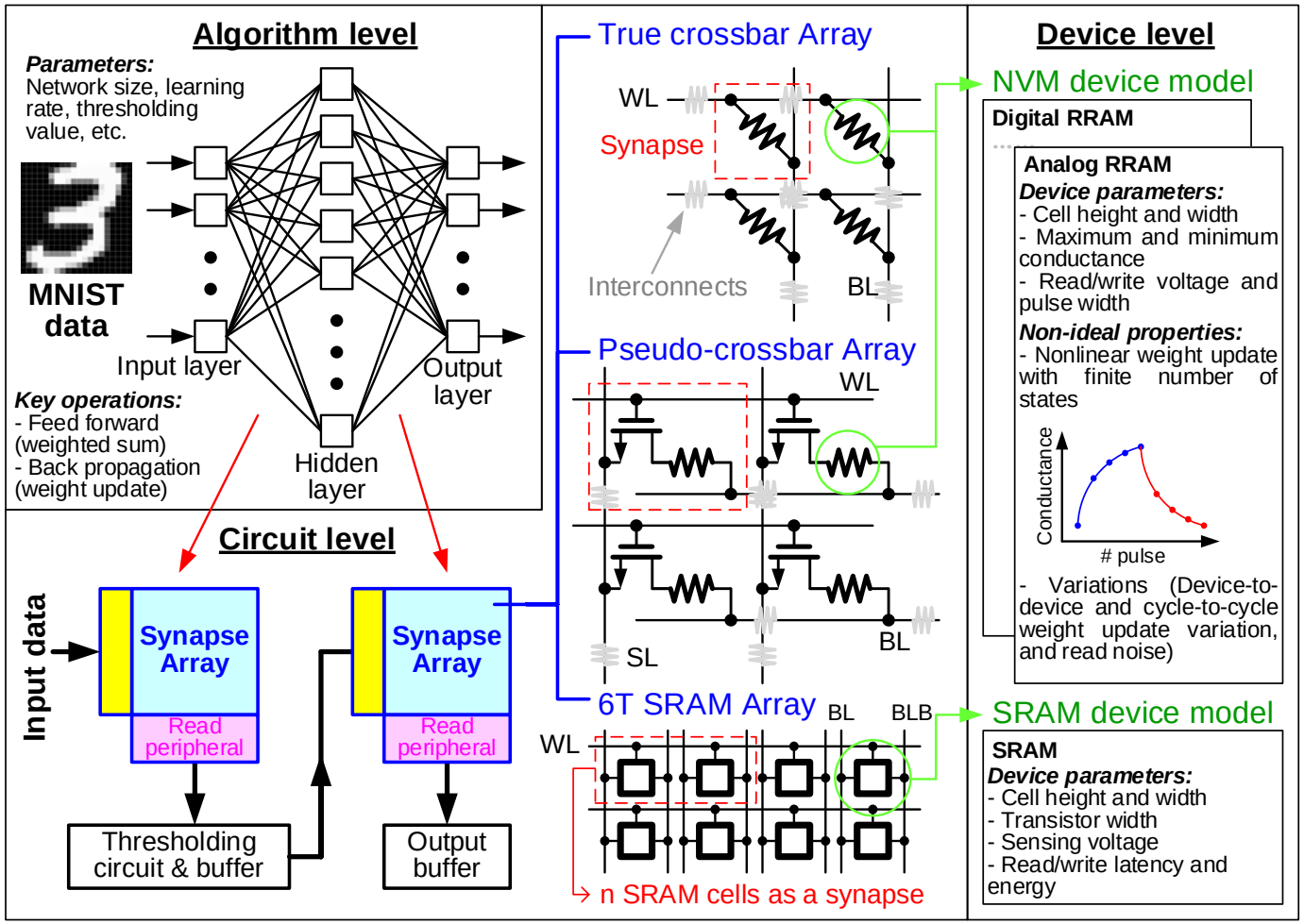




- Nonlinearity in weight update
- Device variations
- Non-zero off-state conductance

How would these non-ideal effects impact learning accuracy?

NeuroSim: A Simulator from Device to Algorithm



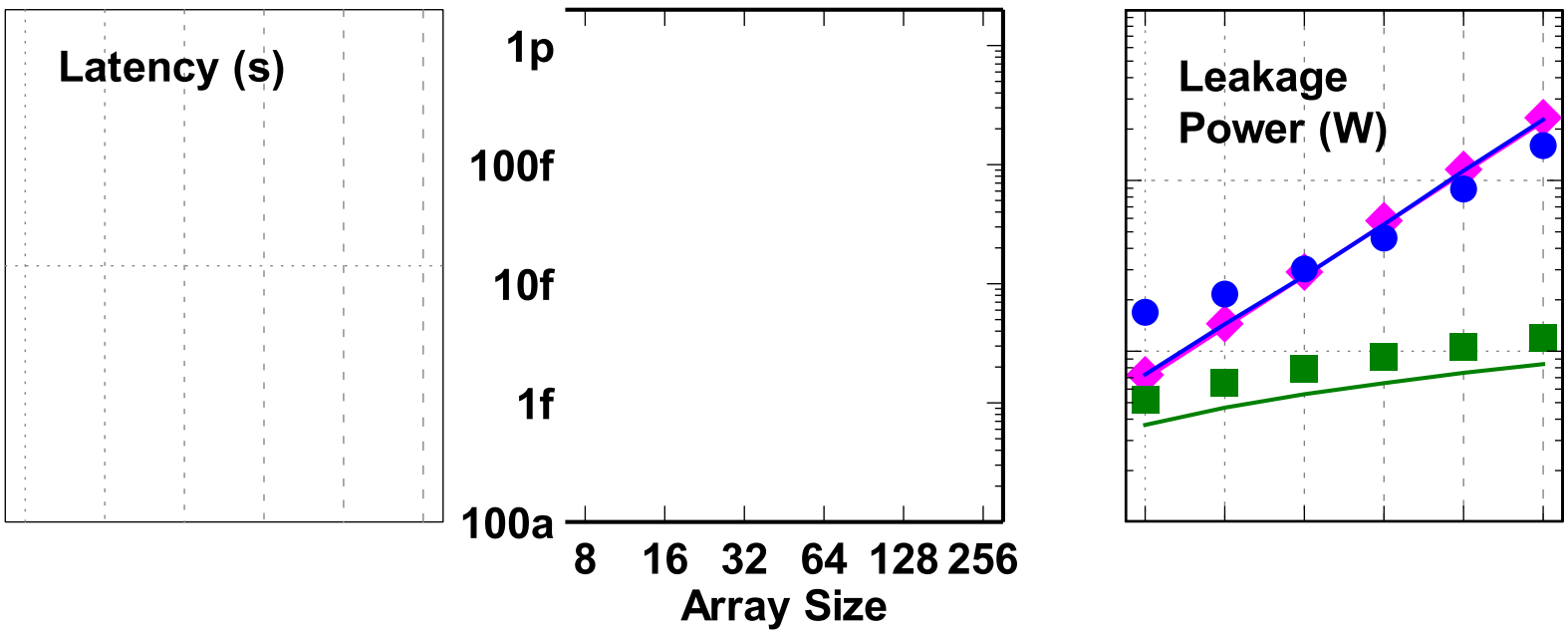
Input:

- Network structure,
- Training/testing traces
- Array type and technology node
- Device type and non-ideal factors

Output:

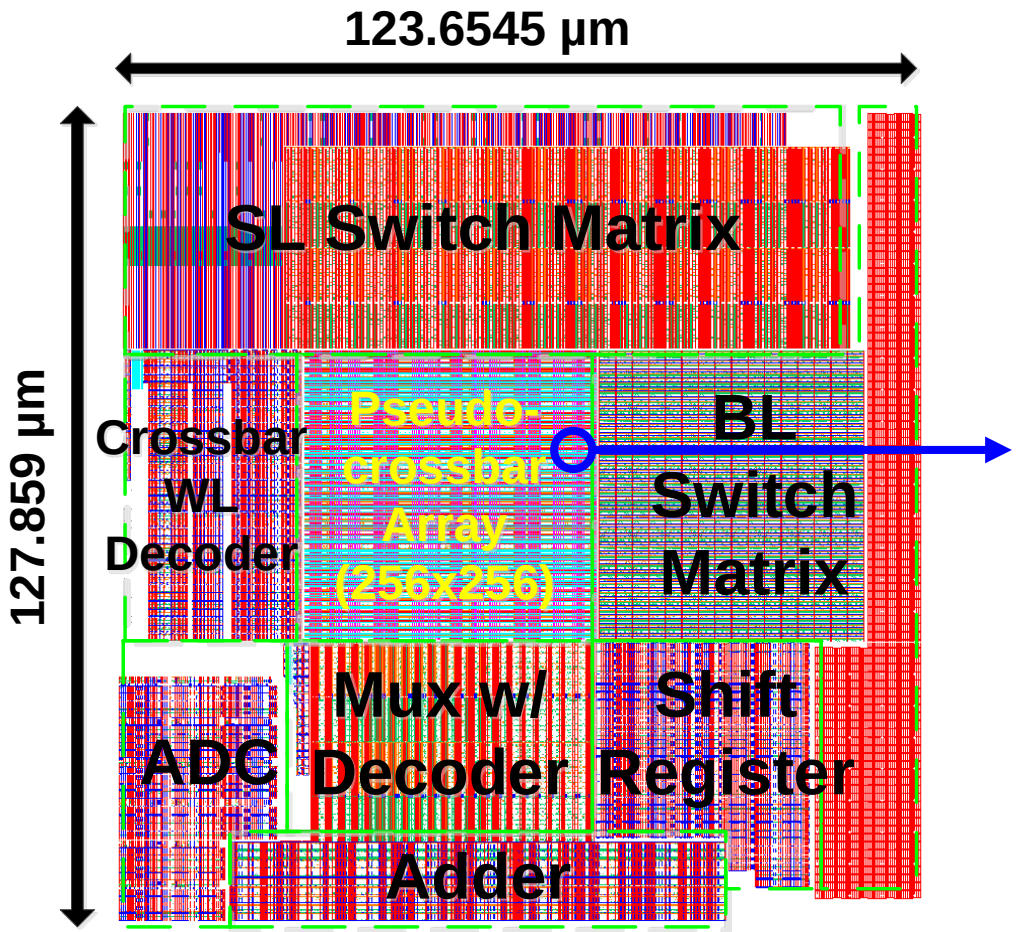
- Area,
- Latency,
- Energy,
- Accuracy

Model Calibration (Latency, Energy, Leakage)

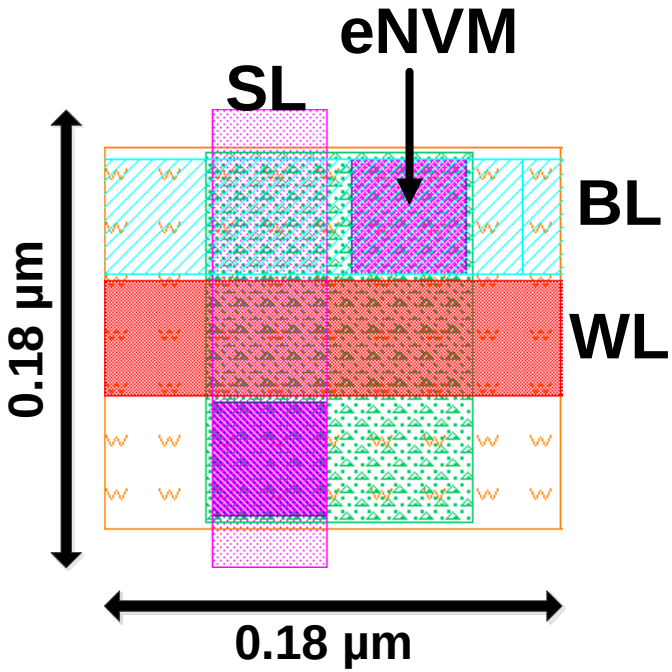


Benchmark at 45 nm with PTM model

Model Calibration (Area)



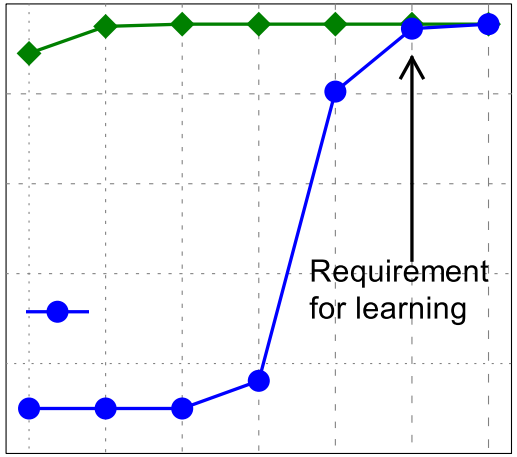
Layout using 45 nm NanGate PDK

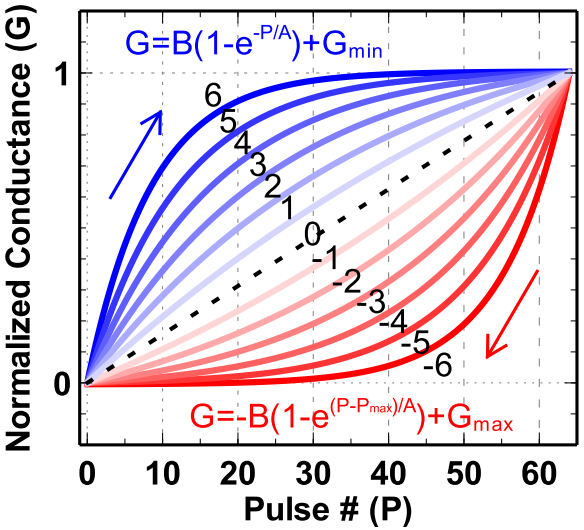


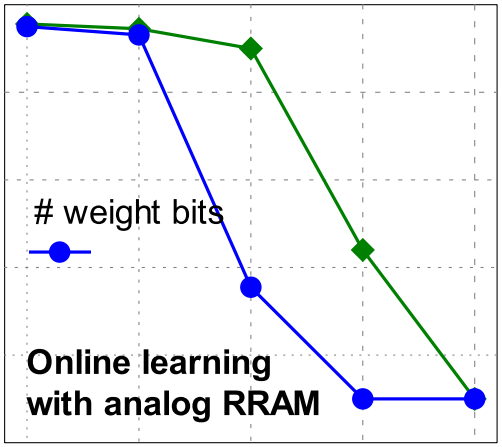
Layout Area: 1.5810E+04 um^2

Model Area: 1.5454E+04 um^2

Impact of Weight Precision and Weight Update Nonlinearity in Analog Synapses







- A multilayer perceptron (MLP) 400-200-10 network is used for benchmarking.
- At least 6-bit is required for MNIST dataset online learning, while 1 or 2-bit may work for offline classification.
- Nonlinearity significantly degrades accuracy for online learning if using analog synapses.

Benchmark of Reported Analog Resistive Synapses

	Reported analog eNVMs for learning				Desired analog eNVMs for learning	
eNVM type	PCMO	Ag:a-Si	TaO_x/TiO_2	AlO_x/HfO_2	Targeted eNVM	Ideal eNVM
# of bits	5	6	6	5	6	6
Nonlinearity (weight increase/decrease)	3.25/5.82	1.13/2.65	1.13/0.72	3/1	1/1	0/0
R_{ON}	23 MΩ	26 MΩ	5 MΩ	16.9 kΩ	200 kΩ	200 kΩ
ON/OFF ratio	6.84	12.5	2	4.43	50	50
Weight update cycle-to-cycle variation (σ)	<1%	3.5%	<1%	5%	2%	0%
Accuracy for online learning	10%	~75%	~10%	~10%	~90%	~94.8%
Accuracy for offline classification	~13%	~51%	~10%	~10%	~94.5%	~94.5%

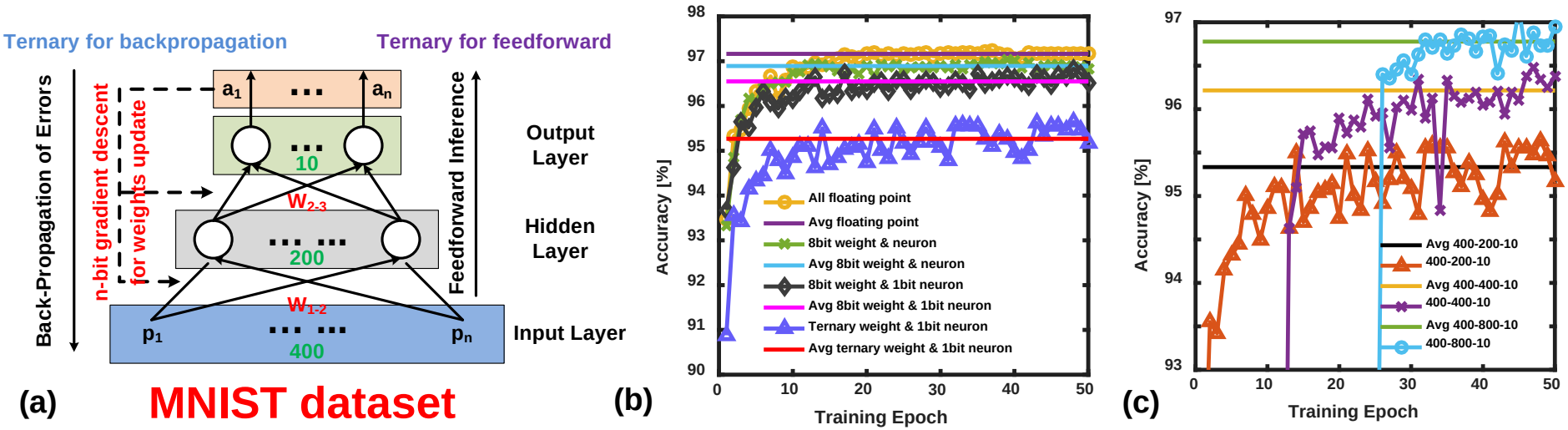
Green: good attributes, Red: major cause of learning failure

Outline

- **Challenges of Analog Synapses and Why We Need Binarize the Neural Network**
- **Binary Neural Network and Its Implementations on Tsinghua's 16 Mb RRAM Macro Chip**
- **Benchmark of Binary and Analog Synapses**
- **Summary**

Binary Neural Network (BNN)

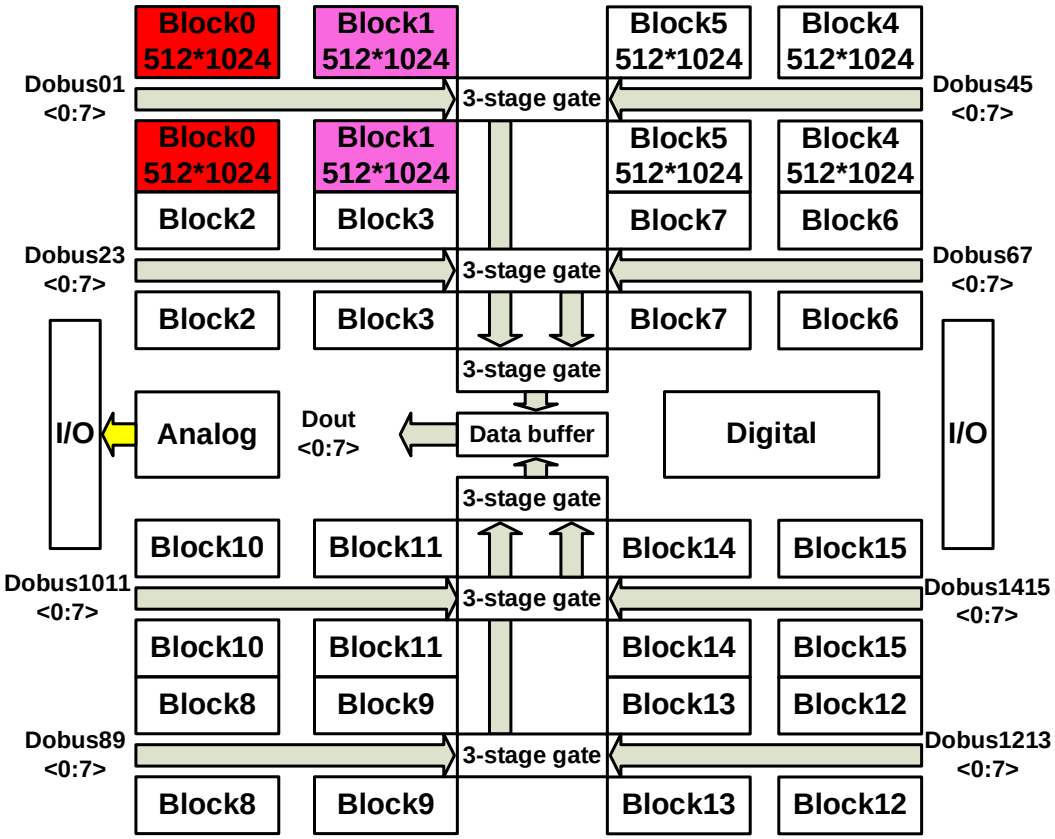
- Precision Reduction to Ternary Weight (+1,0,-1) and Binary Neuron for Propagation
- Higher precision (e.g. 8 bit) is kept for weight update only (because ΔW is small)

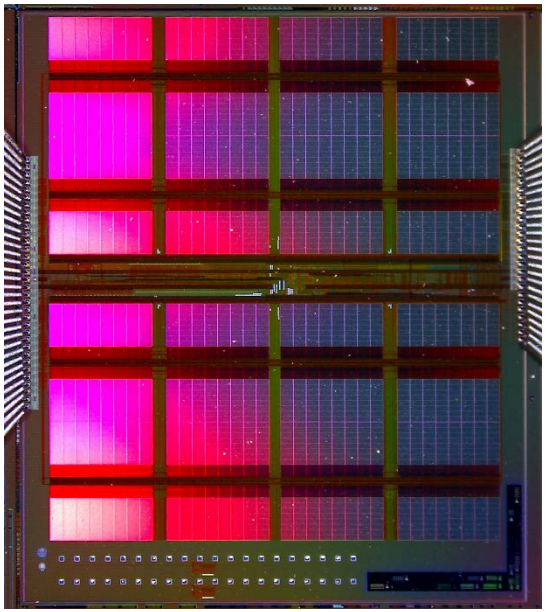


- Followed the recent trends in machine/deep learning, e.g. *BinaryNet* and *XNOR-Net*

S. Yu, et al. IEDM 2016

16 Mb Macro Chip (Tsinghua)

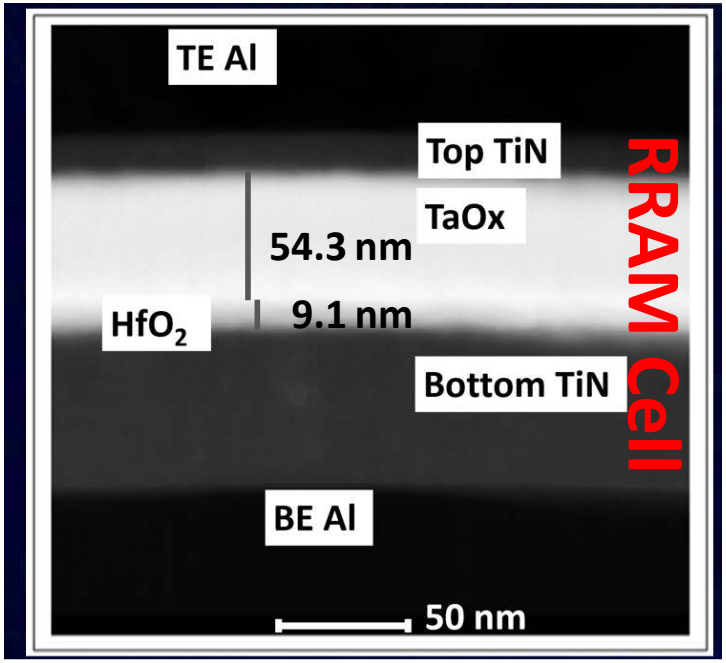


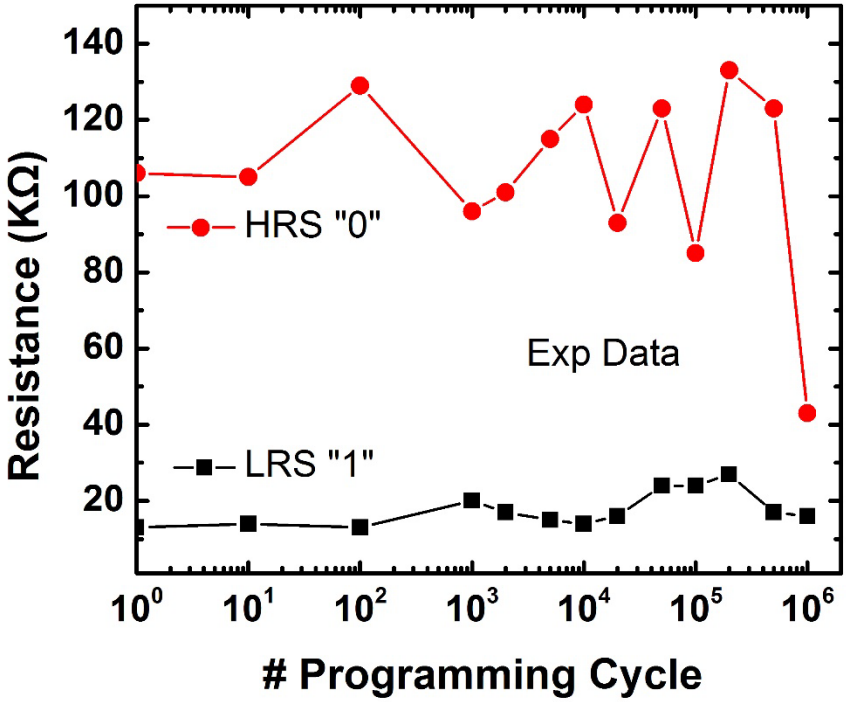


Capacity	16 Mb
Tech Node	130 nm
$V_{DD\_Digital}$	1.8 V
V_{DD_Analog}	5 V
V_{WL_SET}	2-5 V/ 50 ns
V_{BL_SET}	2-3 V/ 50 ns
V_{WL_RESET}	3.5-5 V/ 50 ns
V_{SL_RESET}	2-3 V/ 50 ns
I/O Width	8

Chip designed and fabricated by Huaqiang Wu's group in Tsinghua University

RRAM Stack and Endurance of RRAM



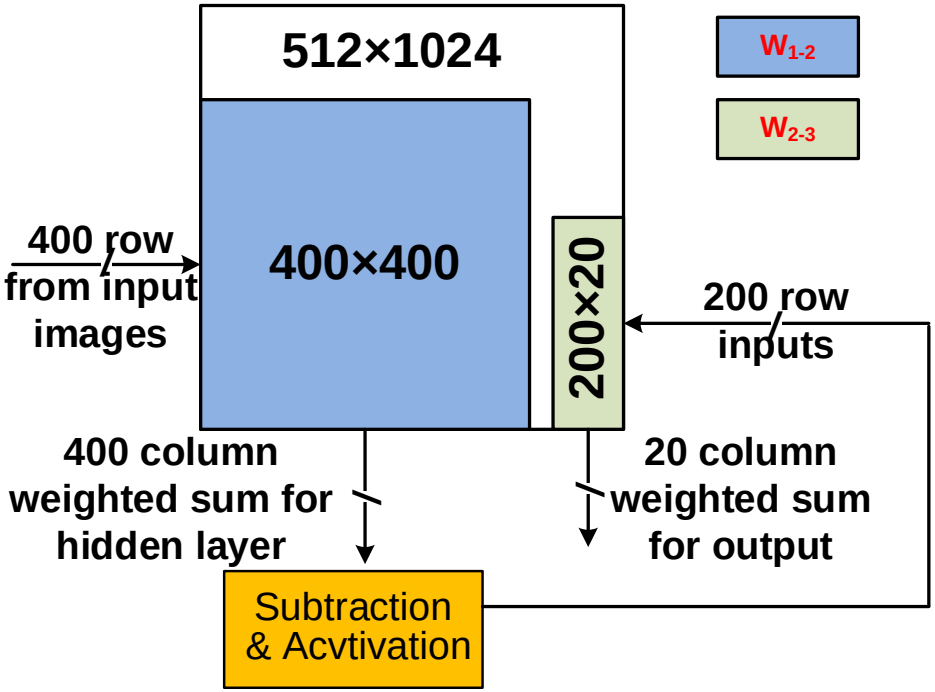


HfOx based RRAM integrated between M4 and M5 on top of CMOS

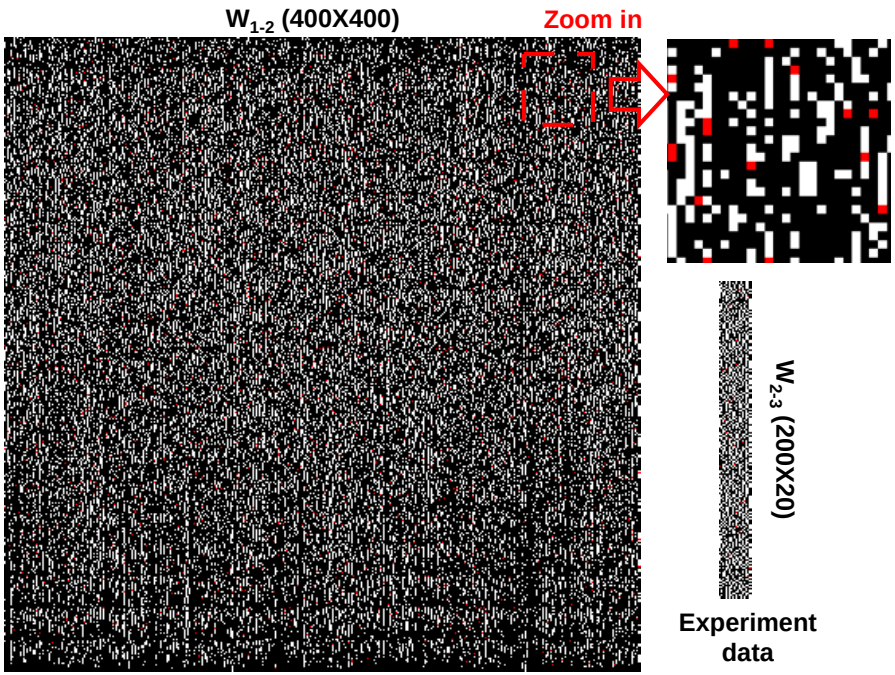
Measured endurance ~1E6 cycles

Courtesy of Huaqiang Wu (Tsinghua University)

Implementation of BNN on 16 Mb RRAM Chip for Offline Classification

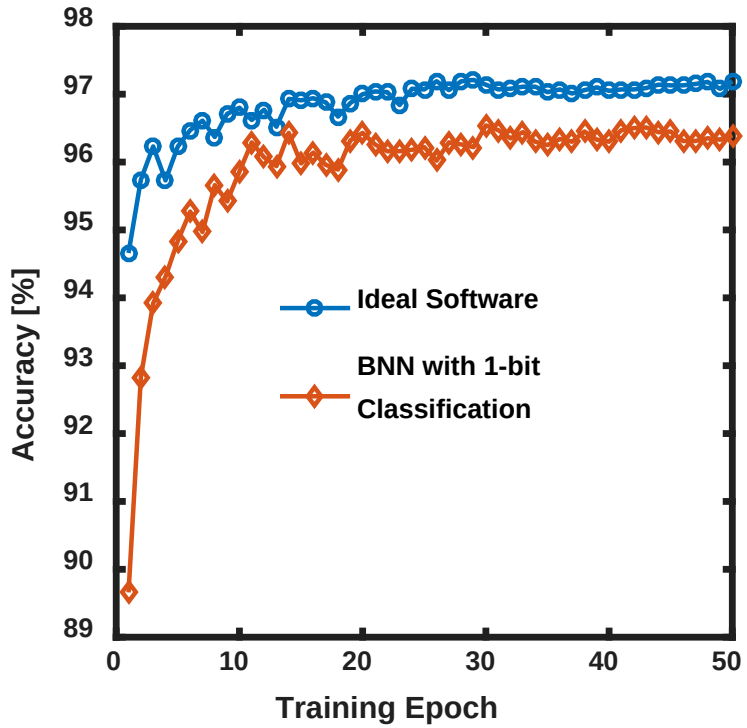


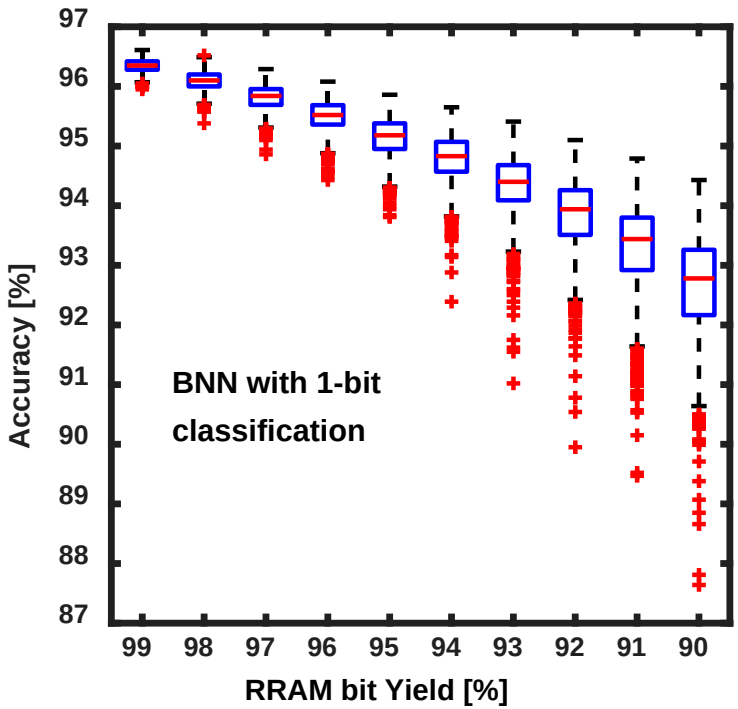
Network topology 400-200-10



Programmed weight matrix pattern on 1 block of 16 Mb chip
Error (in red) occurs, bit yield ~99%

Impact of RRAM Finite Bit Yield for Classification

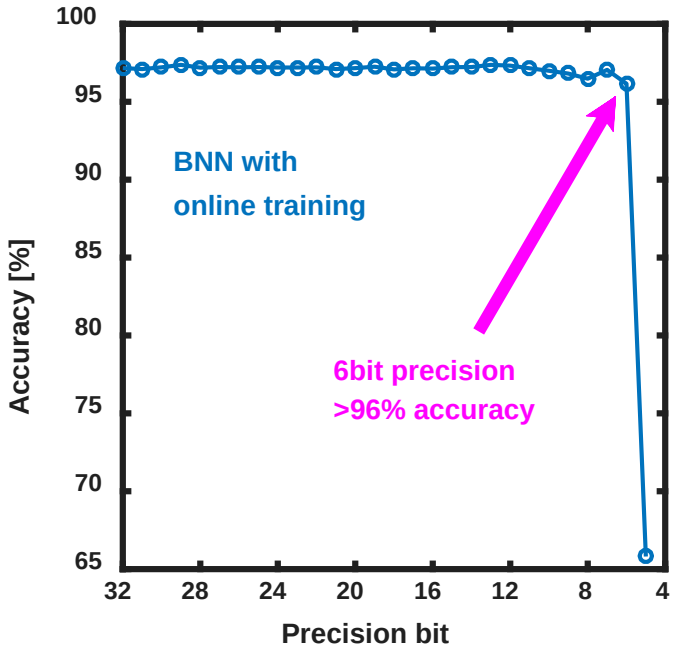


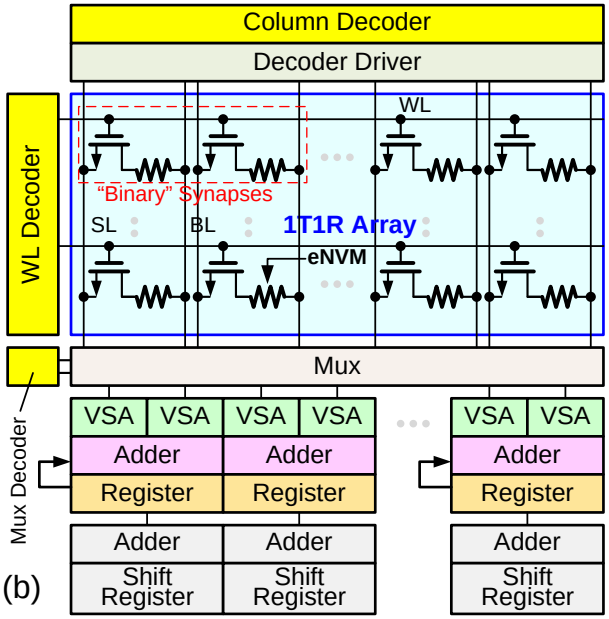


- The software baseline with high precision classification has accuracy ~97%.
- BNN with 1-bit classification (with sign) has accuracy ~96.3%
- For MNIST dataset, 99% bit yield is sufficient to maintain ~96.3%

Precision Reduction for Training

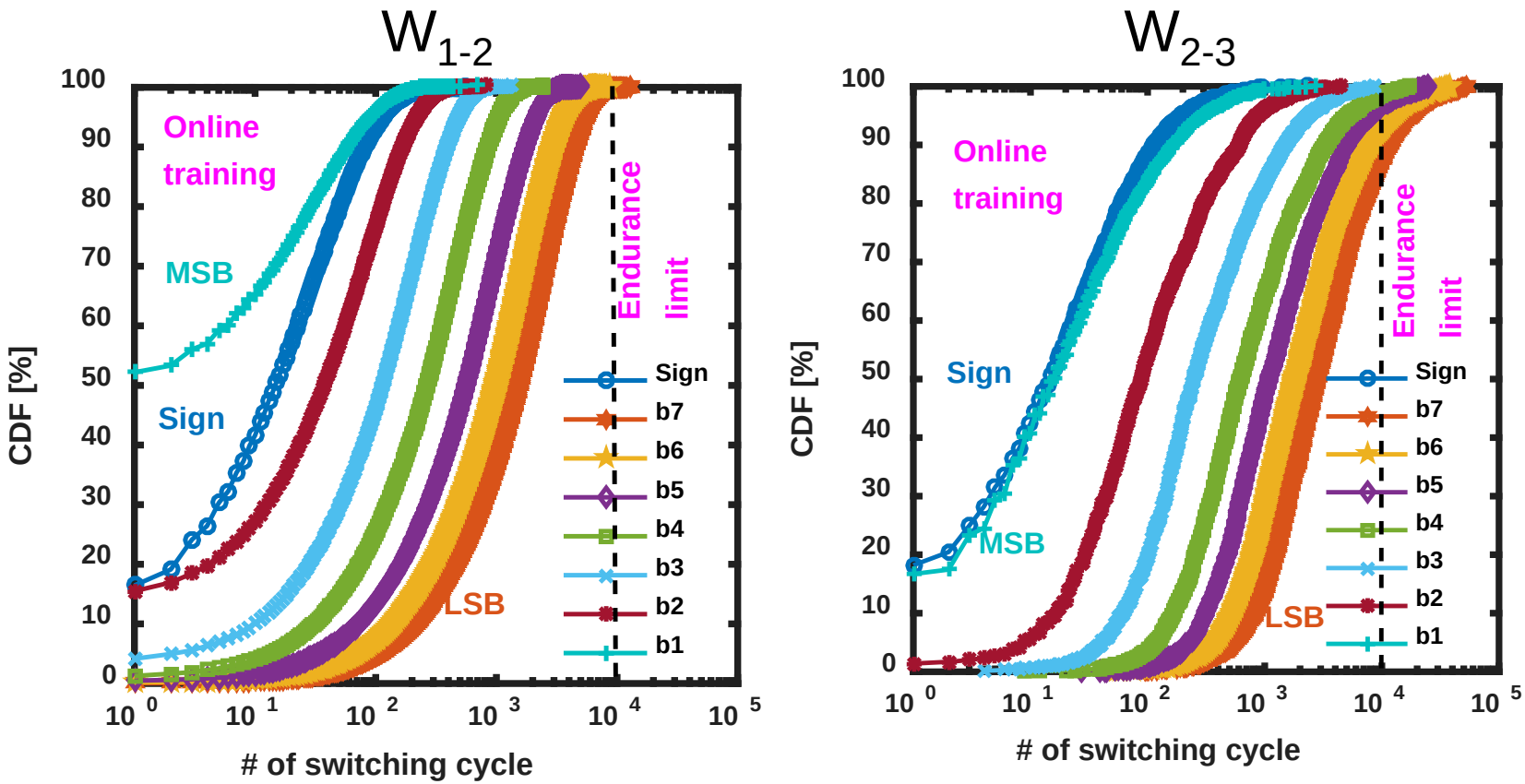
Online training needs higher precision than offline classification, because the small error accumulation is needed in backpropagation





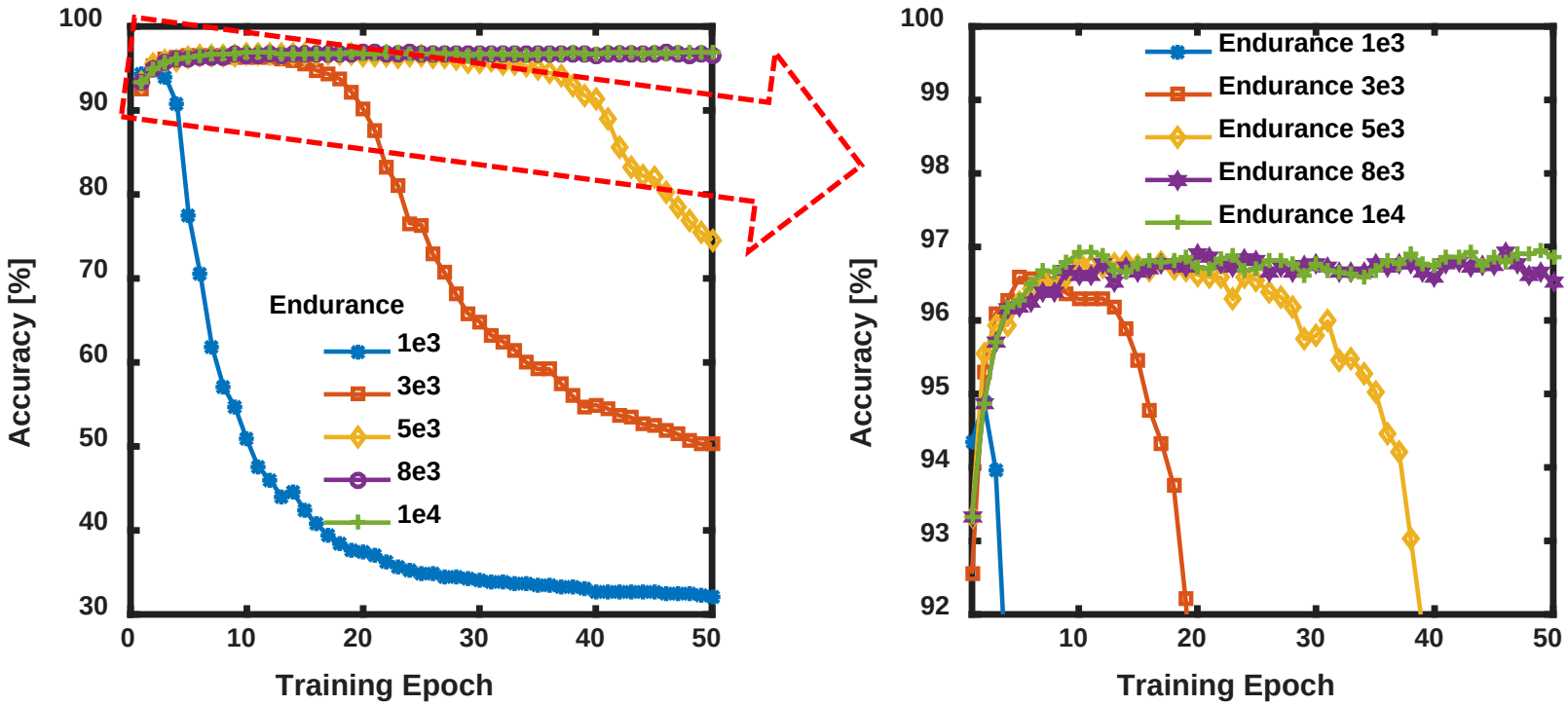
6-bit is needed for MNIST dataset, thus 6 binary RRAM cells are grouped for implementing one synapse

Distribution of RRAM Updates During Training



- Most cells update less than endurance limit (10^4 cycles)
- LSB updates more than MSB, and $W_{2\text{-}3}$ updates more than $W_{1\text{-}2}$

Impact of RRAM Finite Endurance on Training



Lower endurance results in lower peak of accuracy. With 10^4 cycles, ~96.9% accuracy is achievable for online training

Outline

- **Challenges of Analog Synapses and Why We Need Binarize the Neural Network**
- **Binary Neural Network and Its Implementations on Tsinghua's 16 Mb RRAM Macro Chip**
- **Benchmark of Binary and Analog Synapses**
- **Summary**

NeuroSim Simulation Set-up for Analog and Binary Synapses

	"Analog" synapse	"Binary" synapse
# bits	6	6
Nonlinearity (weight increase/weight decrease)	0.72/0.72	--
R_{ON}	200kΩ	200kΩ
ON/OFF ratio	50	50
Read voltage	0.5 V	0.5 V
Write voltage	2 V (for both weight increase and decrease)	2 V
Write pulse width	100 ns per pulse	100 ns
Resistance of access transistor in 1T1R	10kΩ	10kΩ
Read noise	2.89%	--
Array type	Pseudo-crossbar	Traditional 1T1R
Array size	400x100 and 100x10	400x600 and 100x60
Tech node	14 nm	14 nm
Wire width	40 nm	40 nm

Benchmark Results of Analog and Binary Synapses

	"Analog" synapse	"Binary" synapse
Accuracy	82.17%	94.03%
Area	1560.8 μm^2	2678.2 μm^2
Total feed forward latency	1.1044e-01 s	2.7063e+00 s
Total weight update Latency	1.7640e+05 s	3.2283e+03 s
Total feed forward energy	4.4835e-04 J	2.3709e-03 J
Total weight update energy	2.7115e+00 J	8.0447e+00 J
Leakage	26.631 µW	15.397 µW

Binary synapses could be a near-term solution, while a "perfect" analog synapses could bring in many benefits in the long run

Summary

- **Today's RRAM technology (even binary) can support <u>offline classification</u> with low-power, fast and accurate recognition.**
- **For <u>online training</u>, "analog" synapses with continuous weights need to overcome grand challenges such as nonlinear weight update, and slow programming speed (as multiple pulses are needed to tune the weights).**
- **Binarizing neural network with low-precision weights, allows today's binary RRAM for online training with high accuracy, which also shows a good resilience to limited yield and endurance, as shown in our demonstration of 16 Mb RRAM chip.**
- **Trade-offs exist between binary and analog synapse implementations: binary synapses are good for high accuracy and fast training speed, but with overhead in the chip area and dynamic energy.**

Acknowledgement

- **Students: Pai-Yu Chen, Zhiwei Li**
- **Collaborator: Huaqiang Wu, Tsinghua University**

NSF-CCF-1552687: CAREER: Scaling-up Resistive Synaptic Arrays for Neuro-inspired Computing