Signature Driven Post-Manufacture Testing and Tuning of RRAM Spiking Neural Networks for Yield Recovery

Anurup Saha, Chandramouli Amarnath, Kwondo Ma (presenter) and Abhijit Chatterjee

Outline

- Problem motivation
- Background
- Variability modeling
- Signature driven testing
- Signature driven tuning
- Experimental results
- Conclusion

Problem Motivation

- RRAM crossbars enable low latency/energy compute-in-memory.
- Conductance variation degrades performance.
- Conductance variation profile is different for every manufactured chip. For example:

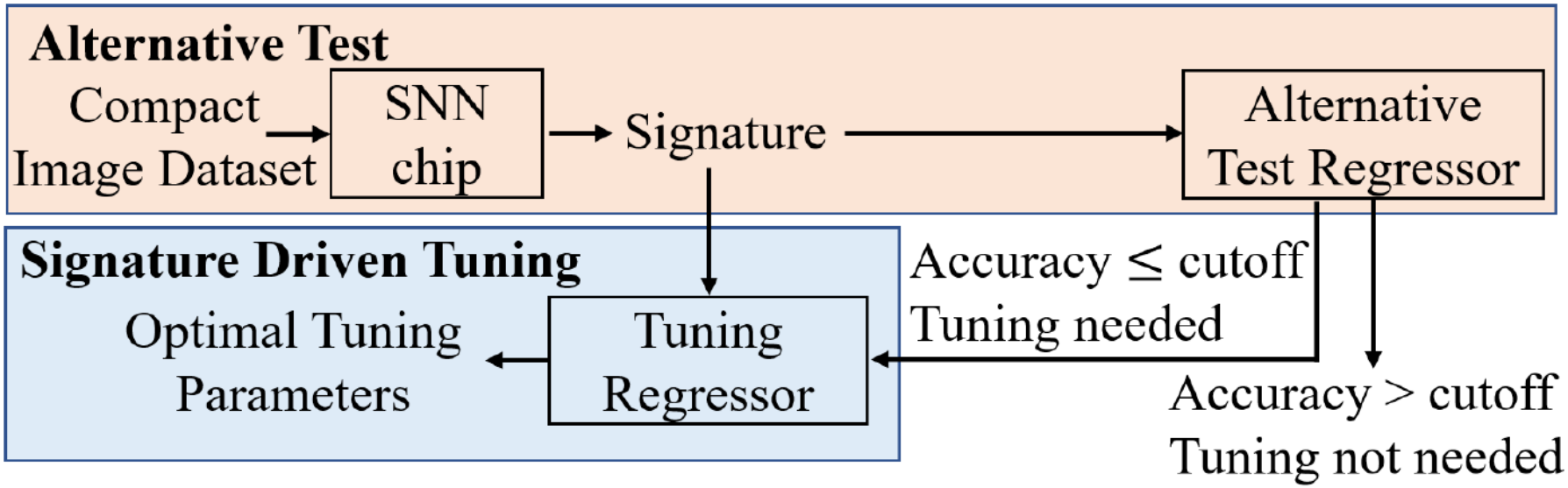
$$W = \begin{bmatrix} 0.21 & 0.33 \\ 0.50 & 0.12 \end{bmatrix}$$ Ideal Model Weights

$$W_1 = \begin{bmatrix} 0.23 & 0.38 \\ 0.54 & 0.14 \end{bmatrix} \quad W_2 = \begin{bmatrix} 0.16 & 0.28 \\ 0.43 & 0.11 \end{bmatrix}$$ Programmed weights in 2 chips

- Performance degradation varies between chips.

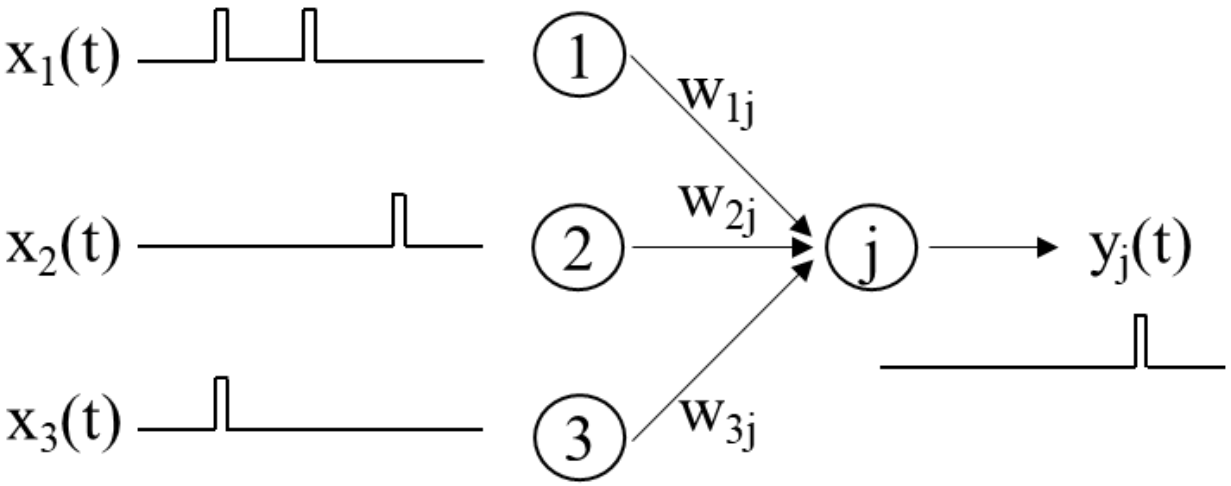
Problem Motivation

- Out-of-spec DUTs need to isolated (Testing).
- Out-of-spec DUTs need to be tuned for performance recovery (Tuning).



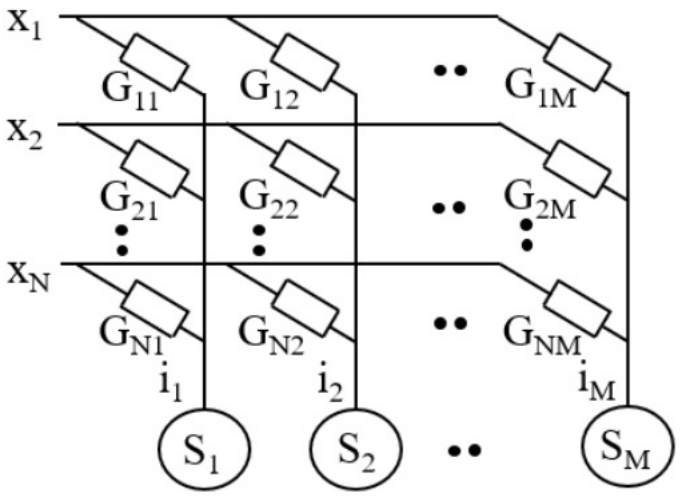
Background: Spiking Neural Networks

- Integrate and fire neuron : neuron emits spike when membrane potential exceeds spiking threshold.
- Rate encoding: rate of emitted spikes represent information.



Background: RRAM Crossbar

- Input spikes are scaled by weights within crossbar.
- Accumulated current is integrated within neurons.



A. Saha et al., "A Resilience Framework for Synapse Weight Errors and Firing Threshold Perturbations in RRAM Spiking Neural Networks," 2023 IEEE European Test Symposium (ETS)

Variability Modeling

1) Inject variations in gap dynamics fitting parameter (γ) of each RRAM device

$$\gamma = \gamma_0 + \Delta\gamma^{sys} + \Delta\gamma^{rand}$$

2) Sample systematic variation once per DUT.
3) Sample random variation once per RRAM device.
4) Calculate effective conductance.
5) Convert the effective conductance to effective weights.

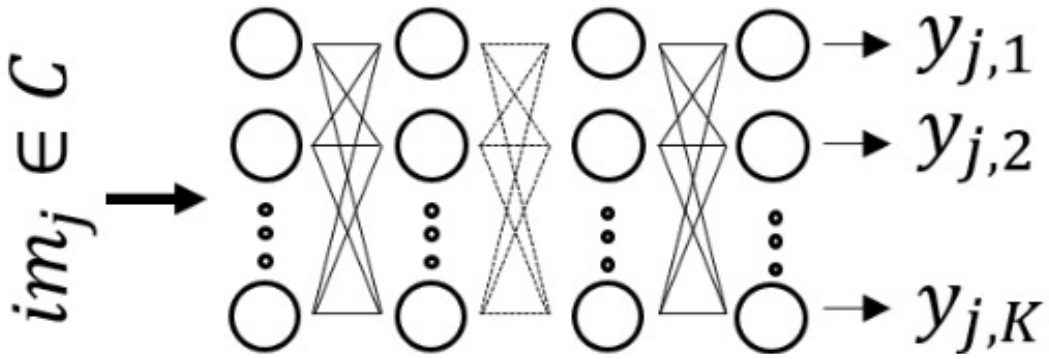
Why Signature Driven Test ?

- How to estimate classification accuracy of a DUT using Exhaustive Test?
  - Apply all images from test dataset.
  - Count number of correctly classified images.
  - Example: if a DUT correctly classifies 8000 out of 10000 applied images, its accuracy is 80%.
- Drawback: Exhaustive test requires inference on very large number of images.
- Key Idea: generate a signature and predict performance using signature.

Signature Driven Test: DUT Signature

- Select a compact image dataset (~32 images).
- Monitor DUT response corresponding to each image $R_j = [y_{j,1} \quad y_{j,2} \quad \dots \quad y_{j,K}]$.
- Stack the DUT responses to create signature

$$sig = [R_1 \quad R_2 \quad \dots \quad R_N]$$

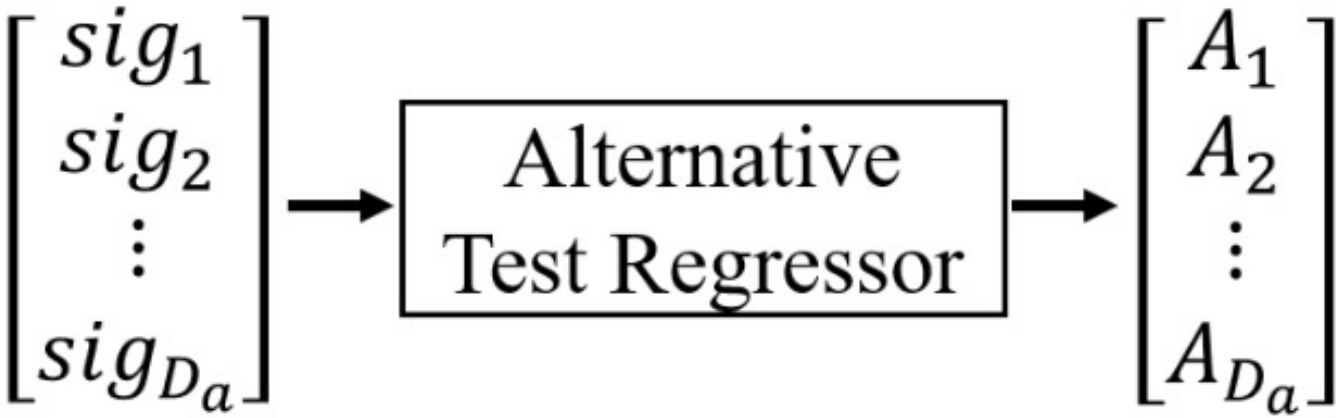


Signature Driven Test: Prediction

- Using the derived DUT signature, the DUT accuracy is predicted by our proposed alternative test regressor.
- The alternative test regressor is trained offline before real-time testing.

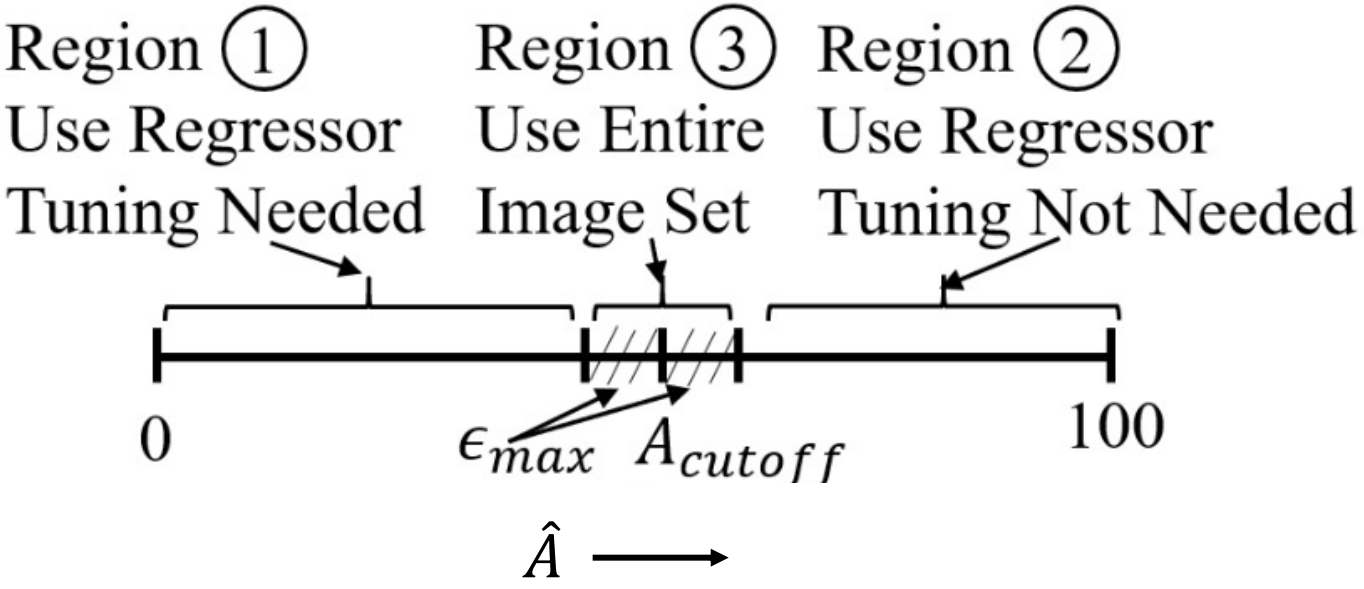
Signature Driven Test: Regressor

- How to train the alternative test regressor?
  - For D_a DUTs, generate signatures.
  - Measure accuracy of these DUTs using exhaustive test.
  - Train a regressor using this data.



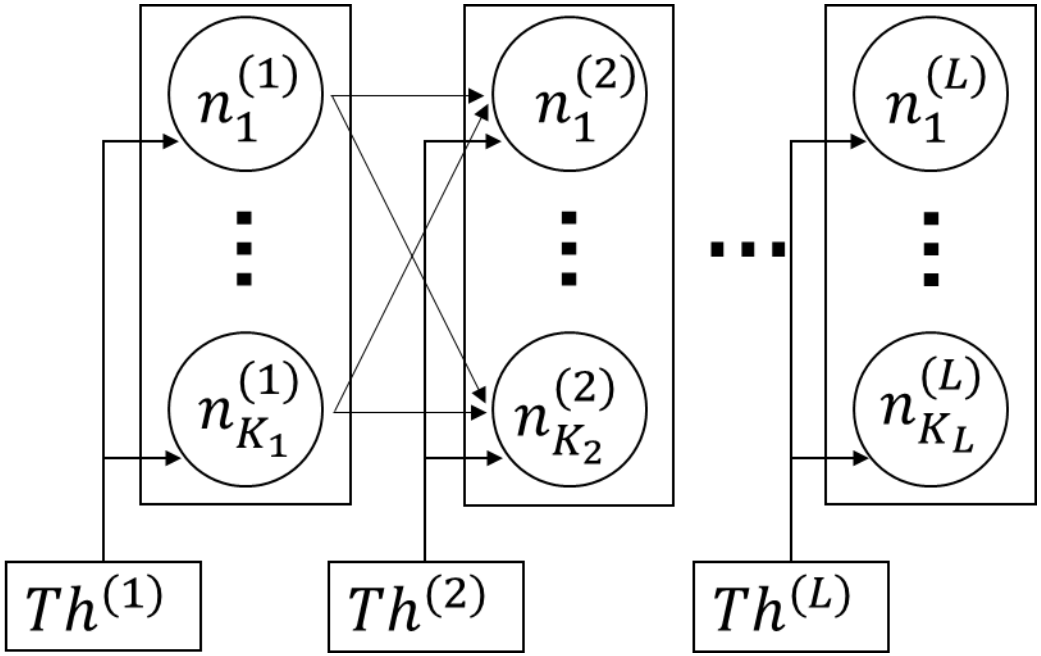
Signature Driven Test: Prediction Error

- Alternative test regressor can have prediction error.
- Example: $A = 79\%, \hat{A} = 81\%, A_{cutoff} = 80\%$.
- If $\left|\hat{A} - A_{cutoff}\right| < \epsilon_{max}$, use exhaustive test.



Signature Driven Tuning: Overview

- Spiking thresholds of each SNN layer are used as tuning knobs.
- Set the spiking thresholds to optimal values for performance recovery.



Signature Driven Tuning: Prediction

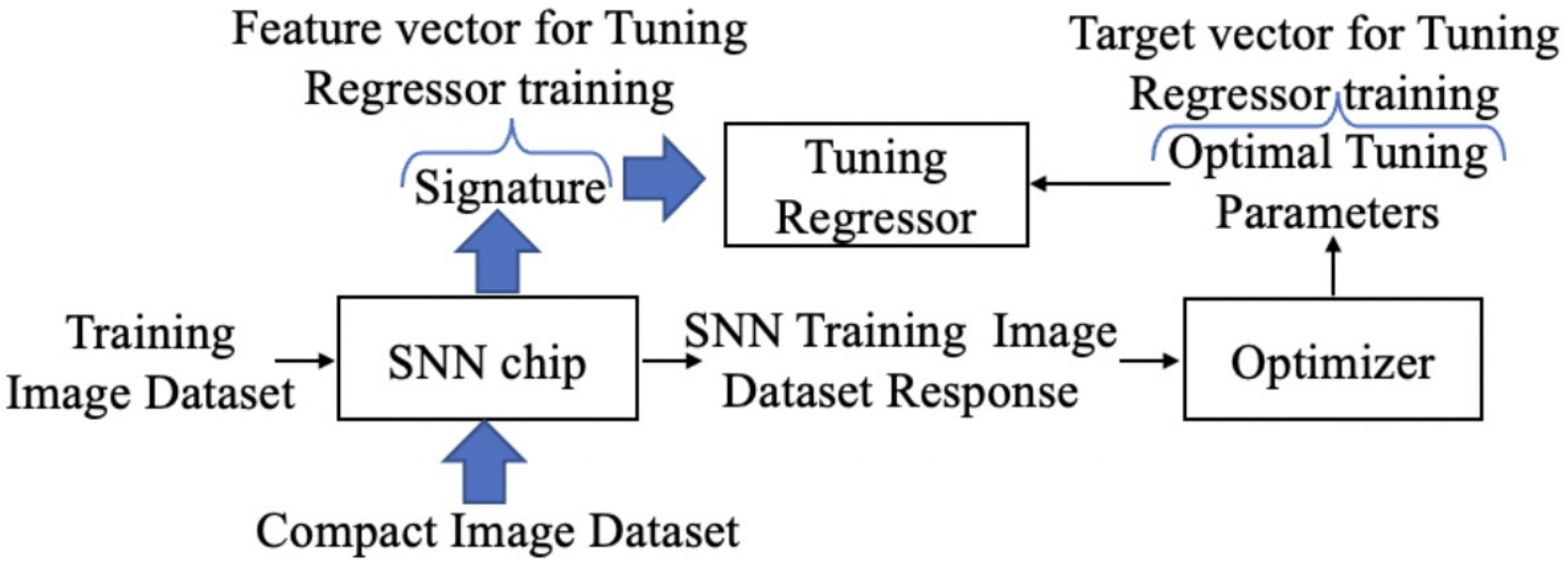
- Tuning regressor is implemented using nearest neighbor regressions.
- For D_b DUTs, the signatures and optimal tuning knobs are calculated using an offline optimization algorithm.
- During tuning, we find the "nearest" DUT as:

$$k^* = argmin_{k \in \{1,2,\ldots,D_b\}} ||sig - sig_k||_1$$

- Use the optimal tuning knob values from "nearest" DUT.

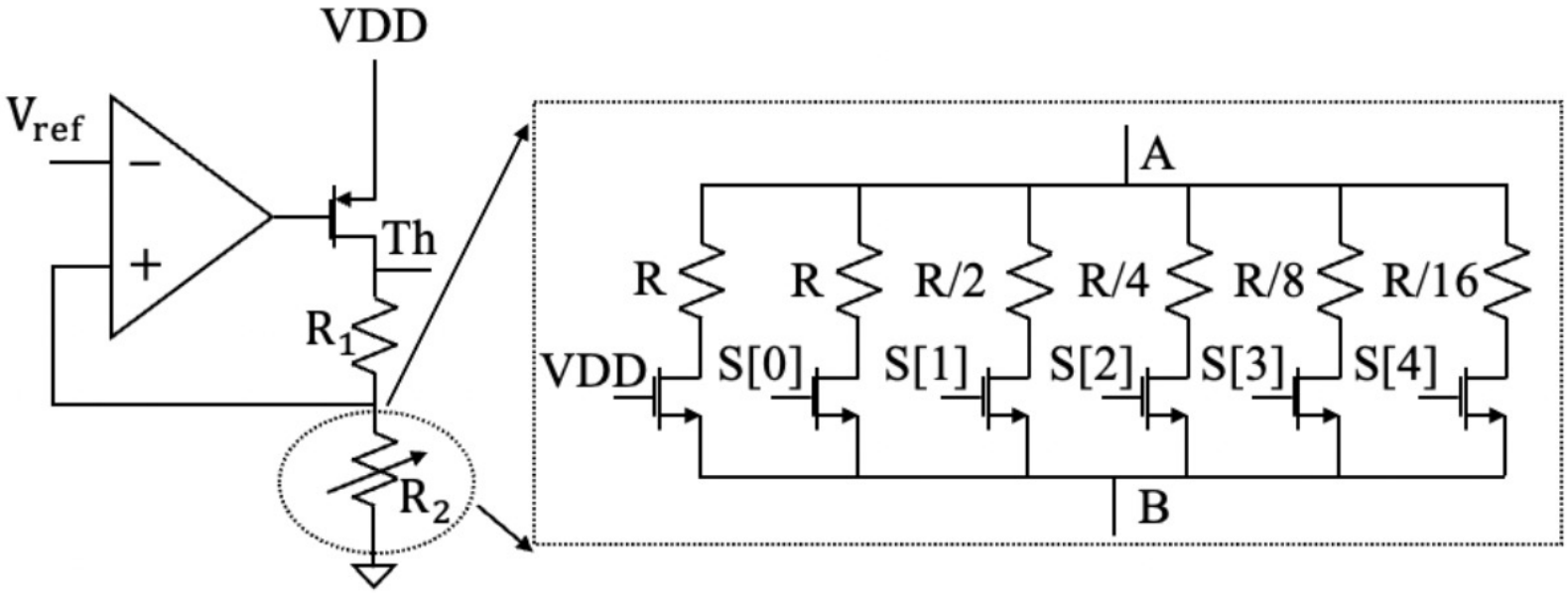
Signature Driven Tuning: Regressor

- For D_b DUTs
  - Find optimal spiking thresholds using backpropagation.
  - Generate signature.



Signature Driven Tuning: Implementation

- We generate $Th = V_{ref}(1 + \frac{R_1}{R_2})$.
- R_2 is implemented as a parallel combination of switches and resistors.

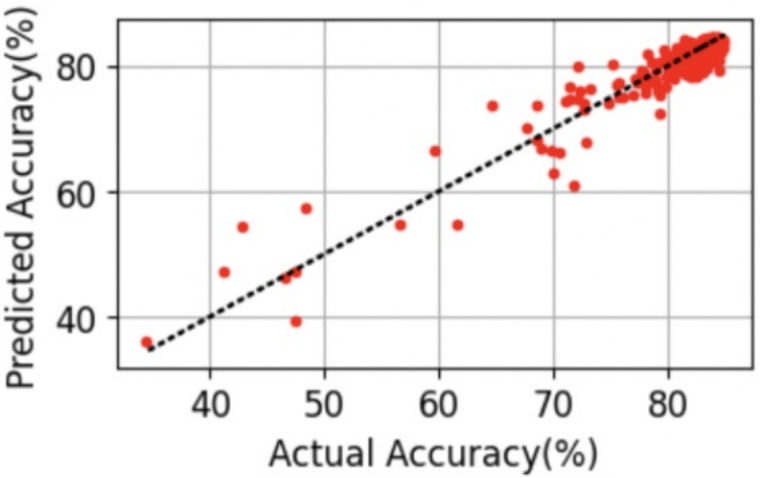


Results

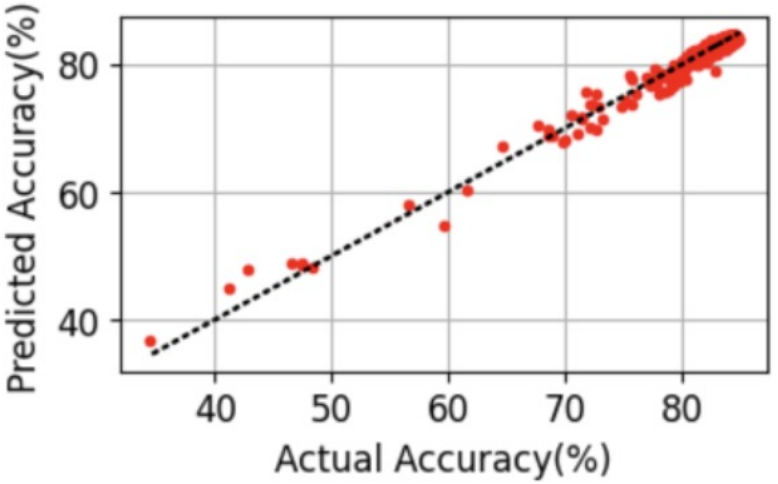
- Simulation Setup
  - Convolutional SNN architecture: VGG9.
  - Dataset: CIFAR10.
  - Baseline accuracy: 85.36%.
  - Variability: 50% systematic, 50% random.
  - Alternative test regressor: Gradient boosting regressor.
- Performance metrics
  - Signature driven testing: Mean absolute error.
  - Signature driven tuning: Yield improvement.
- Evaluated on 500 DUTs.

Results: Signature Driven Testing

- As we increase the size of the test stimulus, prediction becomes more accurate.



N = 4



N = 32

Results: Signature Driven Testing

- As N increases, MAE reduces.
- N = 32 and N = 64 gives similar MAE.
- We choose N = 32.

N	MAE (%)
4	1.03
8	0.77
16	0.70
32	0.54
64	0.52

Results: Signature Driven Tuning

- Lower allowable accuracy drop leads to lower yield.
- Tuning improves yield.

Allowable accuracy drop (%)	Yield without tuning (%)	Yield with tuning (%)	Yield improve-ment (%)
10	93.8	97.2	3.4
5	86.4	91	4.6
4	82.4	88.6	6.2
3	74	82.6	8.6

Results: Signature Driven Tuning

- Recovery of bad devices:

Allowable accuracy drop (%)	# of Bad DUTs before tuning	# of Bad DUTs after tuning	Recovery (%)
10	31	14	54
5	68	45	33
4	88	57	35
3	130	87	33

Conclusion

- A signature driven testing framework is proposed for RRAM spiking neural networks.
- Recalibration of spiking threshold is proposed for performance recovery.
- The testing and tuning frameworks are collaboratively designed to incorporate the unique DUT signature.

Thank You

- Corresponding author:Anurup Saha(asaha74@gatech.edu)