HD-RRAM: Improving Write Operations on MLC 3D Cross-point Resistive Memory

Chengning Wang[1], Dan Feng[1,2], Wei Tong[1,2], Yu Hua[1,2], Jingning Liu[1,2], Bing Wu[1], Wei Zhao[1], Linghao Song[3], Yang Zhang[1], Jie Xu[1], Xueliang Wei[1], and Yiran Chen[3]

[1]*Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China.*

[2]*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, China.*

[3]*Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA.*

***Abstract*—Multilevel cell (MLC), cross-point array structure, and three-dimensional (3D) array integration are three technologies to scale up the density of resistive memory. However, composing the three technologies together strengthens the interactions between array-level and cell-level nonidealities (IR drop, sneak current, and cycle-to-cycle variation) in resistive memory arrays and significantly degrades the array write performance. We propose a nonidealities-tolerant high-density resistive memory (HD-RRAM) system based on multilayered MLC 3D cross-point arrays that can weaken the interactions between nonidealities and mitigate their degradation effects on the write performance. HD-RRAM is equipped with a double-transistor array architecture with multiside asymmetric bias, proportional-control state tuning, and MLC parallel writing techniques. The evaluation shows that HD-RRAM system can reduce the access latency by 27.5% and energy consumption by 37.2% over an aggressive baseline.**

***Index Terms*—memory devices, crossbar, array nonidealities, memristor-based memory, memory array operation scheme.**

I. Introduction

Resistive memory (RRAM) is promising to be used as high-density energy-efficient storage-class memory [2]–[4]. To scale up the density of resistive memory, there are three common technologies: multilevel cell (MLC), cross-point array structure, and three-dimensional (3D) array integration [1], [5], [6]. However, composing the three high-density technologies together strengthens the interactions between array-level and cell-level nonidealities and degrades the array write performance [7]. We first build a dynamic memory array model for multilayered MLC 3D cross-point memory arrays to analyze the interactions between cells, interconnects, and vertical pillar access transistors (VPAT) during write operations. We have three observations. (1) In the conventional single-transistor array (SITA) architecture, the small-size VPAT divides up more than 30% of the array bias voltage, and only a small portion of voltage falls on the selected cell. (2) For the single-side bias scheme, the effective voltage and the corresponding multilevel write latencies are highly non-uniform between cells in the array, due to the interaction between interconnect IR drop and cell sneak currents. The cell-to-cell non-uniformity of effective write voltage makes it hard to share a group of multilevel write latencies among all the multilevel cells. (3) The effective write voltage across the selected cells significantly increases with time during a RESET process, due to the time-varying resistance of the selected cells and the dynamic voltage-dividing effect among the VPAT, the interconnects, and the memory cell along the selected write current path.

This work was supported in part by the National Natural Science Foundation of China under Grant 61832007, Grant 61821003, Grant 61772222, and Grant U1705261; in part by the Fundamental Research Funds for the Central Universities under Grant 2019kfyXMBZ037; in part by the Zhejiang Lab (No. 2020AA3AB07); and in part by the National Science and Technology Major Project No. 2017ZX01032-101. (Corresponding author: Dan Feng. e-mail: dfeng@hust.edu.cn).

The research article [1] of this extended abstract can be viewed at https://ieeexplore.ieee.org/abstract/document/9130780.

II. Design of HD-RRAM System

Overview and Motivation: By weakening the interactions between IR drop and the other two nonidealities and conquering multiple nonidealities in each part of operation design stages respectively, we can cope with each nonideality of MLC 3D cross-point resistive memory during memory access operations. Motivated by (1) the current-dividing effect along the vertical pillars, (2) the pulse-amplitude-dependent feature of cycle-to-cycle write variation, and (3) the pass-through feature of intermediate resistance states, HD-RRAM system improves write performance from three dimensions: (1) improving the resistive-switching velocity under an array write pulse, (2) reducing the average number of write-and-verify iterations in a voltage pulse ramp, and (3) improving the cell-level write parallelism in an array. The performance improvement of multilevel write operations can be decomposed into the product of the three dimensions.

Double-Transistor Array Architecture: The large voltage drop across the VPAT lowers the effective write voltage across the selected cell. To reduce the current driving requirement of the small-size VPATs, we propose a double-transistor array (DOTA) architecture with 2TnR cell organization along pillars for 3D cross-point memory arrays, as shown in Fig. 1. The top-layer and bottom-layer VPATs can be viewed as parallel connected, thus the current drivability of VPATs is improved.

Multiside Asymmetric Bias: To reduce the undesired voltage drop along the selected multidirectional write current paths, we propose a multiside asymmetric bias scheme based on the DOTA architecture. For write operations, MAB connects two ends of the selected pillar electrodes and four sides of the selected plane electrode to the voltage sources simultaneously.

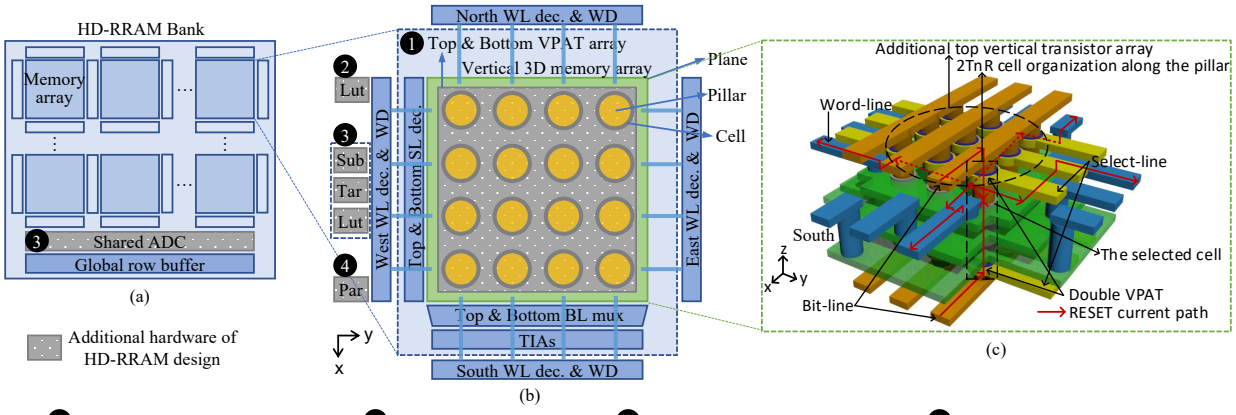

Fig. 1. Overview of HD-RRAM system [1]. (a) array organization in a memory bank. (b) memory array and periphery. (c) double-transistor array architecture.

Meanwhile, MAB only connects one end of unselected pillar electrodes and one side of unselected plane electrodes to the voltage source [1]. Compared with SITA-based MWD [2], DOTA-based MAB can lower the array bias voltage from 2.9 V to 2.55 V with RESET velocity improvement of around 6.3 times and array RESET energy reduction of 6.9 times, owing to the current-dividing effects along the selected pillars.

Proportional-Control Multilevel State Tuning: For multi-level write operations, iterative write-and-verify process can be used to tolerate cycle-to-cycle write variation [8]. We fully utilize the feedback information obtained from the verify pulses to reduce the average number of write-and-verify iterations and the time taken by the verify pulses during multilevel write operations. We propose a proportional-control variation-adaptive multilevel state tuning algorithm at the array level. The key idea is that if the current state resistance is far away from the target state resistance, the memory cell is tolerable to large resistance variation, and we apply large pulses but with large absolute variation to make the current resistance go faster to the target resistance range. If the current state resistance reaches close to the target state resistance range, the memory cell is vulnerable to variation, and we use small write pulses with small absolute resistance variation to refine the resistance change. We regulate the amplitude of the next write pulse in a linear form [1].

Multilevel Cell Parallel Writing: Motivated by the pass-through feature of intermediate resistance states, we further propose an array-level multilevel cell parallel writing method. The method extracts the common write-and-verify iterations among the to-be-written multilevel cells to improve the cell-level write parallelism during write-and-verify processes [1].

Overhead Discussion: We implement the hardware components of HD-RRAM system in Synopsys Design Compiler. The top-layer VPAT arrays and their decoders incur 16.9% extra cost-per-bit. The total area of peripheral CMOS circuitry is reduced by 6.5%, and the power consumption of bank-level peripheral circuitry is reduced by 147.2 mW. The peripheral circuitry including the ADCs, target registers, substracters and lookup tables to implement POST in a bank takes extra 3.9% chip area, 70.2 mW power, and 5.63 ns latency. The peripheral logic to support MPW in a bank consumes extra 1.4% chip area, 12.2 mW power, and 0.2 ns latency, which are considered in the system-level evaluation.

III. System-Level Evaluation Results

To corroborate the effectiveness of the proposed array-level write operation scheme for multilayered MLC 3D cross-point resistive memory, we evaluate and compare the HD-RRAM system with the state-of-the-art design (SITA-based MWD [2] + ESP [9] + IPST [8]) under SPEC CPU2006 benchmarks. Parasitic effects along vertical pillars in resistive memory arrays are also considered in the evaluation. Compared with the state-of-the-art design, HD-RRAM reduces the average write latency and the memory access latency by 34.1% and 27.5% across the benchmarks, respectively. The average IPC improvement of HD-RRAM system is 19.4% across the benchmarks. Additionally, the reductions of energy and energy-delay product of the HD-RRAM system are 37.2% and 54.4% on average across the benchmarks, respectively.

References

[1] C. Wang *et al.*, "Improving multilevel writes on vertical 3D cross-point resistive memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, pp. 1–14, 2020.

[2] C. Xu *et al.*, "Architecting 3D vertical resistive memory for next-generation storage systems," in *Proc. 2014 IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, 2014, pp. 55–62.

[3] Y. Zhang *et al.*, "Tiered-ReRAM: A low latency and energy efficient TLC crossbar ReRAM architecture," in *Proc. 2019 35th Symp. Massive Storage Syst. & Technol. (MSST)*, 2019, pp. 92–102.

[4] W. Wen *et al.*, "Exploiting in-memory data patterns for performance improvement on crossbar resistive memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 10, pp. 2347–2360, 2020.

[5] C. Wang *et al.*, "Design and analysis of address-adaptive read reference settings for multilevel cell cross-point memory arrays," *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5347–5352, 2019.

[6] C. Wang *et al.*, "Improving write performance on cross-point RRAM arrays by leveraging multidimensional non-uniformity of cell effective voltage," *IEEE Trans. Comput.*, pp. 1–15, 2020.

[7] C. Wang *et al.*, "Cross-point resistive memory: Nonideal properties and solutions," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 24, no. 4, pp. 46:1–46:37, 2019.

[8] F. Alibart *et al.*, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, pp. 1–7, 2012.

[9] C. Xu *et al.*, "Understanding the trade-offs in multi-level cell ReRAM memory design," in *Proc. Design Autom. Conf. (DAC)*, 2013, pp. 1–6.