MULTIFUNCTIONAL RRAM CHIP WITH CONFIGURABILITY FOR SPARSITY-AWARE IN-MEMORY ISNG MACHINE

Wenshuo Yue[1], Zhaokun Jing[1], Bonan Yan[1,3], Yaoyu Tao[3], Teng Zhang[1], Ru Huang[1] and Yuchao Yang[1,2,3,4*]*

[1] Beijing Advanced Innovation Center for Integrated Circuit, School of Integrated Circuits, Peking University, Beijing, China.[2] School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China.[3] Center for Brain Inspired Chips, Institute for Artificial Intelligence, Frontiers Science Center for Nano-optoelectronics, Peking University, Beijing, China, [4] Center for Brain Inspired Intelligence, Chinese Institute for Brain Research (CIBR), Beijing, Beijing, China.

*Corresponding Author's Email: tengzhang@pku.edu.cn, yuchaoyang@pku.edu.cn

ABSTRACT

The Ising machine is a kind of annealing processor. When a combinatorial optimization problem is mapped on an Ising graph, the Ising machine calculates the physical evolution of this system and solves the problem. RRAM-based in-memory computing (IMC) is an important technique that can build up Ising machines. However, the high sparsity of the Ising graphs still fundamentally limits the time and energy efficiency. In this work, we propose a multifunctional RRAM chip that suits sparsity-aware in-memory computing Ising machine, which contains RRAM accelerated content-addressable memory (CAM), multiply-accumulate (MAC) unit, and true random number generator (TRNG) to work collaboratively. Such RRAM-based Ising machine provides significant improvement in both computing speed and energy consumption.

INTRODUCTION

Combinatorial optimization problems (COPs) refer to the problems that consist of the combination of discrete variables. The number of combinations of these variables increases exponentially with the number of variables. Thus, COPs are NP-hard problems and cannot be solved by enumeration of variables in a limited time [1]. Traveling salesman problems, graph coloring problems, and max-cut problems are typical COPs[2, 3]. Ising machine has been proposed to find the COP solution by solving the lowest energy state of Ising models via simulated annealing method and has attracted extensive interest recently [4]. Usually, the Ising machine cannot provide the optimal solution of a COP, but it can provide a comparatively good solution that can benefit the actual usage.

Nevertheless, the Ising annealing faces two major challenges: a) the graph aggregation dominates the latency due to frequent memory access, and b) the adjacency matrices of the Ising graph are highly sparse and require excessive memory capacity to store redundant zeros. Previous attempts in building silicon-based digital Ising chips[5, 6] and quantum computers[7] still face difficulties in high hardware cost and low operation temperature.

Here, we demonstrate a novel sparsity-aware IMC Ising machine based on multifunctional TiN/TaO_x/HfO_2/TiN RRAM to conquer the abovementioned computing bottlenecks. The 1T1R RRAM array has been exploited in varied configurations to implement 3 fundamental operations with high efficiency: a) CAM based on 2T2R subarray, b) MAC based on 1T1R subarray, and c) random selector based on 1T1R subarray to enable the simulated annealing process. A device and system co-design strategy is adopted to optimize the speed and energy efficiency, and only nonzero submatrices are processed. Such sparsity-aware in-memory Ising machines can achieve large reductions in computing latency and energy consumption.

RESULTS AND DISCUSSION

RRAM Device and Array

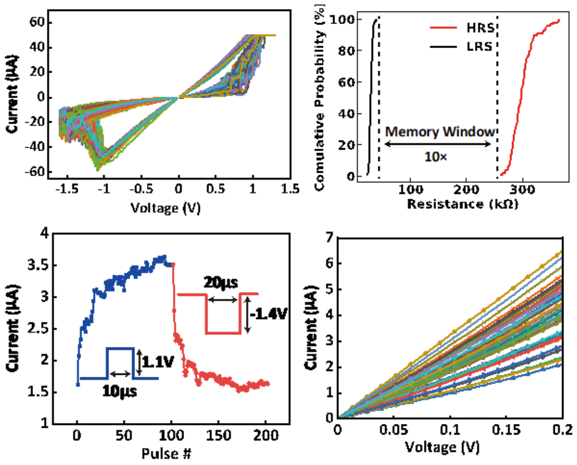

Figure 1: (a) Repeatable bipolar resistive switching of the RRAM device. (b) Distribution of 1T1R cells, showing a large memory window. (c) Analog property under 100 positive/negative biased pulses. (d) High I-V linearity for a wide range of conductance states.

The 36×32 TiN/TaO_x/HfO_2/TiN-based RRAM array was fabricated using a 180 nm foundry process. The array core uses 1T1R architecture with RRAM fabricated between M5 and M6. A customized on-board test platform was built for measurements. The RRAM devices show highly stable bipolar resistive switching upon 1.2V/-1.5 V voltage after forming (Fig. 1a), >10× memory window (Fig. 1b) and programmable analog state (Fig. 1c) that can be fine-tuned for versatile applications. The high I-V linearity (Fig. 1d) is desirable for both MAC and CAM operations. These RRAM properties have formed the physical substrate for building sparsity-aware IMC Ising machines.

To accommodate such a sparsity-aware IMC Ising machine, we have fabricated a 36×32 TaO_x/HfO_2 1T1R array. The array is configured using 3 different bit cell structures and operating conditions. Specifically, the 1T1R bit cell is used for in-memory MAC operation, and two such 1T1R bit cells can form a 2T2R CAM cell, while the 1T1R bit cell with inherent stochasticity can be used as TRNG. This has thus achieved CAM-based addressing, MAC operation, and randomness generation. These 3 different RRAM modules, together with the partial sum registers and global buffer, constitute the hardware of the proposed IMC Ising machine.

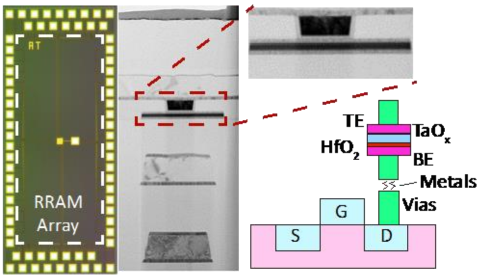

Figure 2: Die photo and TEM of the 1T1R RRAM chip and schematically illustrated cell structure with BEOL $TiN/TaO_X/HfO_2/TiN$ stack.

RRAM for CAM

Since the compression process has extracted nonzero elements from a highly sparse adjacency matrix (with different $\boldsymbol{C_i}$, where $\boldsymbol{i}$ is an integer), the partial sum results from MAC need to be reconstructed to their corresponding positions. The results from different blocks with the same $\boldsymbol{C_i}$ encodings need to be accumulated together. Hence, we use CAM to search $\boldsymbol{C_i}$ and enable the corresponding post-MAC accumulators. In CAM mode, the bit lines (BLs) serve as match lines (MLs), which are pre-charged to high voltages before searching (Fig. 3). The searched words are applied through WLs.

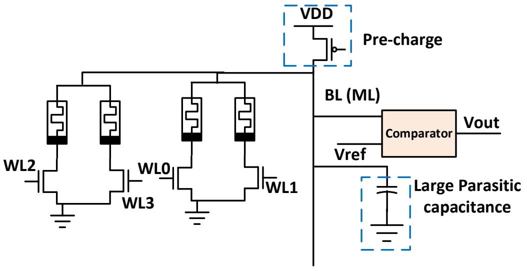

Figure 3: 2T2R CAM structure with comparator.

The CAM cell is experimentally tested using the circuit platform in Fig. 13. According to the measured voltage decay on ML (Fig. 4a), there is a large sensing margin between match and mismatch cases. We deployed an off-chip operational amplifier to amplify the output signal on ML. The difference of sensing delay between match and mismatch was ~40%. Using on-chip sense amplifier is able to reduce the half-time of ML voltage decay to ~10 ns (Fig. 4b). In fact, a tradeoff is found to exist in the CAM design between its speed and energy efficiency. We have explored the design space of RRAM resistance. Lower RRAM resistance makes ML sensing faster but induces higher energy consumption due to increased current, and vice versa for high resistance. Simulation results show that LRS = 30 kΩ is at the cross point and is hence optimal for both energy efficiency and computing speed. Moreover, the energy consumption reduces rapidly as on/off ratio increases to 10, and changes slightly when on/off ratio is >10. The thickness of HfO_2 plays a key role in deciding the device resistance and on/off ratio. Based on the aforementioned tradeoff, we have adopted a HfO_2 thickness of 8 nm, and used such RRAM arrays for subsequent Ising machine applications.

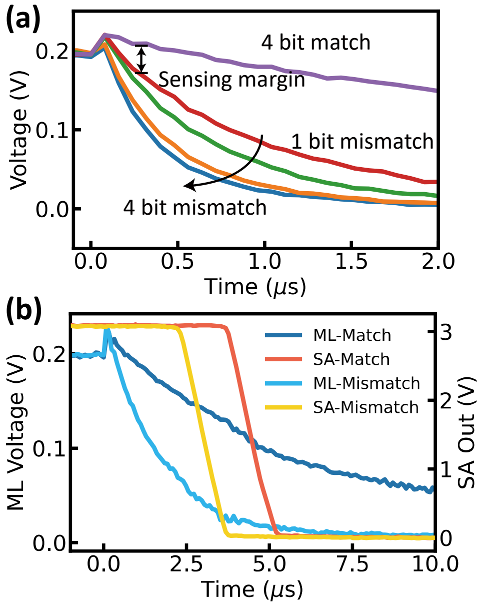

Figure 4: (a) Measured result of ML voltage decay for 4-bit classification. (b) Comparison of output results for match/mismatch using an off-chip operational amplifier.

RRAM for MAC

The Ising graph aggregation is achieved via RRAM-based MAC operations. Measured results of 4×4 MAC, where 5-state output current can be distinguished with large margins. The amplification of the signal was tested using an off-chip operational amplifier. Conventional MAC consumes extensive energy in computing zeros in graphs with high sparsity. Although RRAM cell in HRS only induces 0.1× static current compared to LRS, up to 91% energy will be consumed on

HRS cells when computing a typical Ising graph due to its high sparsity. Our sparsity-aware computing has largely alleviated this problem, where ~81% of zeros are eliminated from computing. This has reduced the energy consumption in MAC by ~11×.

RRAM for Random Selector

A random selector is needed to produce randomness for the simulated annealing algorithm. The random codes are generated from RRAM-based TRNG due to cycle-to-cycle variation. To evaluate the bandwidth of random number generation, we tested the switching speed of RRAM devices (Fig. 5). The RRAM-based TRNG can achieve ~80Mb/s output bandwidth using a moderate programming condition (2.4 V/-2.1 V, 200 ns pulse for set/reset) and 32 BLs. Fig. 5c shows the bitmap of 6.5 kbits random numbers and the evolution of the nodes' color with stochasticity in the Ising machine. Therefore, the RRAM-based TRNG can meet the experiments' bandwidth demands.

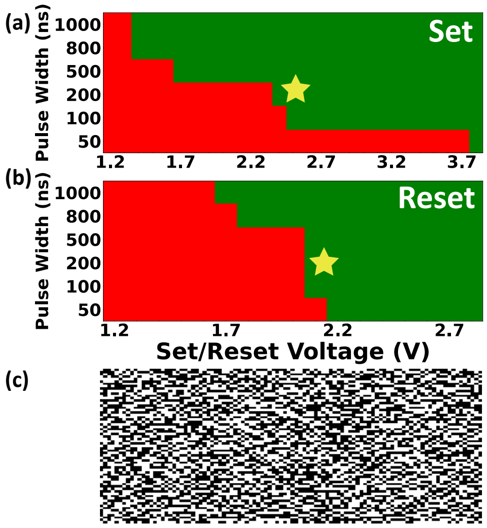

Figure 5: Measured resistance switching time by using different set/rest voltage on RRAM device.

As described above, the fabricated multifunctional 1T1R RRAM array can perform MAC operation, CAM addressing, and random number generating functions for serving as an Ising machine (Fig. 6). It achieves a low-cost way to implement a sparsity-aware RRAM module.

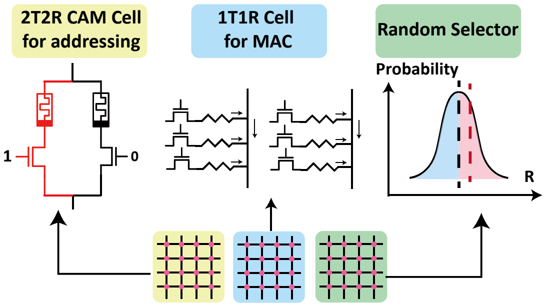

Figure 6: Multifunctional 1T1R Array could serve as CAM for addressing, MAC unit for partial sum and random selector for annealing.

CONCLUSION

By developing the multifunction of the RRAM array, a sparsity-aware IMC Ising machine is implemented using an RRAM chip with high time/energy efficiency. Compared with existing works, our Ising machine has great advantages in sparsity awareness and solving real-world NP-hard COPs.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (61925401, 92064004, 61927901, 92164302) and the 111 Project (B18001). Y.Y. acknowledges support from the Fok Ying-Tong Education Foundation and the Tencent Foundation through the XPLORER PRIZE.

REFERENCES

[1] Korte, B.H., et al., *Combinatorial optimization*. Vol. 1. 2011: Springer.

[2] Benlic, U. and J.-K. Hao, *Breakout local search for the max-cutproblem.* Engineering Applications of Artificial Intelligence, 2013. **26**(3): p. 1162-1173.

[3] Lin, S. and B.W. Kernighan, *An effective heuristic algorithm for the traveling-salesman problem.* Operations research, 1973. **21**(2): p. 498-516.

[4] Lucas, A., *Ising formulations of many NP problems.* Frontiers in physics, 2014. **2**: p. 5.

[5] Su, Y., H. Kim, and B. Kim, *CIM-spin: A scalable CMOS annealing processor with digital in-memory spin operators and register spins for combinatorial optimization problems.* IEEE Journal of Solid-State Circuits, 2022. **57**(7): p. 2263-2273.

[6] Yamamoto, K., et al. *7.3 STATICA: A 512-spin 0.25 M-weight full-digital annealing processor with a near-memory all-spin-updates-at-once architecture for combinatorial optimization with complete spin-spin interactions*. in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. 2020. IEEE.

[7] Johnson, M.W., et al., *Quantum annealing with manufactured spins.* Nature, 2011. **473**(7346): p. 194-198.