Hardware Acceleration of Simulated Annealing of Spin Glass by RRAM Crossbar Array

Jong Hoon Shin[1], YeonJoo Jeong[1], Mohammed A. Zidan[1], Qiwen Wang[1], and Wei D. Lu[1*]
[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, USA
Email: wluee@eecs.umich.edu

***Abstract*—**Simulated annealing (SA) was successfully implemented and accelerated by in-memory computing hardware/software package using RRAM crossbar arrays to solve a spin glass problem. Ta_2O_5-based RRAM array and stochastic Cu-based CBRAM devices were utilized for calculation of the Hamiltonian and decision of spin-flip events, respectively. A parallel spin-flip strategy was demonstrated to further accelerate the SA algorithm.

I. Introduction

The spin glass system is a representative combinational optimization problem (COP) which tries to find the globally optimal object in discrete space. Since COPs such as spin glass systems and the traveling salesmen problem are NP-hard, simulated annealing (SA), a metaheuristic algorithm that effectively search global optima, has been developed and widely used.[1] However, the convergence of SA may be slow because it involves compute-intensive operations within a massively connected interaction network and stochastic search rules that require random number generation (RNG) with an exponentially decaying probability distribution. Recently, there have been significant progress in RRAM-based acceleration of numerical computation such as partial differential equation and neural network using vector-matrix multiplication,[2,3] in-memory computing,[4] and stochastic computing using stochastic bit streams.[5,6] Inspired by the ability of RRAM devices for numerical computation, in this work, we utilized the ability for vector-matrix multiplication of Ta_2O_5 RRAM-based crossbar and stochastic switching observed in Cu-based CBRAM devices to accelerate an SA algorithm that solves a spin glass problem effectively.

II. Spin Glass Problem and Simulated Annealing

Finding the ground state of a two-dimensional (2D) spin glass, from randomly mixed states as shown in Fig.1a is a classical problem in COP. Although the interaction between two spins is simple such that the Hamiltonian is just a multiplication between neighboring spins weighted by the coupling strength, complex interactions between arbitrary spin pairs exist in the spin glass, as illustrated in Fig.1b and make the problem difficult to solve in polynomial time.[7] Fig.2a shows the flowchart of conventional SA that starts from initializing the spin configuration, followed by calculating the change of Hamiltonian ΔH_y due to flip of randomly selected y^{th} single spin, σ_y. The Hamiltonian of the spin glass is given as:

$$H = -J\sum_{<x,y>}\sigma_x\sigma_y = -\frac{1}{2}J\sum_{x,y}N_{xy}\sigma_x\sigma_y \qquad (1)$$

where J is the amplitude of the coupling strength, σ_x and σ_y are the x^{th} and y^{th} spin in the spin glass. $<x,y>$ in Eq (1) indicates that the spin multiplication needs to be conducted only for neighboring spins. The introduction of N_{xy}, a coupling strength (CS) matrix, makes the expression more concise. Elements in N_{xy} are '1' if σ_x and σ_y are neighbors of each other, and '0' for non-neighboring spins. If a spin flip decreases energy, e.g. inversion of σ_y leading to negative ΔH_y, SA accepts the change because it stabilizes the spin system. If on the other hand ΔH_y is positive, the spin flip will happen with a probability proportional to the Boltzmann factor ($P = exp(-\Delta H_y/kT)$) where T is absolute temperature. After a fixed number of attempted spin flips, the temperature T is decreased following a cooling schedule, and the process is repeated at the new temperature. The stochastic hill climbing provided by the Boltzmann factor enables the spin glass to escape from local optima as depicted in Fig.2b, and the escape probability decrease to zero as time increases and temperature cools down.

III. Simulated Annealing Accelerated by RRAM Array and Stochastic CBRAM

During SA, calculations of the inner products in ΔH_y and the probability generated by the RNG function in the Boltzmann factor make the process compute-intensive. To reduce the computational cost and speed up SA, inner products between the spin vector $\vec{\sigma}$ and neighboring spins, as determined by the CS matrix, can be directly obtained in an RRAM array storing the CS matrix N_{xy}, as shown in Fig.3a-b. For example, when the y^{th} spin attempts to be flipped, all x^{th} row ($\forall\sigma_x \in \vec{\sigma}$) in the RRAM array in Fig.3c are applied with a $V_x(=\sigma_x V_{read})$ pulse, and the output current I_y at the y^{th} column is proportional to $\sum_{x,y}N_{xy}\sigma_x\sigma_y$, producing the desired value of ΔH_y. As a result, the inner-products can be readily obtained from read operations through the RRAM array.

Since only nearest neighbor interactions are non-zero, the CS matrix can be very large but sparse. The large CS matrix can be effectively mapped into smaller RRAM arrays where only the non-zero portions are stored, as illustrated in Fig.4,5. Here a 9×9 2D spin glass was chosen as an example. The 81×81 CS matrix of the spin glass represents all-to-all connection, and can be divided into three groups (top-edge row, mid rows, and bottom-edge row), representing the coupling strength of a spin in the top (middle, or bottom) row with its neighbors. The

groups are 9 column wide (corresponding to the 9 spins in each row), and can be further divided into sub-groups of 3 spins (3 columns), for spins at the left-edge, middle columns, and right-edge, producing the patterns shown in Fig. 5. All the possible (non-zero) sub-matrix patterns can then be stored in a three-column RRAM array (11×3), as shown in Fig.5d. Experimentally, the 11×3 RRAM array was fabricated with a Pd/Ta/Ta_2O_5/Pd cell structure. The RRAM crossbar array is then wire-bonded and connected to a custom test board as shown in Fig.6.

Reliable switching characteristics and tight forming, set and reset voltage distribution can be obtained from all devices in the RRAM array (Fig. 7a,b). The cell-to-cell current variations shown in Fig.7c can be significantly improved to be lower than 1% using a write-verify method, as shown in Fig. 7d, enabling robust dot product operations to obtain ΔH_y .[8] The hill climbing probability was also obtained through hardware by using stochastic switching effects in a Cu-based CBRAM, as shown in Fig.8. The CBRAM device shows stochastic switching behavior at low programming voltage, with a switching probability $P(\Delta t) = 1 - \exp(\Delta t/\tau)$ for programming pulse width Δt, where τ is a time constant dependent on the voltage amplitude. A Cu/ALD Al_2O_3/Pd CBRAM structure is used in this experimental implementation, with $\tau = 24.9\text{ms}$ for transition from HRS to LRS. After applying a single SET pulse, the probability of the device staying at HRS then follows the exponential decaying function $\exp(-\Delta t/\tau)$, which follows the Boltzmann factor required for SA, after converting ΔH_y to $\Delta t = \tau(\Delta H_y/kT(t))$.

IV. Experimental Demonstration of RRAM-based Simulated Annealing

The flow chart of implementing SA to simulate a spin glass is shown in Fig.9. Starting from the initial spin configuration, a spin (i^{th} row and j^{th} column in the spin glass) is randomly selected for flip-trial. The spin vector is converted as input pulse vector based on its location and applied to the 11×3 Ta_2O_5 RRAM array. After the current measurement from the selected column I_y, the sign of I_y is compared with σ_y. The flip-event of σ_y is accepted if the signs match (corresponding to negative ΔH_y). If the signs of I_y and σ_y do not match, the flip-event is only accepted if a single SET pulse on a the CBRAM does not change its original HRS state, following discussions above. The data flow is illustrated in Fig.10.

A 15×15 2D ferromagnetic spin glass was tested to prove the concept of RRAM-based SA process. Fig.11 shows one test case with a fixed spin edge condition, where all the edge spins are fixed at the 'up'(+1) state and the rest of the spins are initialized to 'down'(-1) state at time = 0. Because the edge spins are always fixed, the only possible ground state of this problem is "all-up" configuration. The SA parameters such as *J, T(t)*, and N_T for the experiment are 1.0, $5/\sqrt[3]{t+1}$, and 100, respectively. As time flows, the initially down-spins get affected by the edge spin states due to ferromagnetic interaction that favors spins with same orientations. Note some of the down-spins surrounded by other down spins are also flipped to up-spin (e.g. at time=5), although this event increases the total *E*. This is an example of hill climbing phenomenon which can speed up the optimization process by escaping from the local optima, as discussed in SA. The ground state is achieved at ~ time=200. Other cases with multiple ground states, i.e. initially random configurations without any fixed edges, were also tested using the RRAM-based SA, as shown in Fig.12. Due to the existence of two possible ground states with "all-up" and "all-down" spin configurations, the same initial condition can evolve to opposite results, as verified by the experiments. Note that the two solutions also show similar proportions of majority spin during the evolutions (e.g. at time=150), since the SA strategy leads to similar dynamic progress towards the respective ground state. Comparison between the experimental RRAM-based SA results and software results verifies the *E* and magnetization (*M*) of both cases show similar dynamics that converge to global optima near time = 200, further proving the successful experimental implementation of RRAM-based SA.

V. Parallel Spin-flip Strategy using Memritive Simulated Annealing

To further accelerate the RRAM-based SA, it is possible to flip multiple non-neighboring spins together simultaneously to take advantage of the parallel vector-matrix multiplication (vs. vector-vector inner product) offered by RRAM arrays, as illustrated in Fig 14. The flipped spins have to be non-neighboring to not affect the energy calculations compared with consecutive spin flips. The parallel spin-flip strategy was also implemented in the RRAM-based hardware. Comparisons of the experimental results obtained from the conventional single spin-flip and the parallel double spin-flip schemes are shown in Fig. 15, for the fixed edge test case. The *E* and *M* from double spin-flip scheme (red) show faster convergence than the single spin-flip scheme (blue). The single spin-flip scheme even fell into a local minimum near time=100 for a while before finally escaping, while the double spin-flip method already reached its ground state. Since the double spin flip should be equivalent to two consecutive spin flips (at the same temperature), the results are compared with another experiment where 2x iterations (i.e. $2N_T$=200) are attempted at each time step using the single spin-flip scheme (black curves). This approach indeed produced results similar to those obtained from the double spin-flip experiments, and suggested possibility of further acceleration of SA with an N spin-flip scheme that can be calculated simultaneously in RRAM-based array.

Acknowledgment

This work was supported in part by NSF through grant CCF-1617315

References

[1] Kirkpatrick, S., et. al., Science 220 (1983): 4598, 671-680
[2] Zidan, M. A., et. al., Nature Electronics 1.7 (2018): 411-420
[3] Chang, C., et al., 2017 IEDM, San Francisco, CA (2017):11.6.1-11.6.4
[4] W. Chen et al., 2017 IEEE IEDM, San Francisco, CA (2017): 28.2.1-28.2.4
[5] Gaba, S., et. al., Nanoscale 5.13 (2013): 5872-5878
[6] Knag, P., et. al., IEEE Trans. Nanotechnology 13.2 : 283–293
[7] F. Barahona, J. Phys. A: Math. Gen., 15 (1982) : 3241-3253
[8] Zidan, M. A., et. al., IEEE Trans. Multi-Scale Comput. Syst. (2017)

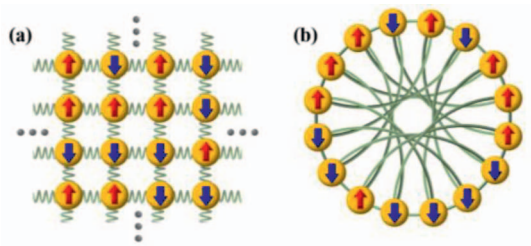

Fig.1. A 2D spin glass and the spin interactions represented by (a) connections to neighboring spins and (b) circular graph showing the complex couplings.

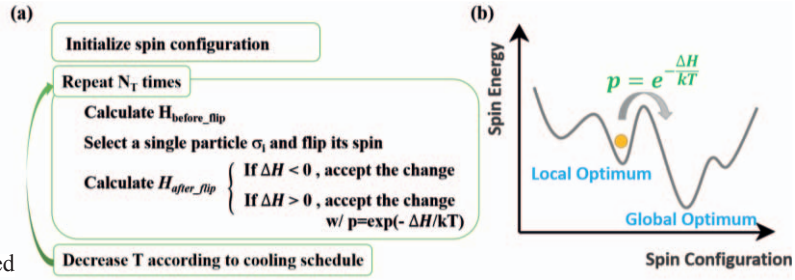

Fig. 2. (a) Flow chart of the SA algorithm. (b) Schematic showing finite spin flip probability even for positive $\Delta \mathbf{H}$ can help the system escape from local optima.

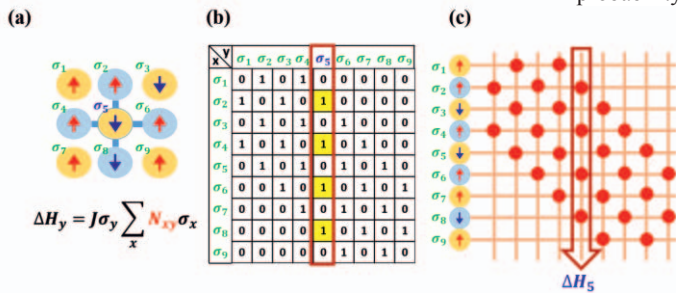

Fig. 3. (a) $\Delta \boldsymbol{H_y}$ due to the change of $\boldsymbol{\sigma_y}$ surrounded by its neighbor spins. (b) CS matrix where the 5th column represents interaction between 5th spin and all the other spins (c) Schematic of inner product between the 5th CS column vector and spin vector $\vec{\boldsymbol{\sigma}}$ conducted by RRAM array.

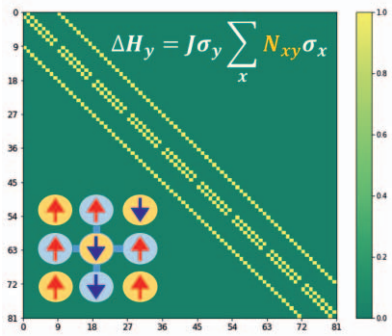

Fig. 4. $\mathbf{81 \times 81}$CS matrix of a $\mathbf{9 \times 9}$ 2D spin array. The large but sparse CS matrix can be sliced to fit into a smaller RRAM array.

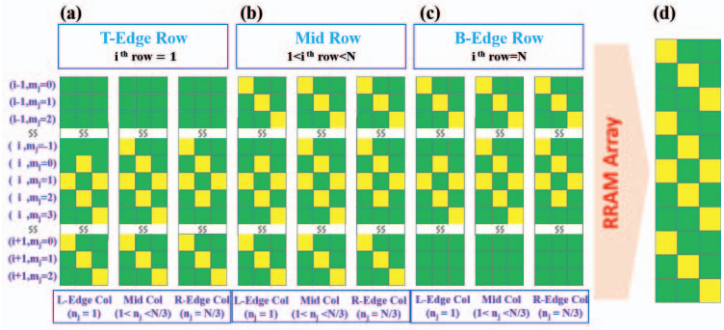

Fig. 5. 9 sub-patterns with three columns each from the $\mathbf{81 \times 81}$ CS matrix, depending on the position of the spin in the 2D spin glass. (a) Top-Edge Row case, (b) Mid Row case, and (c) Bottom-Edge Row case. (d) All the non-zero and unique patterns in (a-c) can be stored in a single $\mathbf{11 \times 3}$ RRAM array.

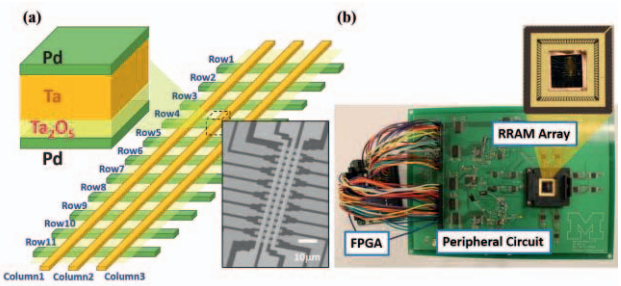

Fig. 6. (a) Schematic of the Ta_2O_5-based RRAM cell and array structure. SEM image of the RRAM crossbar array. (b) Test board comprised of FPGA, peripheral circuit, and the RRAM array chip for experimental implementation of simulated annealing.

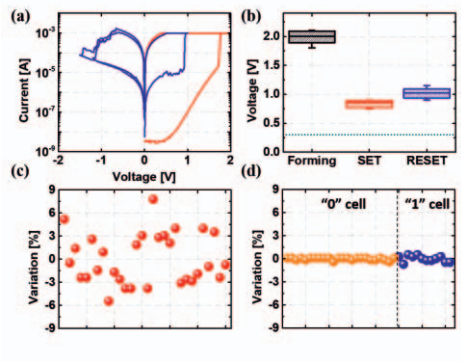

Fig. 7. (a) I-V curves showing the forming (red) and subsequent switching (blue) processes. (b) Distribution of $V_{Forming}$, V_{SET}, V_{Reset} of the 33 cells in the RRAM array. (c-d) Variation of device current without (c) and with (d) write-verify pulse method.

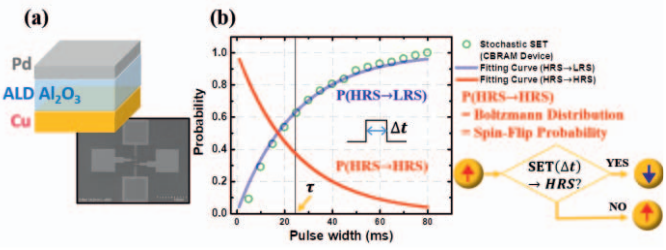

Fig. 8. (a) Structure and SEM image of Cu-based CBRAM devices. (b) Experimentally measured probability of HRS→LRS switching (blue). The Boltzmann factor (red) can be obtained by the probability of the device staying at HRS after applying a single SET pulse with pulse width $\Delta \boldsymbol{t}$.

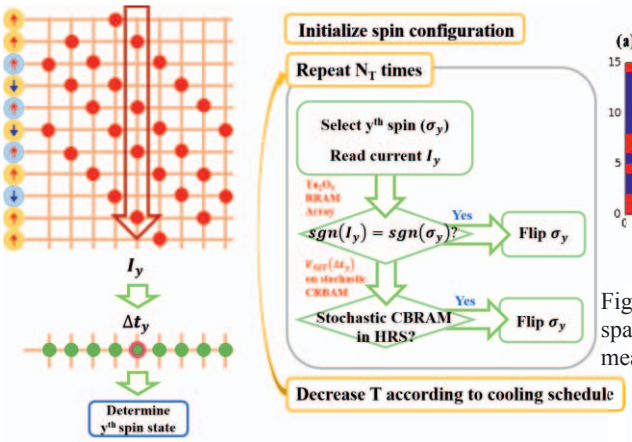

Fig. 9. Flowchart of implementing the SA algorithm using RRAM array for the 2D spin glass problem.

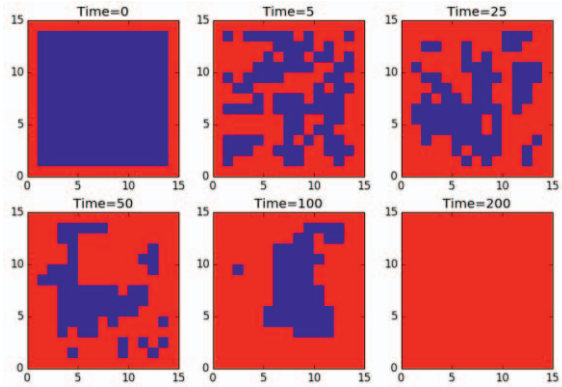

Fig. 11. Evolution of the spin configuration at different time steps for the fixed spin-edge case. Data obtained experimentally from the RRAM array-based hardware system.

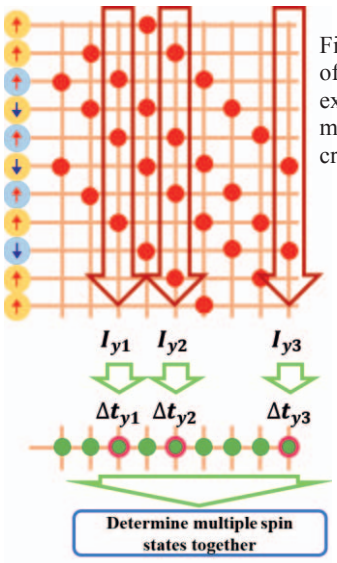

Fig. 14. Schematic illustration of multi-spin flip method that exploits parallel vector-matrix multiplications in RRAM crossbar array.

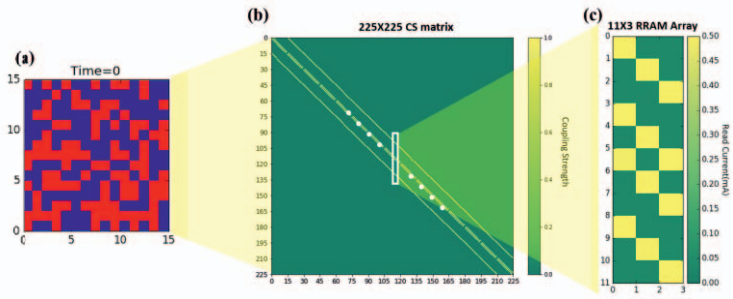

Fig. 10. (a) Randomly initialized $\mathbf{15 \times 15}$ spin array (with 225 spins). (b) The sparse $\mathbf{225 \times 225}$ CS matrix. (c) Coupling strength patterns stored and measured from the RRAM array used in the experimental setup.

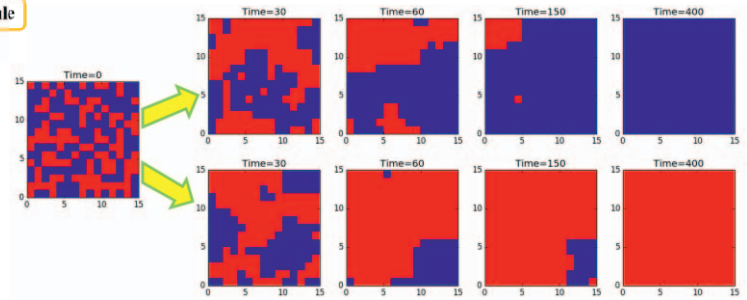

Fig. 12. Time-dependent evolution of the spin glass system solved by the RRAM hardware, for random initial states with no fixed spins. Two ground states with global energy minima, 'all-up' state and 'all-down' states, can be generated from the same initial state in different runs.

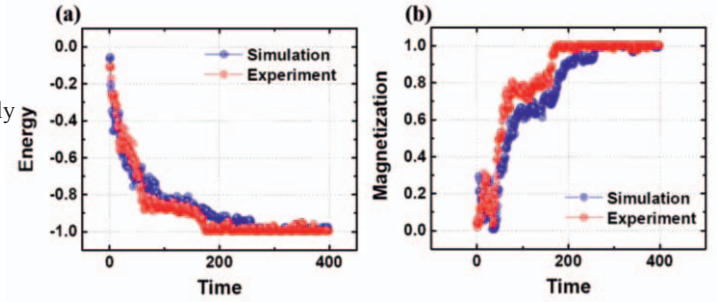

Fig. 13. (a) Average energy and (b) magnetization as a function of cooling schedule. Conventional software version of SA (red) and experimental SA results obtained from the RRAM array (blue) are compared.

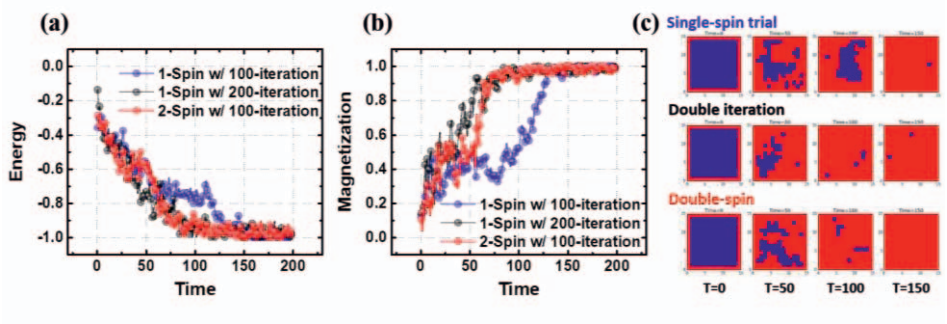

Fig. 15. Comparison of (a) energy, (b) magnetization, and (c) spin configuration snap shots, for results obtained using the single-spin method with 100 iterations per time step (blue), single-spin method with 200 iterations per time step (black), and double-spin method with 100 iterations per time step (red). All results are obtained from the RRAM hardware setup.