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

- 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)

Proposed Self-Terminating Write Scheme

- Reuse peripherals to implement on-the-fly monitoring, and terminate automatically once the target is achieved
- Three modules in peripherals:
  - A: TIA module for current-to-voltage converter
  - B: SAR-ADC module for comparison
  - C: verdict control module (new add)
- Three operation modes control:
  - Computing mode (EN high)
  - Set and Reset termination mode (EN low)

ReRAM MLC STW Schematic

Programming Waveform
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

Latency comparison between different schemes:

- STW scheme shows 4.7x speedup (conservative)
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

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

Refresh Frequency and Accuracy Balancing:
- The lower the refresh frequency, the lower the proportion of refresh delay, but the lower the accuracy
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