Arizona State University

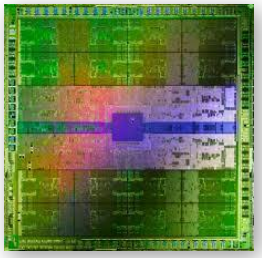
Impact of Non-Ideal Resistive Synaptic Device Behaviors on Neuromorphic System Performances

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

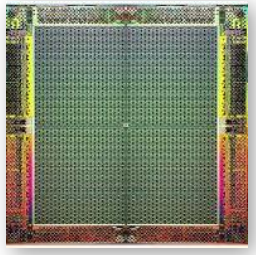
http://faculty.engineering.asu.edu/shimengyu/

Hardware Acceleration Platforms

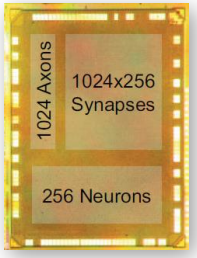
- **10^3 – 10^5** speedup over CPU required to achieve real-time learning, e.g. feature extraction for an HD image at 30 frames/second



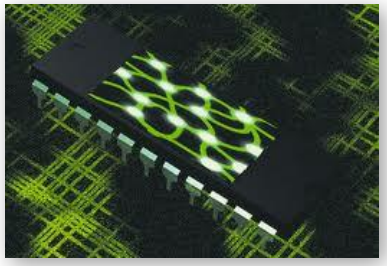
GPU
10 – 30 X



FPGA
10 – 30 X



CMOS ASIC
10^2 – 10^3 X

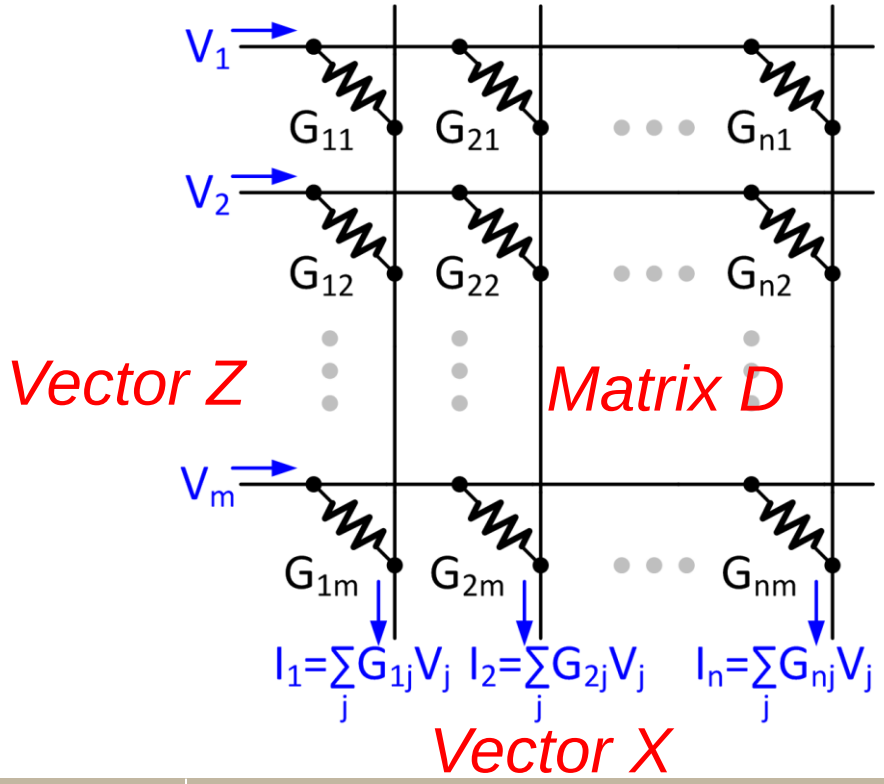


Beyond CMOS
>10^3 X

- **Solution**: beyond CMOS with emerging non-volatile memory
  - Maximizing the **parallel** operation in hardware
  - Our goal: improving computing speed and energy-efficiency. Do not strictly follow the biological principles, such as spike-timing dependent plasticity (STDP)

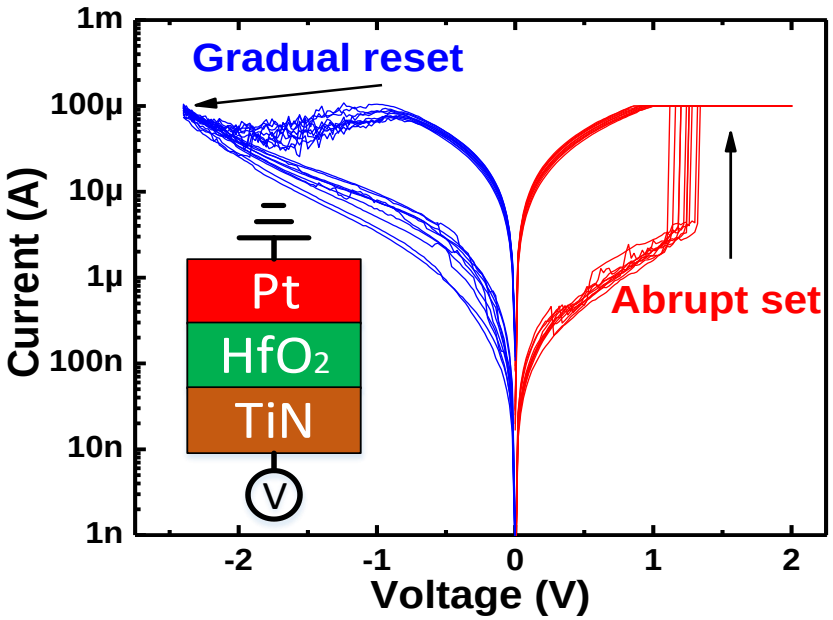
Cross-point Architecture for Accelerating Weighted Sum and Weight Update

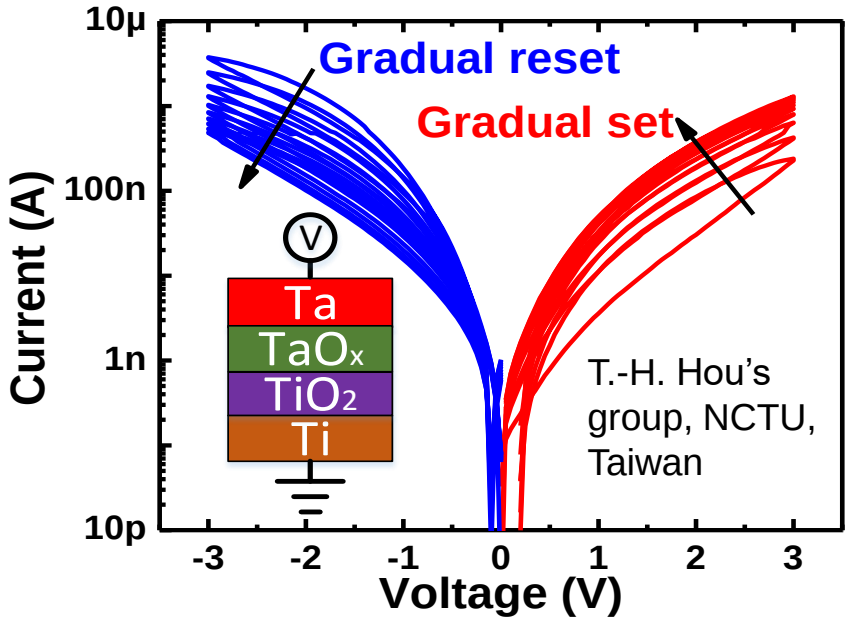
- **Direct mapping weight matrix in neuro-algorithms on crossbar array**
- **All cells are used in parallel, no sneak path problem for read.**
- **Selectors needed for minimizing write power if not fully parallel write**



Task	Operations
$\boldsymbol{D \cdot Z}$	$I_{x,i} = \sum_i G_{ij} \cdot V_{Z,j}$
$\boldsymbol{D^T \cdot X}$	$I_{Z,j} = \sum_j G_{ij} \cdot V_{x,i}$
$\boldsymbol{D}$ **update**	$\Delta G_{ij} = \eta \cdot V_{x,i} \cdot V_{z,j}$

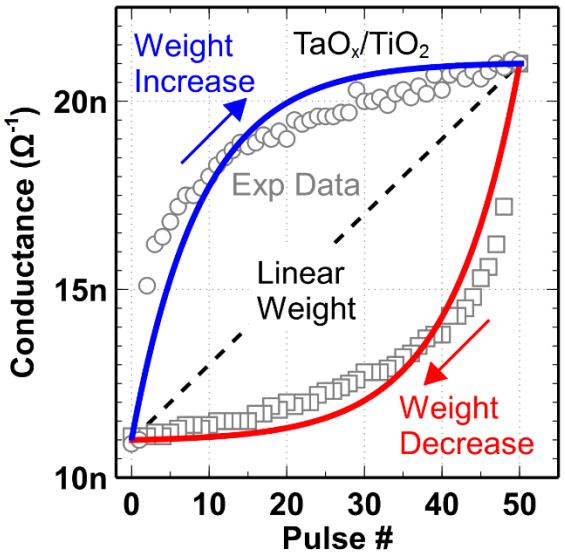
Resistive Devices for Offline and Online Training

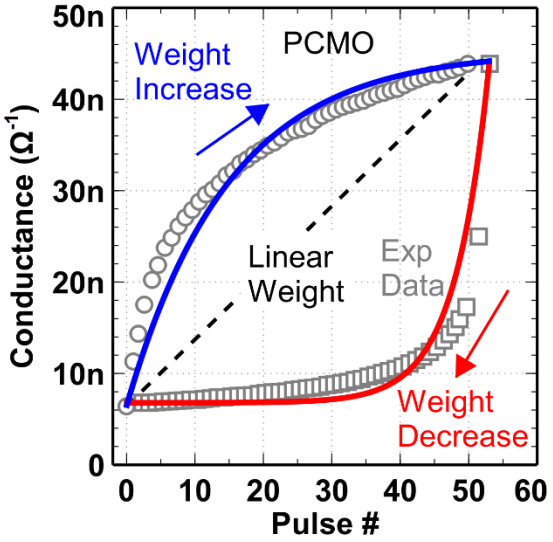


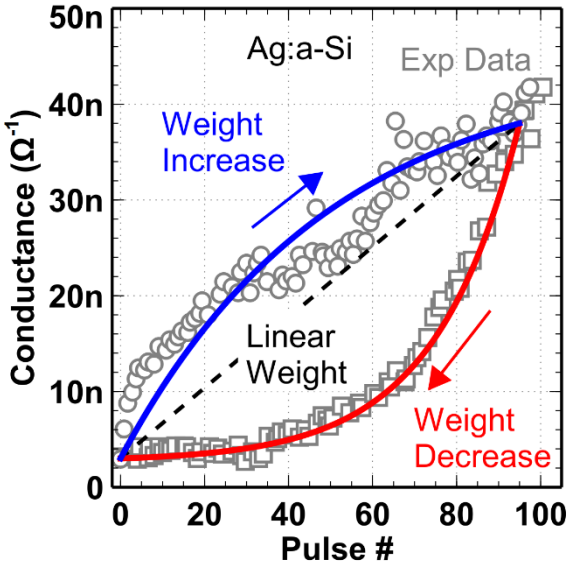


- **Offline training:** weights are pre-defined by software training, just need one-time loading to the array→ Conventional RRAM with gradual reset only is good enough
- **Online training:** weights are updated during run-time→ Special RRAM with both smooth set and reset is needed

Realistic Device's Weight Update Behaviors





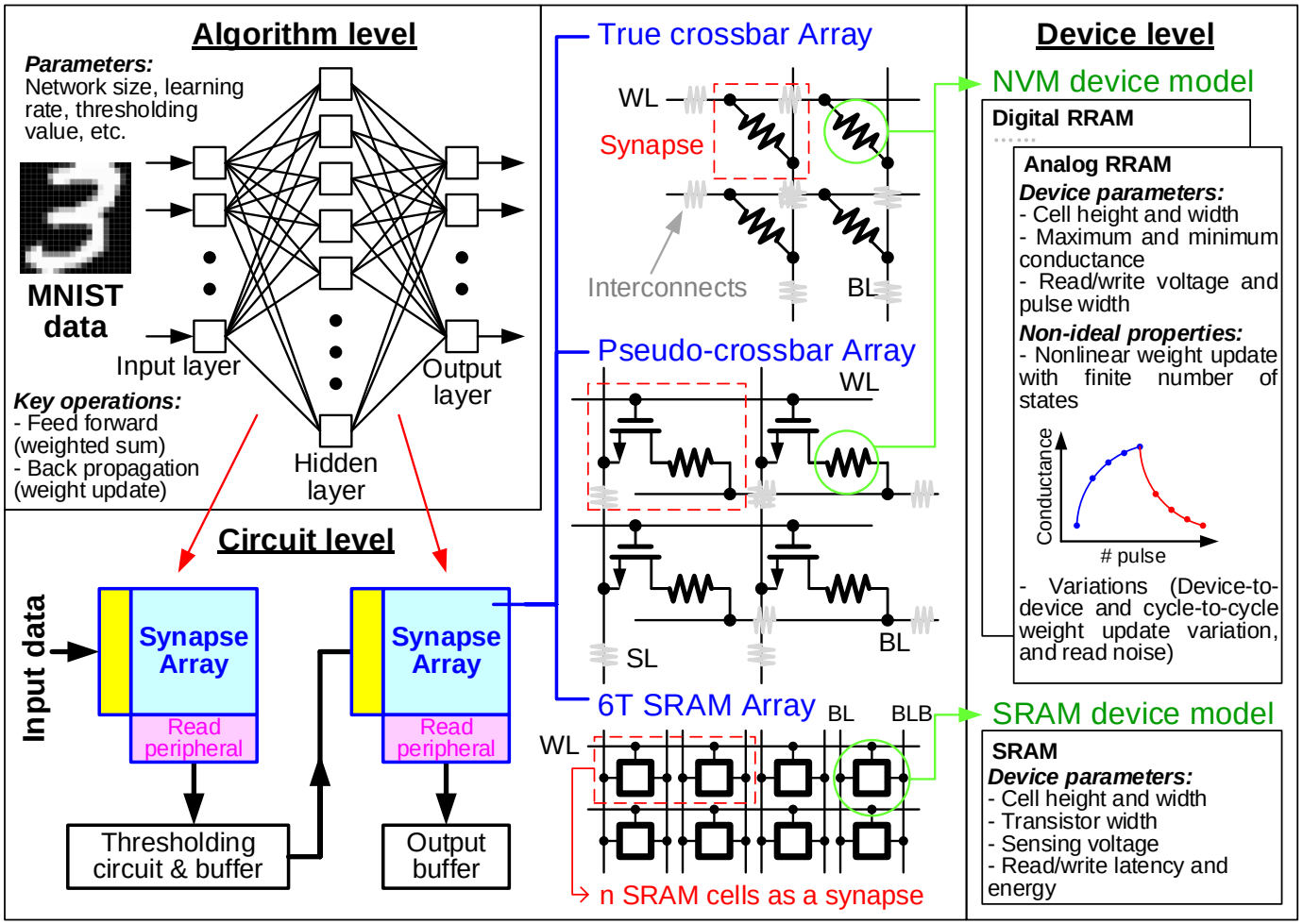


- Nonlinearity in weight update
- Device variations
- Non-zero off-state conductance

How would these non-ideal effects impact learning accuracy?

S. Yu, et al, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: challenges and prospect," IEDM 2015

NeuroSim: A Simulator from Device to Algorithm



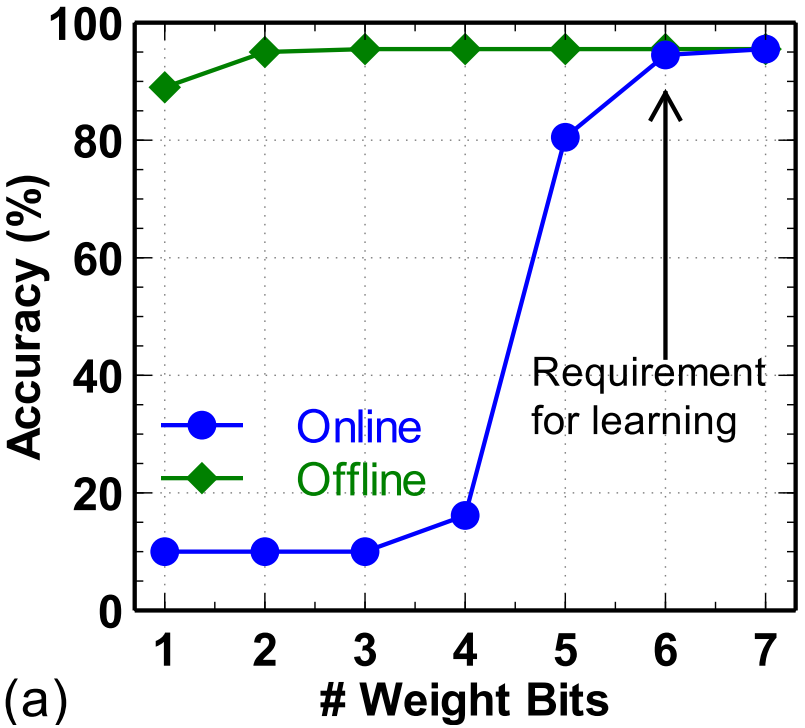
Input:

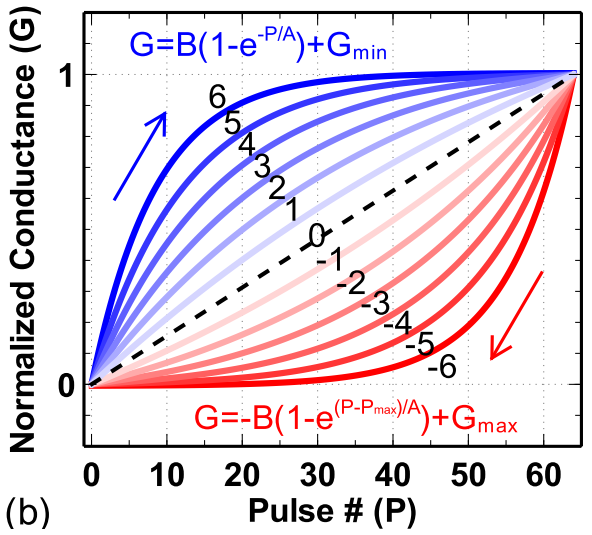
- Network structure,
- Training/testing traces
- Array type and technology node
- Device type and non-ideal factors

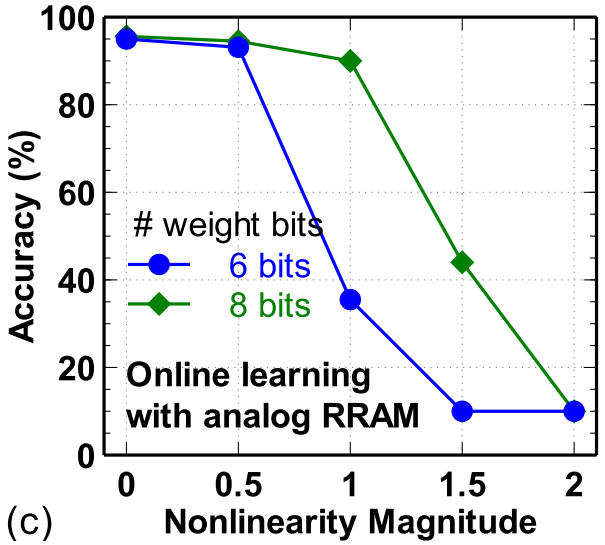
Output:

- Area,
- Latency,
- Energy,
- Accuracy

Weight Precision and Weight Update Nonlinearity



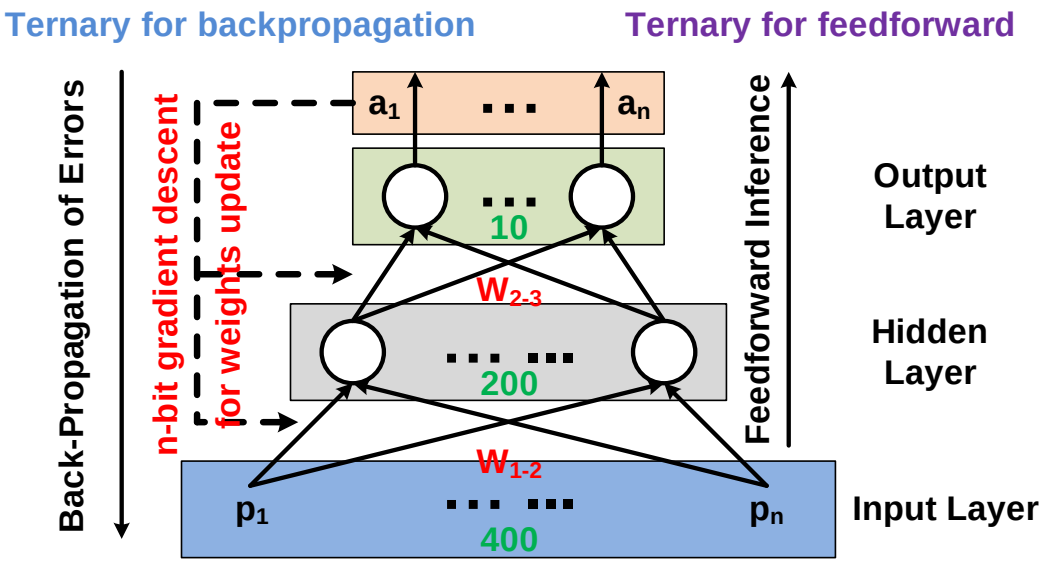


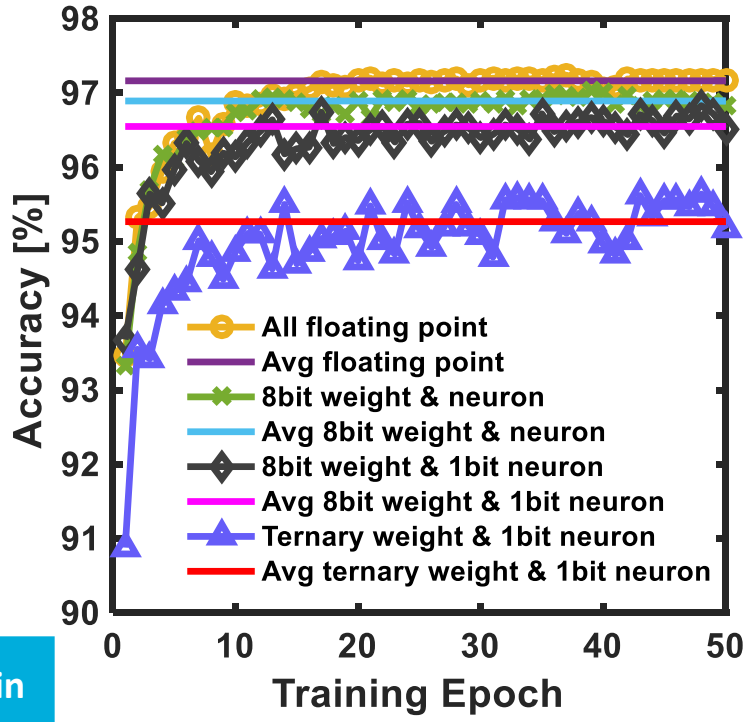


At least 6-bit is required for online learning, while 1 or 2-bit may work for offline classification. Nonlinearity significantly degrades accuracy for online learning.

Ternary Neural Network (TNN): Precision Reduction to Ternary Weight (+1,0,-1) for Feedforward

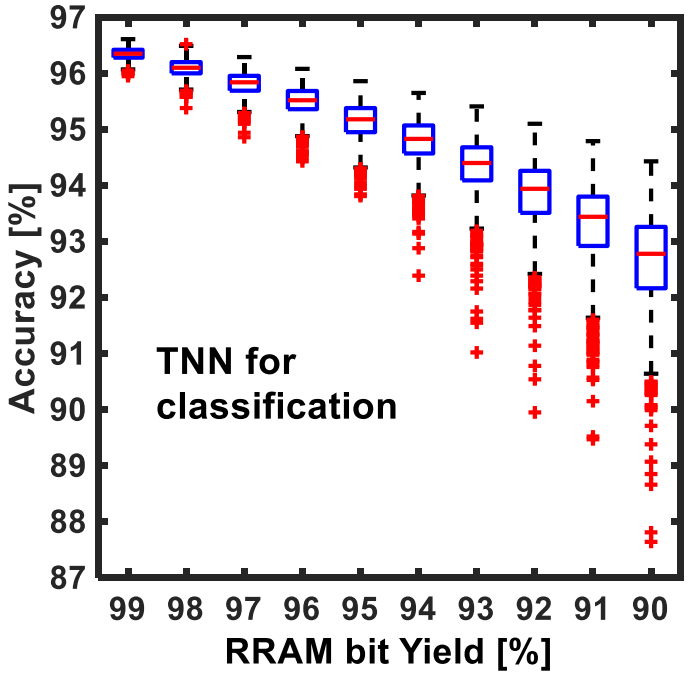
To allow the conventional digital (1-bit) RRAM work as binary synapse

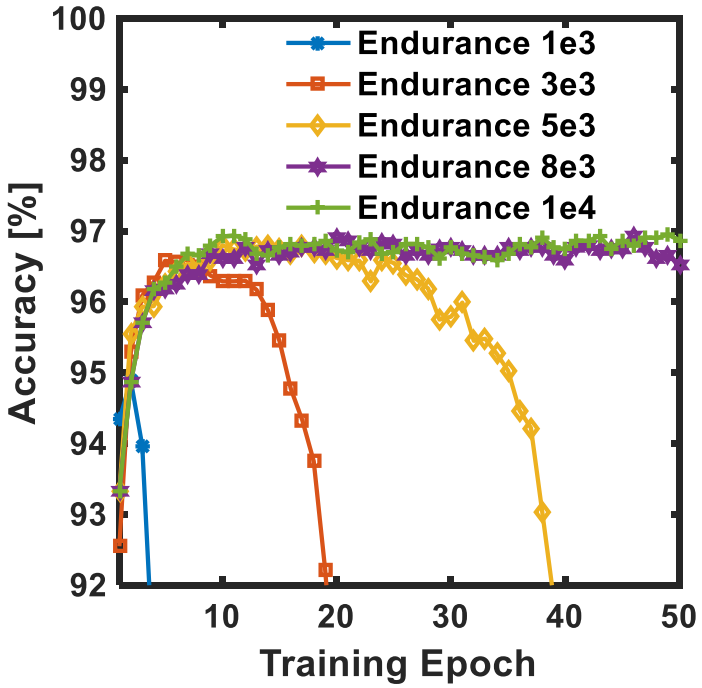




Binary expression b0b1..bn	Decimal expression	Normalized value	Round in decimal	Round in binary
011001	+25	+25/31	+1	01
100111	-25	-25/31	-1	10

Impact of RRAM Finite Yield and Endurance





For MNIST dataset, 99% bit yield and 1E4 cycling endurance is sufficient

Summary

- **Resistive devices can be tuned to the targeted multilevel (possibly by iterative programming), and offline classification is most suitable application scenario that achieves both low-power, fast and accurate recognition.**
- **For online training, "analog" synapses with continuous weights need to overcome challenges such as nonlinear weight update, and further improve on/off ratio and programming speed**
- **Digitalizing neural network with low-precision weights (e.g. ternary +1, 0, -1), allows today's "digital" RRAM arrays for online training and offline classification with high accuracy, which also shows good resilience to limited yield and endurance.**

Sponsor