Self-Terminating Write of Multi-Level Cell ReRAM for Efficient Neuromorphic Computing

Zongwu Wang (Speaker)

Zhezhi He*, Rui Yang, Shiquan Fan, Jie Lin, Fangxin Liu, Yueyang Jia, Chenxi Yuan, Qidong Tang and Li Jiang*

Shanghai Jiao Tong University

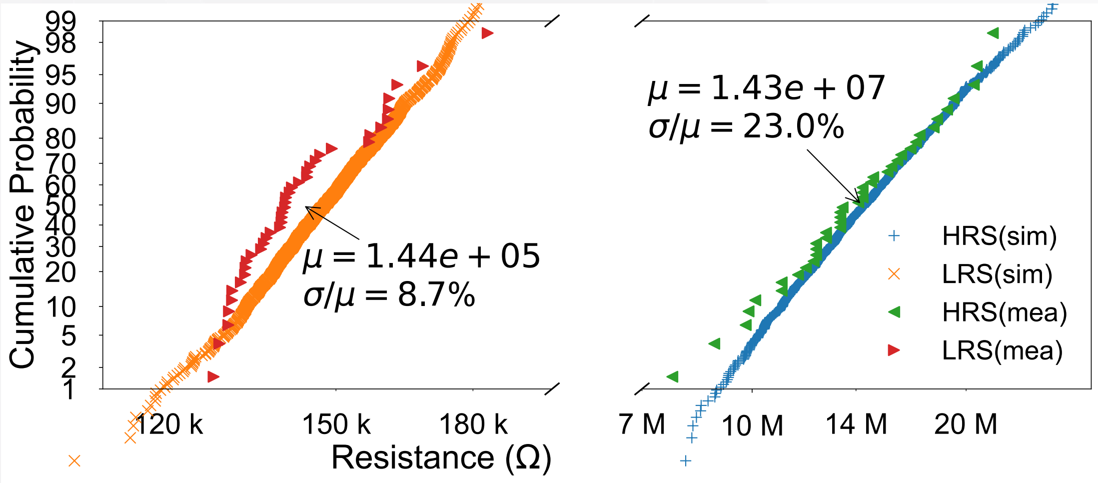
2022年2月28日

Self-Terminating Write Scheme Overview

Challenges In ReRAM-based PIM

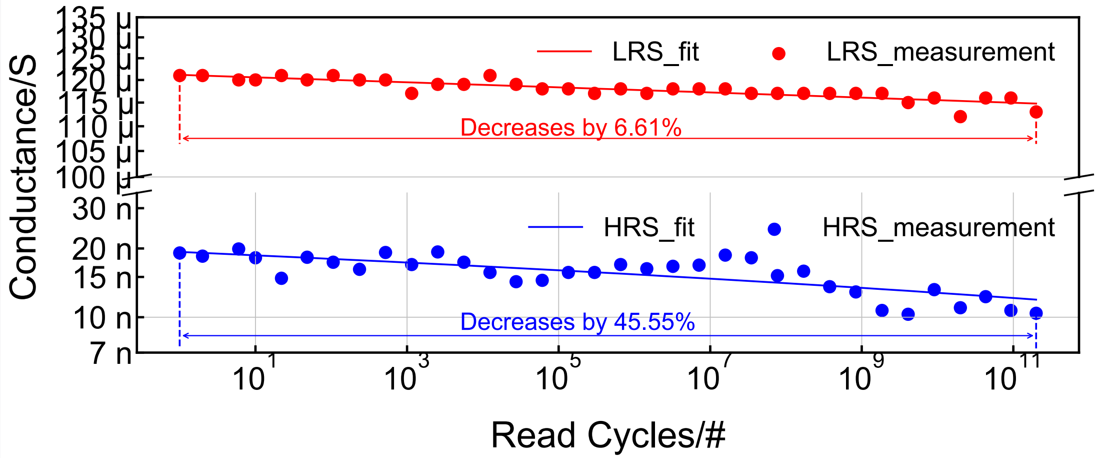
- **ReRAM has intrinsic write variation**
- **Read disturb induces resistance drifting**
- **Write-verify scheme is relative slow**

Write variation (CDF vs resistance)



- Measurement and simulation results comparison
- C2C variation exists in programming
- Set and Reset have significant write variation (8% and 23%, respectively)

Read disturb (conductance vs. #read times)



- ReRAM suffers from read induced drifting
- In-Memory computing equivalents to read
- Reliability test shows 6.6% and 45.6% drifting

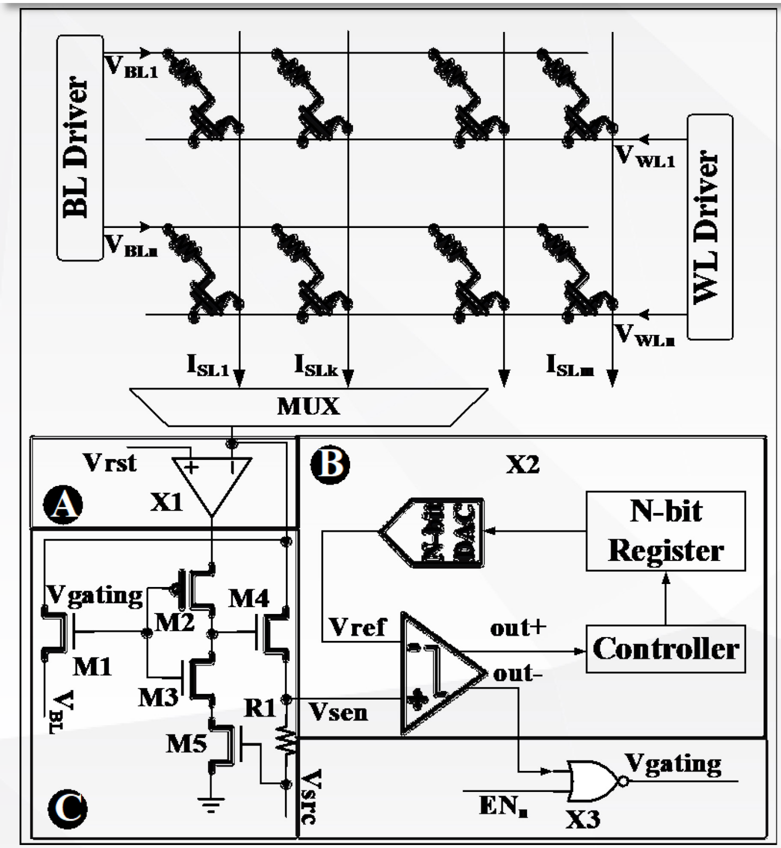
Proposed Solution

- **Heavily peripherals reuse achieves precise self-terminating scheme(2-bit)**
- **Pick appropriate programming range according to circuit design**
- **Compare to Write-verify scheme, Reduce the latency and energy by 4.7x and 2x, respectively**

Self-Terminating Write Scheme Design

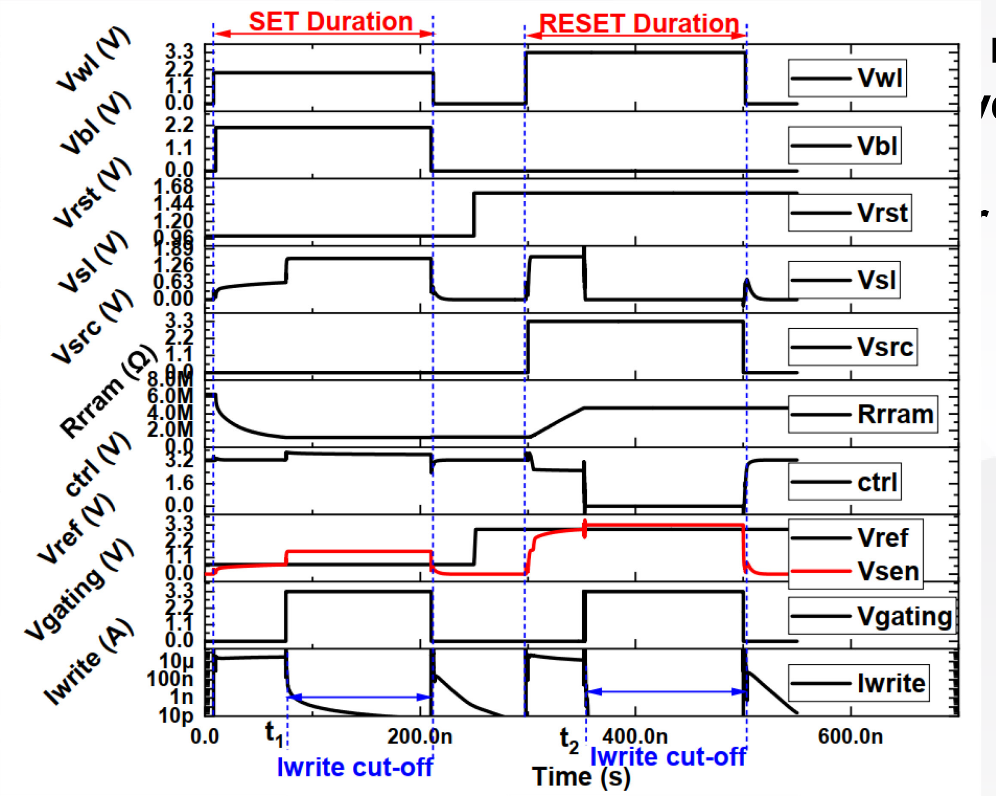
Proposed Self-Terminating Write Scheme

- Reuse the peripherals in ReRAM-based PIM system
- Implementing both Set and Reset termination with circuit sharing
- Compact design achieves low cost and fast feedback (high precision)



ReRAM MLC STW Schematic

- Reus… monitoring, and
terminа… ved
- Thre…
  - A: …
  - B: …
  - C: …
- Thre…
  - Co…
  - Set…



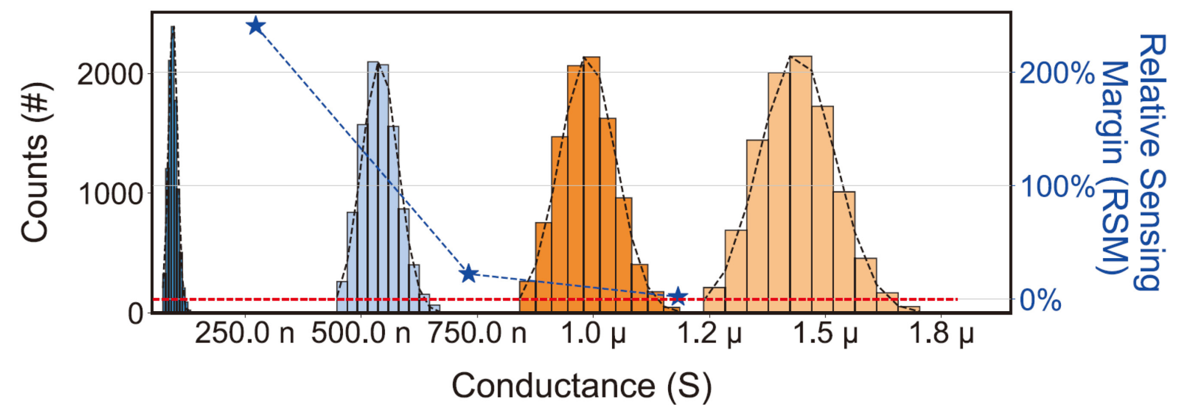
Programming Waveform

Self-Terminating Write Scheme Evaluation

	Structure	Area	Terminate	Precision
This work	**2Amp+5T+NOR**	**Medium**	**both**	**2 bits**
JSSC-2013 [10]	2Amp+R+30T +DelayUnit+others	Large	both	1 bit
ISSCC-2014 [24]	4T	Small	set	1 bit
IEDM-2017 [6]	RESET: Amp+4SW+6T SET: 5T	Medium	both	1 bit
ISSCC-2021 [25]	2Amp+R+5T+3INV +AND+Delay Unit	Large	set	1 bit

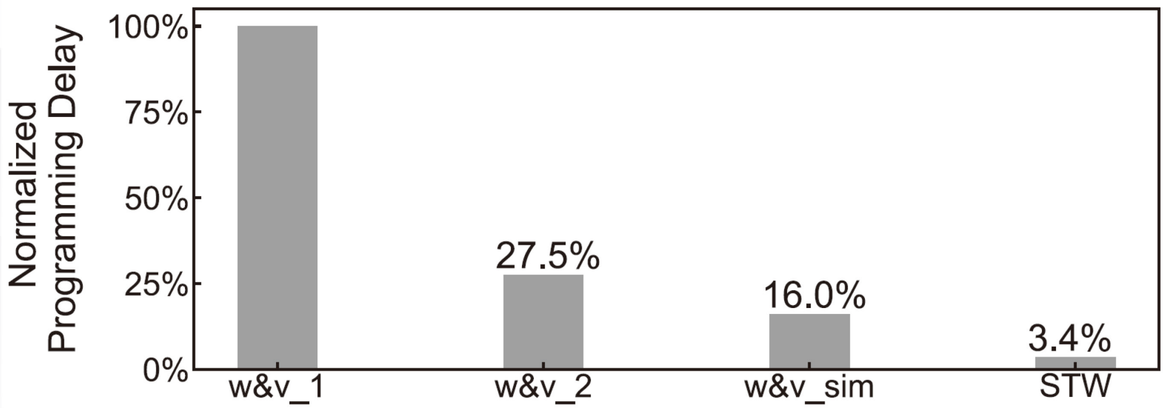
Comparison with previous works (area, programming polarity and precision) :

- Reduces area overhead by peripherals reuse
- Supports both Set and Reset termination
- Achieves 2-bit MLC self-terminating



10^4 trials MC simulation with range selection algorithm

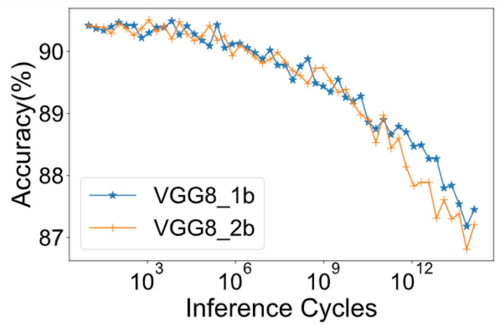
- the proposed STW scheme achieves 2-bit precision



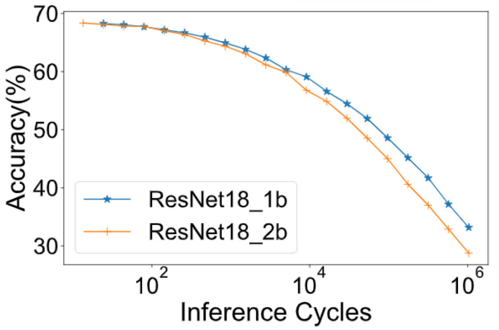
Latency comparison between different schemes

- STW scheme shows 4.7x speedup (conservative)

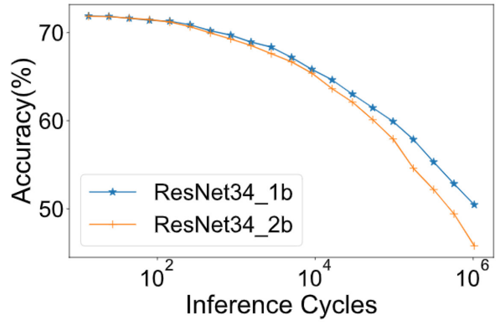
Self-Terminating Write Scheme Evaluation



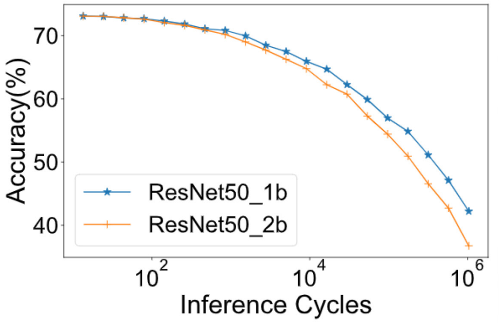
(a) VGG8 on CIFAR-10



(b) ResNet-18 on ImageNet



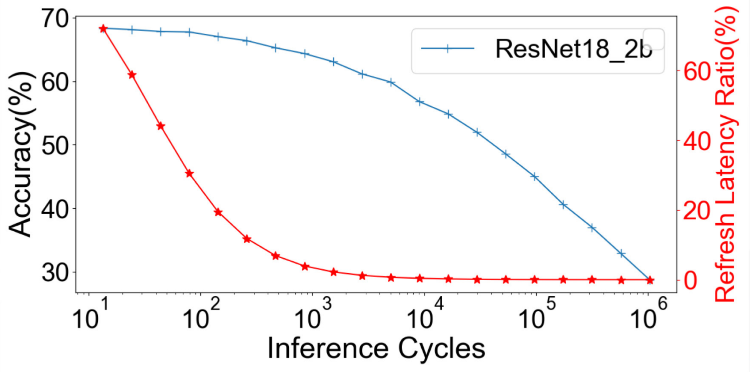
(c) ResNet-34 on ImageNet



(d) ResNet-50 on ImageNet

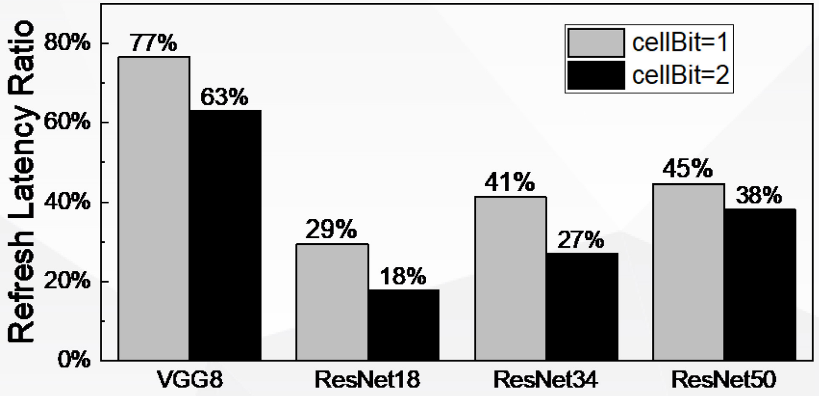
Impact of Read disturb on inference accuracy:

- ➢ Accuracy loss with the continuous inference after the network deployed
- ➢ MLC can reduce the storage/computation cost, but it is more vulnerable to read disturb



Refresh Frequency and Accuracy Balancing:

- ➢ The lower the refresh frequency, the lower the proportion of refresh delay, but the lower the accuracy



Proportion of delay on different networks

- ➢ Ratio of refresh latency is low on compact networks
- ➢ From the perspective of deployment cost, programming delay is an important factor

Conclusion

1. An auto-calibrate Framework

- Provides easy-use and confidence ReRAM compact model

2. A valid self-terminated programming scheme for MLC

- Heavily reuses the original peripheral
- Compact design achieves low cost and high precision
- Reduce the latency and energy by 4.7x and 2x, respectively

3. Cross-layer simulation (device/circuit/system) to validate the design