Annealing Accelerator for Ising Spin Systems based on In-memory Complementary 2D FETs

Amritanand Sebastian
Pennsylvania State University https://orcid.org/0000-0003-4558-0013
Sarbashis Das
Pennsylvania State University https://orcid.org/0000-0003-4553-5693
Saptarshi Das (sud70@psu.edu)
Pennsylvania State University https://orcid.org/0000-0002-0188-945X

Article

Keywords:

Posted Date: April 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-331394/v1

Annealing Accelerator for Ising Spin Systems based on In-memory Complementary 2D FETs

Amritanand Sebastian[1], Sarbashis Das[2] & Saptarshi Das[1,3,4]*

[1]*Deparment of Engineering Science and Mechanics, Penn State University, University Park, PA 16802*

[2]*Department of Electrical Engineering, Penn State University, University Park, PA 16802*

[3]*Department of Materials Science and Engineering, Penn State University, University Park, PA 16802*

[4]*Materials Research Institute, Pennsylvania State University, University Park, PA 16802*

Abstract

Metaheuristic algorithms such as simulated annealing (SA) has been implemented for optimization in combinatorial problems, especially for discreet problems. SA employs a stochastic search, where high-energy transitions ("hill-climbing") are allowed with a temperature-dependent probability to escape local optima. Ising spin glass systems have properties such as spin disorder and "frustration" and provide a discreet combinatorial problem with high number of metastable states and ground-state degeneracy. In this work, we exploit subthreshold Boltzmann transport in complementary two-dimensional (2D) field effect transistors (*p*-type WSe_2 and *n*-type MoS_2) integrated with analog, non-volatile, and programmable floating-gate memory stack to develop in-memory computing primitives necessary for energy and area efficient hardware acceleration of SA for the Ising spin systems. We experimentally demonstrate > 800X search acceleration for 4×4 ferromagnetic, antiferromagnetic, and a spin glass system using SA compared to an exhaustive search using brute force trial at miniscule total energy expenditure of ~120 nJ. Our hardware realistic numerical simulations further highlight the astounding benefits of SA in accelerating the search for larger spin lattices.

Introduction

Combinatorial optimization problems represent a set of problems where finding the best solution using an exhaustive search is often unfeasible. Interestingly, such problems appear in applications ranging from supply-chain management, airline scheduling, industrial resource allocation to artificial intelligence, applied mathematics, and theoretical computer science. The travelling salesman problem (TSP) is a representative optimization problem where given a number of cities, the shortest route connecting all the cities must be found [1]. In TSP the time complexity is in the order of $n!$, where n is the number of cities. The optimal solution can be found for a small n by an exhaustive search using the brute force trial (BFT) where every combination is evaluated. However, BFT becomes impractical for large n. For example, 40 cities would require 10^{49} trials! While the BFT fails, various metaheuristic algorithms such as simulated annealing (SA) [2], genetic algorithm [3], and ant colony optimization [4] have been utilized to obtain an approximate solution which is very close to the actual solution for TSP and other optimization problems. Among these, SA is an excellent optimization technique in discrete problems where multiple local minima exist. SA provides a simple framework which can be implemented on systems with arbitrary energy landscapes, and it statistically guarantees an optimal solution. SA has hence been employed to solve optimization problems in a wide variety of domains such as circuit design [5], data analysis [6], imaging [7], neural networks [8], geophysics [9], finance [10], and Ising model of magnetism [11].

SA draws inspiration from physical annealing where a material is heated above its recrystallization temperature to allow atoms to rearrange and is slowly cooled down to improve its crystallinity and reach a low energy state. SA is an optimization algorithm where the free energy (H) of a system

is minimized by employing a stochastic search. It is similar to other optimization methods such as gradient descent [12] where transitions lowering H are accepted. However, unlike the gradient descent method, it allows transitions increasing H ("hill-climbing") determined by the annealing temperature [13-16]. This "hill-climbing" feature of SA makes it highly attractive for systems with multiple local minima in their energy landscape (Fig 1a). The cost associated with a state transition leading to ΔH change in the free energy of the system is evaluated using the Boltzmann factor and accepted if it falls below a predefined threshold, say P, following Eq. 1.

$$\exp\left(\frac{\Delta H}{k_B T}\right) < P \quad [1]$$

Here, k_B is the Boltzmann constant, and T is the temperature. This particular SA approach is also referred to as the threshold accepting method and is widely used for its simpler structure [17]. The basic principle of SA is illustrated in Fig. 1a. To minimize the free energy function $H(x)$ corresponding to the argument set $x = \{x_1, x_2, x_3, \ldots x_N\}$, where N is a *large number*, a random x_i is initialized. At each iteration, a new random point (x_k) is chosen and ΔH associated with the transition is evaluated. If $\Delta H < 0$, the transition is automatically accepted, whereas if $\Delta H > 0$, the transition is accepted i.e. "hill-climbing" is allowed following Eq 1. SA algorithm starts at a high T, where significant "hill-climbing" is allowed and T is progressively reduced following an annealing schedule. At sufficient lower T the system finds the global minima or arrives close to the same.

Most of the work on optimization algorithms like SA is done in software [11, 13-16], with a few hardware demonstrations [18, 19] based on von-Neumann architecture rendering them area and energy inefficient owing to the physical separation of memory and compute. Non-von Neumann hardware accelerators such as graphics processing units (GPUs) [20, 21], and field-programmable

gate arrays (FPGAs) [22] have become increasingly popular in the later years offering speedup, energy efficiency and smaller physical footprint [23]. Quantum annealing, a variant of SA, where the "hill-climbing" is achieved by quantum mechanical tunneling has been implemented in D-wave processor, a quantum computer [24]. More recently, memristive crossbar arrays have been used to implement standalone SA [25] and SA in Hopfield neural networks for various optimization problems [26-28]. While memristors offer in-memory computing, they have limitations owing to sneak paths and narrow dynamic ranges limiting the number of conductance states, high power consumption, and metal migration in the ON state due to high currents. To improve their performance, memristors are often integrated with an access transistor in a 1T1R structure and also require peripheral circuits based on silicon complementary metal oxide semiconductor (CMOS) technology. Finally, annealing schedule is usually realized through physical cooling of the system which significantly adds to the energy overhead.

In this work, we explore SA using in-memory complementary field-effect transistors (FETs) based on ultra-thin body two-dimensional (2D) semiconductors, i.e. *p*-type WSe_2 and *n*-type MoS_2 FETs. First, unlike quantum computers operating at cryogenic temperatures, our demonstration is based on room temperature, and second, we exploit analog subthreshold conduction and analog programmability to design unique computing primitives and annealing schedule which achieve better energy and area efficiency compared to large memristive cross-bar arrays. Note that earlier works have already demonstrated the scalability and technological viability of 2D materials [29-31]. Additionally, in-memory and near-memory technologies based on 2D materials have been used recently for various computing tasks overcoming the von-Neumann bottleneck [32-34]. This work further advances the development of 2D FET based in-memory computing platform. To the

best of our knowledge this is the first demonstration of hardware acceleration of SA for the Ising spin glass systems using emerging materials and devices.

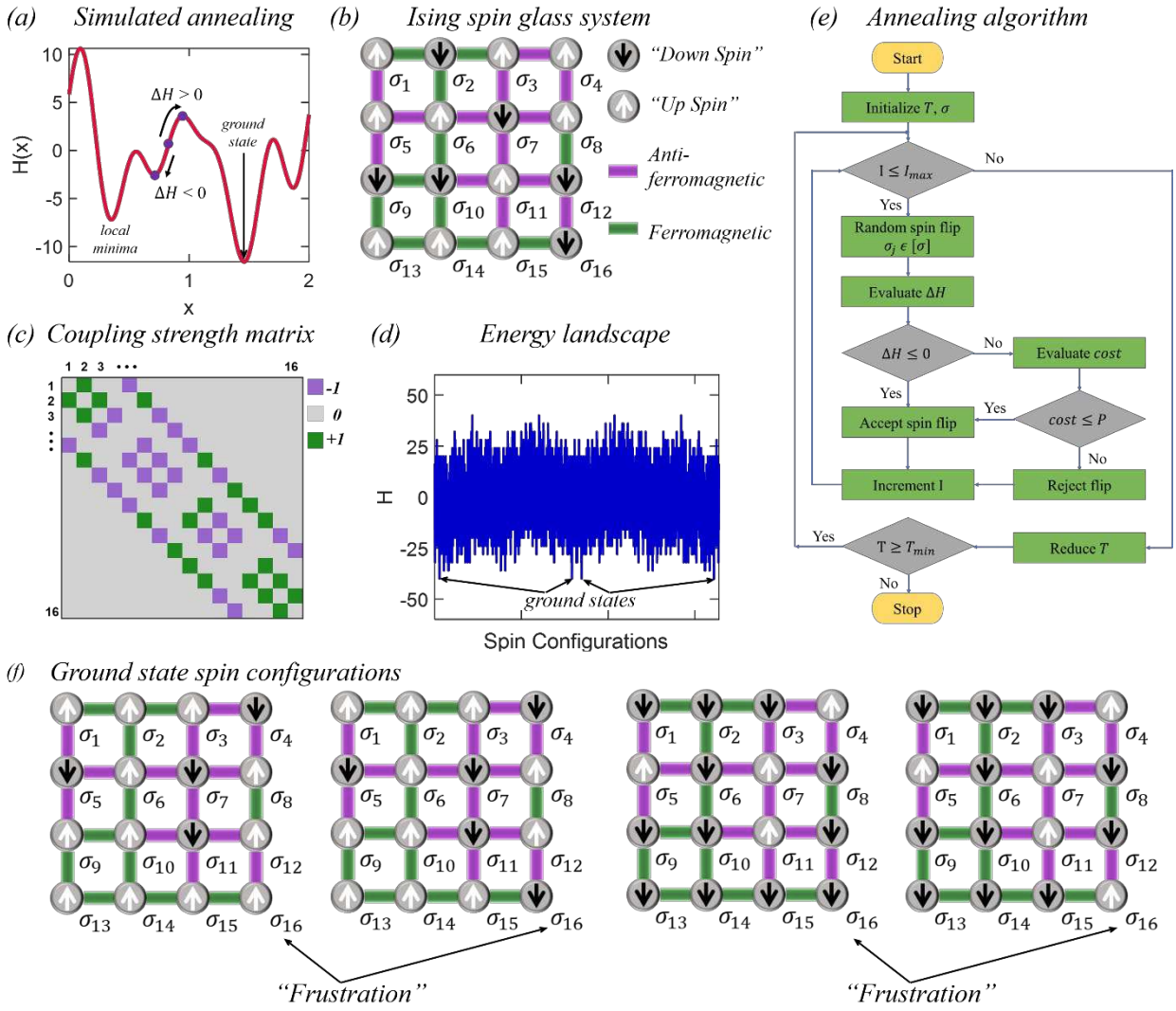

Figure 1. Simulated annealing (SA) and Ising spin glass system. *a) Illustration of the basic principle of SA used for finding the ground state(s) or lowest energy state(s) of a system in a large search space with multiple local minima. Unlike many other optimization methods, SA accepts transitions increasing H ("hill-climbing") based on the annealing temperature (T). b) A 4×4 Ising spin glass system with 16 randomly oriented spins in up (white) or down (black) direction. The neighboring spin interactions can be either ferromagnetic (green) or antiferromagnetic (purple). c) The corresponding coupling strength matrix* $[CS]$*. d) Free energy (H) corresponding to the* 2^{16} *possible states for the spin glass system in (b). The energy landscape shows multiple local minima with ground state degeneracy of 4. The number of brute force trials increases exponentially as the size of the spin lattice increases, qualifying the spin glass system as a challenging combinatorial optimization problem, where SA can accelerate the search. e) Flowchart of the SA algorithm for a spin glass system. A predetermined cooling schedule is used for* T*. At each* T*, the algorithm runs for a predetermined number of iterations (*I_{max}*). During each iteration, a random spin is flipped and* ΔH *associated with the spin flip is evaluated. For* $\Delta H \leq 0$*, the spin flip is always accepted. For* $\Delta H > 0$*, the spin flip is accepted if the cost is less than a predetermined value,* P*. For a sufficiently large number of iterations, the system converges to one of the f) 4 ground states for the spin glass systems in (b).*

Spin glass system

Since the early years of SA, the Ising spin glass problem has been extensively studied since it offers combinatorically huge number of outcomes with multiple statistically equivalent ground states. Similar to the disordered nature of atomic positions in glass, in a spin glass system, the magnetic spins are disordered [35, 36]. The spatial disorder in a glass is set by quenching, where its atoms are frozen in a disordered state. Similarly, spin glass is a system of quenched magnets with disorder due to frozen spin states and interactions. Spin glasses demonstrate interesting properties such as frustration. At low temperatures, a spin glass system has roughly equal number of ferromagnetic and antiferromagnetic interactions or bonds which are frozen. However, spins can be flipped to obtain a low energy state which satisfies the frozen spin interactions. This leads to "frustration" in spins trying to settle between competing ferromagnetic and antiferromagnetic interactions [37]. A large number of metastable states with high degeneracy corresponding to different spin orientations are observed in the spin glass systems. This leads to a non-monotonic energy landscape with multiple local minima. The transition from a metastable state to the lowest energy state requires specific spin flips and represents an optimization problem where high energy transitions should be allowed to escape from local minima; hence, spin glass is an ideal system to implement SA.

A $K \times K$ square spin lattice with $K = 4$ is shown in Fig. 1b. The nature of spin interactions i.e., ferromagnetic (green) *versus* antiferromagnetic (purple) are also shown. The corresponding spin vector $[\sigma]$ consisting of $K^2 = 16$ spins is given by $[\sigma] = [\sigma_1, \sigma_2, \ldots \sigma_{K^2}]$. In the Ising spin model, +1 represents "up" spin and -1 represents "down" spin, respectively. The Hamiltonian representing

the free energy of the spin glass system is given by the zero-field Edward-Anderson (EA) model described in Eq. 2 [38].

$$H = -\sum_{<i,j>} \sigma_i J_{ij} \sigma_j \; = -\frac{1}{2}\sum_{i,j} \sigma_i J_{ij} N_{ij} \sigma_j = -\frac{1}{2}\sum_{i,j} \sigma_i CS_{ij} \sigma_j = -\frac{1}{2}[\sigma][CS][\sigma]^T \quad [2]$$

Here, σ_i and σ_j are the i^{th} and j^{th} elements of σ. J_{ij} denotes the nature of interaction between σ_i and σ_j, i.e., J_{ij} = +1 for ferromagnetic interaction and J_{ij} = -1 for antiferromagnetic interaction. The operator <> denotes that only the nearest neighbor interactions are taken into account, i.e., N_{ij} = +1, if σ_i and σ_j are immediate neighbors and N_{ij} = 0, otherwise [25]. Note that both $[J]$ and $[N]$ matrices are sparse matrices of size $K^2 \times K^2$. These two matrices are combined into a single $K^2 \times K^2$ matrix, referred to as the coupling strength matrix, $[CS]$ (Fig. 1c). The light squares represent +1, the dark squares represent -1, and the gray squares represent 0. Note that the $K \times K$ spin glass system with K^2 number of spins, each with two possible orientations ("up"/"down"), has 2^{K^2} possible states. The free energy (H) corresponding to each of these 2^{K^2} states are shown in Fig. 1d for the spin glass system shown in Fig. 1b. Clearly, the energy landscape is non-monotonic with multiple local minima and degenerate global minima or ground states. Furthermore, with increasing size of the spin lattice (K) the search for spin configurations that satisfy all interactions become infeasible using BFTs qualifying the spin glass system as a challenging combinatorial optimization problem, where SA can speed-up the search process.

Simulated annealing for spin glass system

The SA algorithm to find the ground state of a spin glass system is shown in Fig. 1e. To find the orientations of K^2 = 16 spins that result in minimum free energy, a random spin in the $K \times K$ spin lattice is flipped and the corresponding ΔH is evaluated using Eq. 2. If ΔH < 0, the free energy of

the system is lowered and hence the spin flip is accepted. However, if $\Delta H > 0$, the spin flip increases the free energy of the system. In this case the spin flip can still be accepted ("hill-climbing") if the associated cost obtained from Eq. 1 is lower than a predetermined value, P. If both conditions are not satisfied, then the spin flip is rejected. At a given T, a predetermined number of iterations (I_{max}) are performed to traverse the energy landscape of the spin system. This process is then repeated by progressively cooling the system i.e., by reducing T following a predefined cooling schedule. It is found that within a reasonable number of iterations and at a sufficiently lower temperature the system converges to its ground state (Fig. 1f). Note that the spin glass systems typically exhibit ground-state degeneracy. See ***Supplementary Fig. 1a-d*** for the ground states and the free energy landscapes of 4×4 ferromagnetic and antiferromagnetic spin systems, respectively. In a ferromagnetic system, the ground state degeneracy is 2 since all interactions are satisfied if all spins are pointing "up" or "down" (***Supplementary Fig. 1b***). An antiferromagnetic system also possesses ground state degeneracy of 2 since all interactions are satisfied if adjacent spins are oriented in opposite directions (***Supplementary Fig. 1d***). However, in spin glass systems, the ground-state degeneracy is increased as the ground state is unable to satisfy all interactions due to the phenomenon of "frustration". For example, "up" and "down" configurations are equally valid for σ_{16} for the spin glass system to be in its ground state (Fig. 1f). ***Supplementary Video 1*** shows SA for 4×4 ferromagnetic, antiferromagnetic, and "frustrated" spin glass systems.

Hardware realization of simulated annealing

Hardware implementation of SA requires: 1) a random number generator for random spin flip, 2) a computational unit to calculate the change in free energy (ΔH) of the spin system associated with

the random spin flip following Eq. 2, 3) a computational unit to determine the cost of "hill-climbing" if $\Delta H > 0$ to accept or reject the spin flip following Eq. 1, and finally 4) a hardware mechanism equivalent to the annealing schedule or cooling in metallurgy. While we use software code to generate the random numbers, all other computational units including the mechanism for annealing are realized in hardware. For example, a multiplier module is designed using complementary 2D FETs along with a resistor and a capacitor module to evaluate ΔH. Similarly, co-location of analog memory and analog computing enabled by non-volatile and programmable floating-gate MoS_2 FETs is used to determine the cost of "hill-climbing" and achieve an annealing schedule equivalent to changing the T .

The schematic and optical image of a back-gated *p*-type WSe_2 FET with 50 nm Al_2O_3 as the gate dielectric is shown in Fig. 2a-b, respectively. Undoped WSe_2 demonstrates ambipolar transport with both electron (*n*-type) and hole (*p*-type) conduction owing to the pinning of metal Fermi-level near the middle of the WSe_2 bandgap [39-41]. However, for the design of the multiplier module, unipolar *p*-type WSe_2 is preferred. Hence, WSe_2 is doped *p*-type using surface charge transfer doping with sub-stochiometric WO_{3-x} [40]. Doping is achieved on a multilayered WSe_2 flake by converting its top layers to WO_{3-x} by exposure to mild O_2 plasma as discussed in our earlier reports [40, 42]. As shown in Fig. 2a before O_2 plasma exposure, the flakes are patterned to obtain a *p-i-p* structure i.e., *p*-doped source and drain extension regions with the middle region left intrinsic [40] (see ***Methods*** section for details on *p*-type WSe_2 FET fabrication). The length of the intrinsic region is designed to be 500 nm. The corresponding transfer characteristics, i.e., drain current (I_{DS}) *versus* back-gate voltage (V_{BG}) for different drain-to-source voltage (V_{DS}) and output characteristics, i.e., I_{DS} *versus* V_{DS} for different V_{BG} of the WSe_2 FET are shown in in Fig. 2c-d,

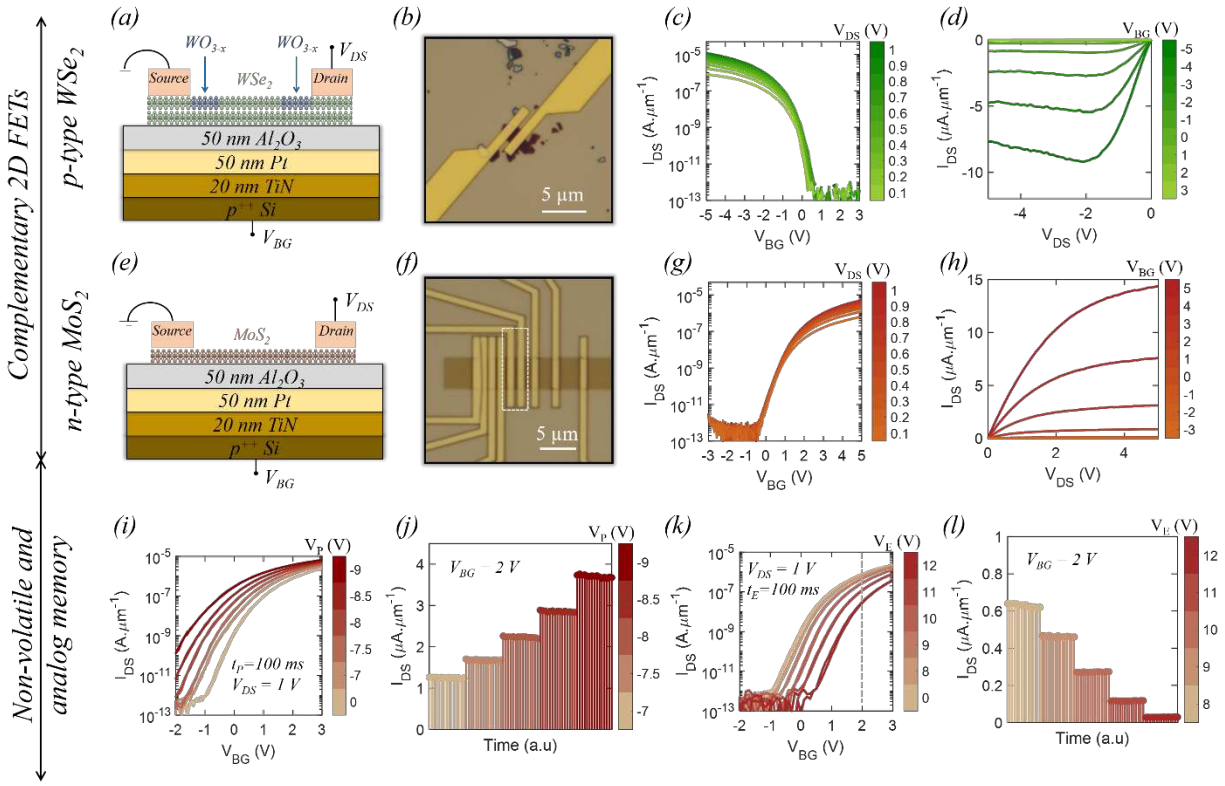

Figure 2. Analog in-memory complementary 2D field effect transistors (FETs). *a) Schematic and b) optical image of a back-gated p-type WSe_2 FET. The channel is selectively exposed to mild O_2 plasma to form the p-i-p structure with the length of the intrinsic region as 500 nm. Corresponding c) transfer and d) output characteristics for WSe_2 FET. e) Schematic and f) optical image of back-gated n-type MoS_2 FET with channel length as 500 nm. g) Transfer and h) output characteristics for MoS_2 FET. The p^{++}-Si/TiN/Pt/Al_2O_3 stack offers analog and non-volatile memory where the threshold voltage of the FETs can be adjusted by applying a programming pulse to the back gate. i) Transfer characteristics of post-programmed MoS_2 FET by applying negative programming voltages of different amplitudes for $t_P = 100$ ms. j) Retention characteristics, i.e., post-programmed I_{DS} versus time measured at $V_{BG} = 0$ V. i) Transfer and k) corresponding retention characteristics of post-erased MoS_2 FET after applying positive erase voltages of different amplitudes for $t_E = 100$ ms.*

respectively. Clearly, the WSe_2 FET demonstrates unipolar *p*-type conduction with current on-off ratio of ≈10^8, subthreshold slope, $SS = 200\ mV.dec^{-1}$, hole mobility, $\mu_P = 13\ cm^2\ V^{-1}s^{-1}$, and on-current, $I_{ON} = 11\ \mu A.\mu m^{-1}$ for an inversion carrier density of $4.4 \times 10^{12}\ cm^{-2}$.

The schematic and optical image of a back-gated MoS_2 FET with the same 50 nm Al_2O_3 gate dielectric and with a channel length of 500 nm is shown in Fig. 2a-b, respectively. We have used monolayer MoS_2 grown using metal organic chemical vapor deposition technique as described in our earlier reports [32, 43, 44] (see ***Methods*** section for details on *n*-type MoS_2 FET fabrication).

For MoS_2, the metal Fermi-level pins closer to the conduction band at the source/drain contact interfaces resulting in unipolar *n*-type conduction [41, 45]. The corresponding transfer characteristics and output characteristics are shown in Fig. 2c-d, respectively. For monolayer MoS_2 FET, we extracted current on-off ratio of $\approx 10^7$, subthreshold slope, $SS = 310\ mV.dec^{-1}$, electron mobility, $\mu_N = 15\ cm^2\ V^{-1}s^{-1}$, and $I_{ON} = 6.7\ \mu A.\mu m^{-1}$ for an inversion carrier density of $3.7 \times 10^{12}\ cm^{-2}$.

The unique stack of p^{++} Si/TiN/Pt/Al_2O_3 offers analog and non-volatile memory, i.e. the threshold voltage of the FETs can be adjusted by applying a programming voltage (V_P) to the back-gate electrode for a programming pulse time (t_P). The programmability is shown in Fig. 2e using the transfer characteristics of a MoS_2 FET at $V_{DS} = 1\ V$ after the application of negative programming voltages of different amplitudes for $t_P = 100\ ms$. Fig. 2f shows the retention characteristics, i.e. post-programmed I_{DS} *versus* time measured at $V_{BG} = 2\ V$ confirming non-volatile and analog programmability of the MoS_2 FET. Similarly, by applying positive erase voltages (V_E) for an erase pulse time, $t_E = 100\ ms$, the programmed states can be erased and the transfer characteristics can be shifted in the opposite direction as shown in Fig. 2g. The corresponding non-volatile retention characteristics are shown in Fig. 2h. Note that, programming (erase) operation can also be achieved by varying t_P (t_E) for a fixed V_P (V_E) as shown in ***Supplementary Fig. 2a and 2b***. The working principle of the analog and non-volatile back-gate memory stack has been described in detail in our earlier report [32].

The above demonstration of *p*-type WSe_2 FET, *n*-type MoS_2 FET, and the analog compute and analog non-volatile storage capability allow us to design the computational primitives necessary

for the hardware realization of SA. Note that ΔH due to random spin flip event can be computed from the difference in the free energy of the spin glass system before (H) and after (H') the spin flip. According to Eq. 2, evaluation of H and H' require multiplication of the spin vector $[\sigma]$ of size $1 \times K^2$ with the $[CS]$ matrix of size $K^2 \times K^2$ followed by multiplication with the transpose of the spin vector, i.e., $[\sigma]^T$ of size $K^2 \times 1$. However, there are several challenges: first, while vector matrix multiplication can be realized using cross-bar architectures, the $[CS]$ matrix contains negative elements which is difficult to realize using conductance states and requires additional circuitry (note that earlier demonstration of SA [25] using resistive random access memory only implemented ferroelectric interactions); and second, the size of the $[CS]$ matrix can become substantial even for relatively low values of K imposing heavy area and energy overhead for the computation. Instead, the computational load can be significantly reduced by acknowledging the fact that only one spin is allowed to randomly flip during each iteration of SA, simplifying the computation of ΔH following Eq. 3

$$\Delta H = H' - H = \left(-\frac{1}{2}\sum_{i,j} CS_{ij}\sigma_i\sigma_j{}'\right) - \left(-\frac{1}{2}\sum_{i,j} CS_{ij}\sigma_i\sigma_j\right) = \left[-\frac{1}{2}\Delta\sigma_j\right]\left[\sum_{i=1}^{K^2} CS_{ij}\sigma_i\right] \quad [3]$$

Here, we assume that the j^{th} spin is flipped. Note that the first term inside the square bracket, i.e., $\left[-\frac{1}{2}\Delta\sigma_j\right]$ is computationally equivalent to the initial spin state of σ_j. For example, if σ_j flips from +1(-1) to -1(+1), $\Delta\sigma_j$ = -2(+2) and hence $\left[-\frac{1}{2}\Delta\sigma_j\right]$ = +1(-1). Therefore, following Eq. 3, ΔH can be obtained just by summing the result of the dot product of the spin vector $[\sigma]$ with the j^{th} row vector of the $[CS]$ matrix, i.e., $[CS]_j$ and by multiplying the sum with σ_j. Fig. 3a-d show the circuit modules used for computing ΔH. The multiplication module (M1) in Fig. 3a comprises of a WSe_2 FET (T1) and MoS_2 FET (T2), connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, $V_{\text{in}-1}$ and $V_{\text{in}-2}$. V_{in-1} is applied to the

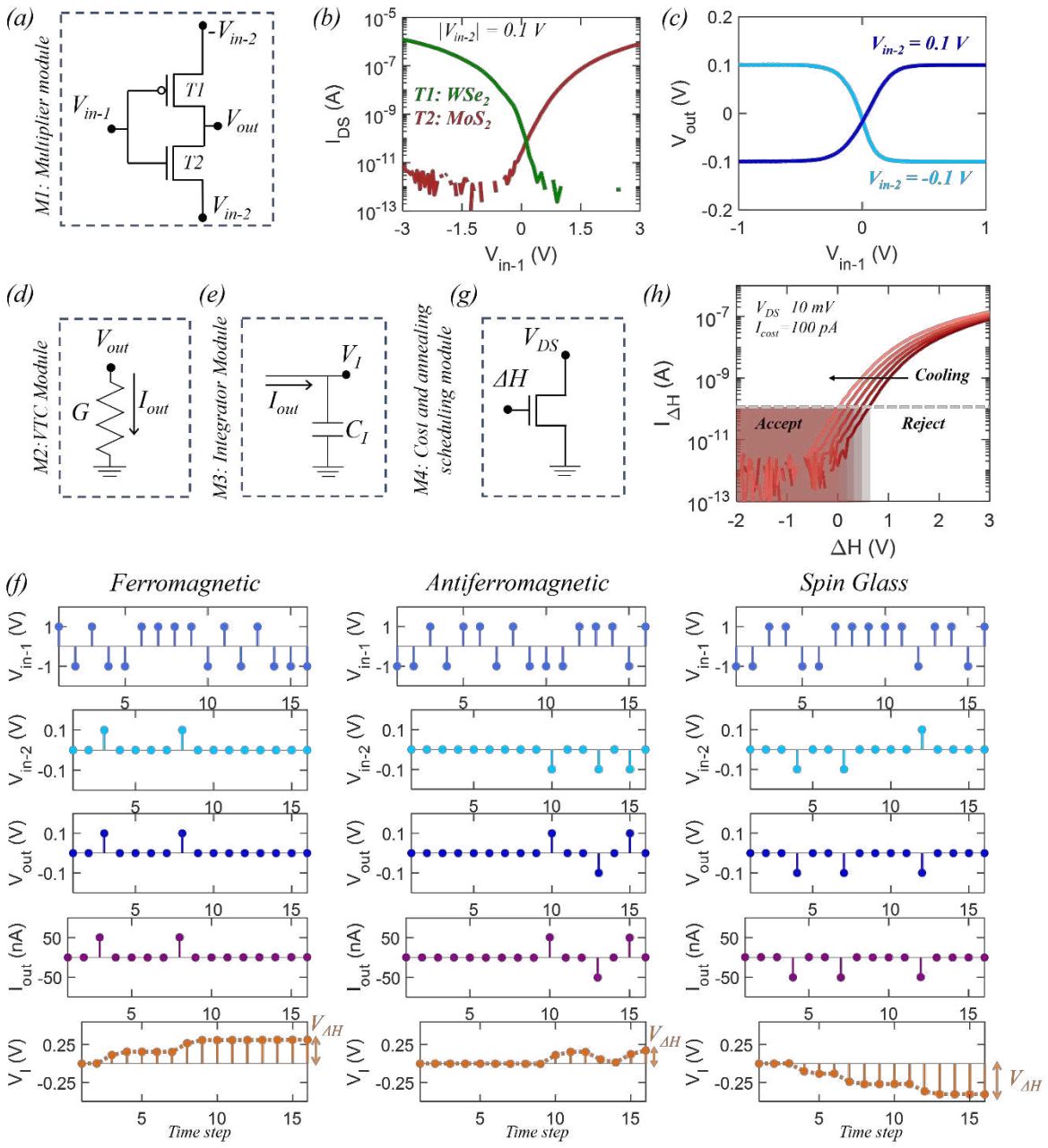

Figure 3. Circuit modules for hardware acceleration of SA. *a) The multiplier module (M1) has a p-type WSe_2 FET (T1) and an n-type MoS_2 FET (T2), connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, V_{in-1} and V_{in-2}. V_{in-1} is applied to the common-gate terminal, V_{in-2} is applied to the drain terminal of T1, and $-V_{in-2}$ is applied to the drain terminal of T2. b) Transfer characteristic of T1 and T2. c) Transfer characteristics of M1 i.e., output voltage, V_{out} versus V_{in-1} for V_{in-2} = ±0.1 V. Using M1 the product between the i^{th} elements of $[\sigma]$ and $[CS]_j$ is obtained at the i^{th} time step ($i\tau_p$) as $V_{out}(i\tau_p)$ by applying, $V_{in-1}(i\tau_p) = V_1\sigma_i$ and $V_{in-2}(i\tau_p) = V_2 CS_{ij}$. Note that $i = 1{:}1{:}K^2$, $K = 4$. We have used τ_p =60 ms, V_1 = 1 V, and V_2 = 0.1 V resulting in $V_{out}(i\tau_p) = 0.1\times CS_{ij}\sigma_i$. d) The voltage to current converter module (M2) transforms V_{out} from M1 into current, I_{out} following $I_{out} = GV_{out}$ with $G \approx 0.5$ μS. e) The integrator module (M3), a capacitor (C_I = 20 nF), sums I_{out} from M2 over K^2 time steps into voltage, $V_{\Delta H}$. f) V_{in-1}, V_{in-2}, $-V_{in-2}$, V_{out}, I_{out}, and the output from M3, i.e., V_I for representative ferromagnetic, antiferromagnetic, and a spin glass system during a given iteration of SA. $V_{\Delta H}$ and σ_j are multiplied to obtain ΔH. g) Schematic and h) transfer characteristics of a programmable MoS_2 FET used for evaluating the cost associated with the state transition as well as for realizing cooling schedule in hardware. The subthreshold conduction governed by Boltzmann statistics is exploited to evaluate the cost of "hill-climbing" by applying $V_{BG} = \Delta H$ and the spin flip is accepted if $I_{\Delta H} < I_{cost}$ = 100 pA. The cooling schedule is implemented by shifting the threshold voltage of the FET through back-gate programming.*

common-gate terminal, V_{in-2} is applied to the drain terminal of T1, and $-V_{in-2}$ is applied to the drain terminal of T2. Note that, T1 and T2 demonstrate dominant *p*-type and *n*-type conduction, respectively, as shown in Fig. 3e. The transfer characteristics of M1, i.e., output voltage, V_{out} *versus* V_{in-1} for V_{in-2} = ±0.1 V are shown in Fig. 3f. For V_{in-1} = 1 V, $V_{out} = V_{in-2}$, whereas for V_{in-1} = -1 V, $V_{out} = -V_{in-2}$. This is expected since for V_{in-1} = 1 V, the *n*-type MoS_2 FET (T2) is more conductive than the *p*-type WSe_2 FET (T1) allowing V_{out} to follow V_{in-2} and *vice versa* for V_{in-1} = -1 V (Fig. 3e). Using M1 the product between the i^{th} elements of $[\sigma]$ and $[CS]_j$ is obtained at the i^{th} time step ($i\tau_p$) as $V_{out}(i\tau_p)$ by applying, $V_{in-1}(i\tau_p) = V_1\sigma_i$ and $V_{in-2}(i\tau_p) = V_2 CS_{ij}$. We have used τ_p =60 ms, V_1 = 1 V, and V_2 = 0.1 V resulting in $V_{out}(i\tau_p)$ = 0.1×$CS_{ij}\sigma_i$. Next V_{out} is converted to I_{out} using a voltage to current converter module, M2 (Fig. 3b), comprising of a resistor, following $I_{out} = GV_{out}$ with $G \approx 0.5$ μS. Finally, the contribution due to all spins are summed over K^2 time steps using an integrator module, M3 (Fig. 3c), comprising of a capacitor ($C_I \approx 20$ nF) resulting in $V_{\Delta H}$ given by Eq. 4.

$$V_{\Delta H} = \frac{\tau_p}{C_I}\sum_{i=1}^{K^2} I_{out}(i) = \frac{\tau_p}{C_I} G V_2 \sum_{i=1}^{K^2} CS_{ij}\sigma_i \qquad [4]$$

Fig. 3g shows V_{in-1}, V_{in-2}, $-V_{in-2}$, V_{out}, I_{out}, and the output of the integrator, V_I for representative ferromagnetic, antiferromagnetic, and spin glass systems during a given iteration. Finally, $V_{\Delta H}$ and σ_j are multiplied to obtain ΔH.

Following the evaluation of ΔH, it must be determined if the spin flip is accepted. Traditionally, this is done using two separate steps: one to determine the sign of ΔH and another one to determine the cost of "hill-climbing" if ΔH is positive. In hardware, both steps are combined in M4 (Fig. 3d) by exploiting the subthreshold conduction in an FET where the carrier injection from the source

contact into the semiconducting channel is given by Boltzmann statistics. The cost of "hill-climbing" is, therefore, obtained by applying $V_{BG} = \Delta H$ and the spin flip is accepted if $I_{\Delta H} < I_{cost}$ = 100 pA (Fig. 3h) following Eq. 5.

$$I_{\Delta H} = I_0 \exp\left(-\frac{q\Delta H}{mk_BT}\right) \quad [5]$$

Here, q is the electronic charge, I_0 is the static leakage current, and m is the body factor, which determines the SS following Eq 6.

$$SS = \frac{mk_BT}{q} ln(10);\ \ m = \left(1 + \frac{C_S}{C_{OX}} + \frac{C_{IT}}{C_{OX}}\right) \quad [6]$$

Here, C_S is the semiconductor capacitance, C_{IT} is the interface trap capacitance, and C_{OX} is the oxide capacitance. For fully depleted ultra-thin-body FETs, $C_S = 0$. We extracted SS of 430 $mV.dec^{-1}$ and hence m of 7.1. Note that in Fig. 3h the transfer characteristics is represented as a plot of $I_{\Delta H}$ *versus* ΔH. Also note that all negative ΔH (low-energy transition) naturally leads to $I_{\Delta H} < I_{cost}$, whereas, positive ΔH up to $\Delta H_{\max}$ satisfies the "hill-climbing" criterion of $I_{\Delta H} < I_{cost}$.

Next, we implement cooling schedule in hardware. While temperature can be used to the change the current $I_{\Delta H}$ for same ΔH following Eq. 5, physically changing the temperature of a system requires excessive energy. Alternatively, $\Delta H_{\max}$ can be modulated by programming the MoS_2 FET in different states as shown in Fig. 3h. As the annealing temperature is lowered, the transfer characteristics is shifted towards the left. This ensures acceptance of more positive ΔH at higher temperature and no "hill climbing" at T = 0 K, i.e. $\Delta H_{\max}$ = 0 V. As shown in Fig. 3h, our annealing schedule involved five (5) different $\Delta H_{\max}$ = 0.6 V, 0.45 V, 0.3 V, 0.15 V, and 0 V analogous to temperature in metallurgical annealing with maximum number of iteration, $I_{\max}$ = 5, 15, 15, 30, and 30 at the respective "temperatures".

Fig. 4a-c, respectively, show the representative experimental demonstration of SA leading to the convergence of randomly initiated ferromagnetic, antiferromagnetic, and a spin glass system to their respective ground states. See ***Supplementary Figures 3-5*** and ***Supplementary Video 2-4*** for SA experiments performed on ferromagnetic, antiferromagnetic and a spin glass system initiated with 20 randomly oriented spin configurations. 11 ferromagnetic (55%), 9 antiferromagnetic (45%) and 8 spin glass systems (40%) converged and reached their respective ground states at the end of the total 95 iterations. Additionally, multiple systems are either 1 or 2 spin flips away from their ground state, and hence they are expected to converge for an increased number of iterations. We also observed "frustration" in the spin glass system. Remarkably, compared to an exhaustive search using BFT that requires a maximum of 2^{K^2} (= 65536 for $K = 4$) spin flips, SA accelerates the search requiring orders of magnitude lower spin flip events. We define acceleration as the ratio of maximum number of spin flips using exhaustive search to the maximum number of SA spin flips to reach the ground state. The highest acceleration for ferromagnetic, antiferromagnetic and the spin glass system were found to be ~1365X, ~1260X, and ~1310X, respectively, whereas, the average acceleration for the systems that converged to their ground states were ~850X, ~900X, and ~810X, respectively. Evolution of the free energy (H) for representative ferromagnetic, antiferromagnetic, and spin glass systems are shown in Fig. 4d-f, respectively (see ***Supplementary Figures 6*** for the evolution H for all 60 spin systems used in our experiments). The signature of SA can be seen in the energy landscapes, i.e. significant "hill-climbing" occurs during the initial iterations when the system is at higher "temperature", whereas at the lowest "temperature", the free energy decreases monotonically.

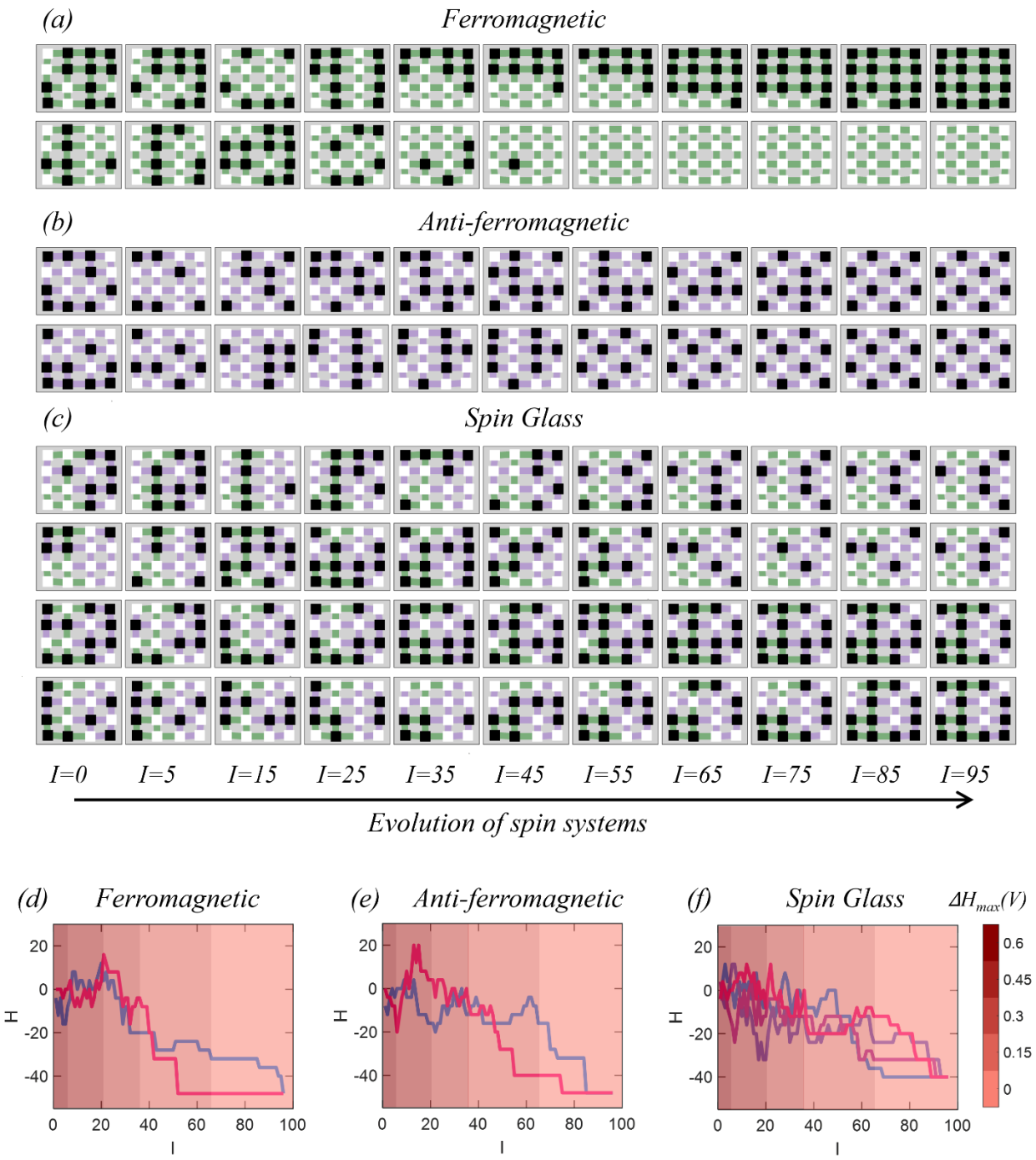

Figure 4. Experimental demonstration of SA. *Randomly initiated a) ferromagnetic, b) antiferromagnetic, and c) a spin glass system converge to their ground states as successive iterations (I) are performed using SA in hardware. Compared to an exhaustive search using BFT that requires a maximum of* 2^{K^2} *(= 65536 for* $K = 4$*) spin flips, SA accelerates the search by > 800 X. Evolution of free energy (H) with I for d) ferromagnetic, e) antiferromagnetic and f) spin glass systems for different ground states. The signature of SA can be seen in the energy landscapes. At higher "temperature", more "hill-climbing" is allowed, whereas at the lowest "temperature", the free energy decreases monotonically without any "hill-climbing".*

To obtain further insight, hardware-realistic simulation of SA is performed using the virtual source (VS) model developed in our earlier work to capture the 2D FET characteristics [46, 47]. Fig. 5a-c, respectively, show the convergence accuracy of 1000 randomly initiated 4×4 ferromagnetic, antiferromagnetic, and a spin glass system subjected to SA for different total number of iterations (I). All ferromagnetic and antiferromagnetic systems converged after ~600 and ~750 iterations resulting in average acceleration of ~110X and ~90X, respectively, compared to exhaustive BFTs. Note that these average acceleration numbers are lower than the experimental findings since more iterations are performed (200 at each temperature) at higher temperatures. The spin glass system, however, shows ~80% convergence accuracy after ~700 iterations. This is owing to the fact that any spin glass system is more prone to getting stuck in a metastable state. Fig. 5d shows the convergence accuracy for 100 randomly oriented $K \times K$ ferromagnetic spin systems as a function

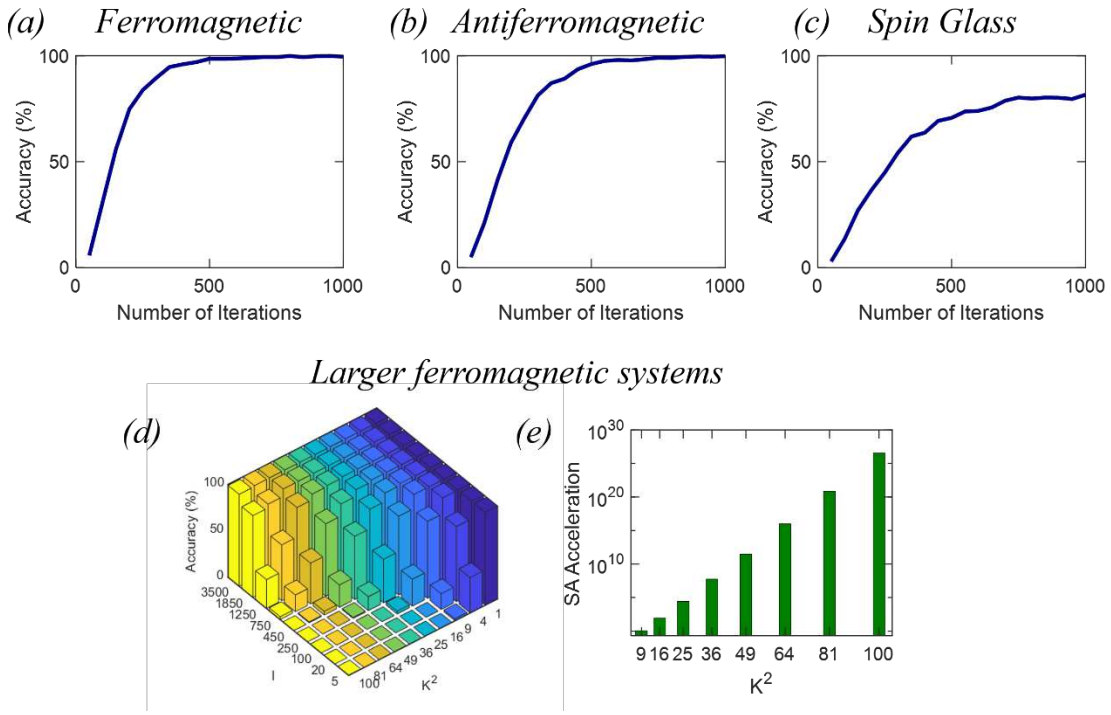

Figure 5. Hardware-realistic simulation of SA using virtual source model. *Convergence accuracy of 1000 randomly initiated 4×4 a) ferromagnetic, b) antiferromagnetic, and a c) spin glass system subjected to SA for different total number of iterations (I). d) Convergence accuracy for 100 randomly oriented $K \times K$ ferromagnetic spin systems as a function of total number of spins (K^2) and total number of SA iterations. e) Acceleration factor i.e., the ratio of maximum number of spin flips using brute force trail to the maximum number of SA spin flips for 100% convergence accuracy as a function of K^2.*

of total number of spins (K^2) and total number of SA iterations. Even for $K = 10$, SA requires only ~ 3500 iterations or maximum number of spin flips in comparison to ~10^{30} maximum number of spin flips required for an exhaustive search demonstrating the tremendous improvement in acceleration (~ 10^{26}X). Fig. 5e shows the acceleration for ~100% convergence accuracy as a function of K^2. Clearly, as K increases the benefits of SA becomes even more astounding.

Finally, we evaluate the energy expenditure (E_{SA}) by our annealing accelerator during each iteration following Eq. 7.

$$E_{SA} = E_{M1} + E_{M2} + E_{M3} + E_{M4}$$

$$= \sum_{i=1}^{K^2} \left\{ \left[(I_{\mathrm{T1}} + I_{\mathrm{T2}}) V_{in-2} \tau_p \right]_{M1} + \left[I_{\mathrm{out}} V_{out} \tau_p \right]_{M2} + \left[\frac{1}{2} C (\Delta V_I)^2 \right]_{M3} \right\} + \left[I_{\Delta H} V_{DS} \tau_p \right]_{M4} \quad [7]$$

Here, E_{M1}, E_{M2}, E_{M3}, and E_{M4} are the energy consumption by M1, M2, M3, and M4, respectively, and I_{T1} and I_{T2} are the current in T1 and T2, respectively. ***Supplementary Fig. 7*** shows E_{M1}, E_{M2}, E_{M3}, E_{M4}, and E_{SA} averaged over all 60 spin systems as a function of I for 4×4 spin lattice. The average energy expenditure for the hardware module was found to be miniscule ~1.3 nJ/iteration, which corresponds to maximum total energy expenditure of ~120 nJ for finding the ground state of any 4×4 spin system. Note that due to the limitations imposed by our measurement instruments, we have used τ_p = 60 ms. However, it is possible to scale τ_p and thereby reduce the energy expenditure even further. Also note that our energy calculations exclude the software operations performed using MATLAB such as the generation of random numbers.

Conclusion

This work successfully demonstrates hardware acceleration of SA for the Ising spin system by exploiting subthreshold conduction and analog programmability of complementary 2D FETs integrated with non-volatile floating-gate memory stack. By designing in-memory computing primitives and annealing schedule equivalent of cooling, we were able to achieve > 800X acceleration for 4×4 ferromagnetic, antiferromagnetic, and a spin glass system, experimentally, at frugal average energy expenditure of ~ 120 nJ. Our numerical simulations show more striking benefits of SA for search acceleration of larger spin lattices.

Methods

Device fabrication: Back gated MoS_2 and WSe_2 FETs are fabricated using e-beam lithography. MOCVD grown MoS_2 is transferred on to 50 nm Al_2O_3 substrate with PMMA (polymethyl-methacrylate) assisted wet transfer process. The substrate is spin coated with PMMA and baked at 180 $^{\circ}C$ for 90 s to define the channel region. The PMMA photoresist is then exposed to e-beam and developed using 1:1 4-methyl-2-pentanone (MIBK) and 2-propanol (IPA) mixture. Using sulfur hexafluoride (SF_6) at 5 $^{\circ}C$ for 30 s, the monolayer MoS_2 film is subsequently etched. Next the sample is rinsed in acetone and IPA to remove the photoresist. In order to fabricate the source/drain contacts the substrate is spin coated with MMA and PMMA followed by the e-beam lithography, developing using MIBK and IPA, and e-beam evaporation of 40 nm Ni/30 nm Au stack. Finally, the photoresist is rinsed away by lift off process using acetone and IPA.

For WSe_2, micromechanical exfoliation is performed to obtain optimally thin WSe_2 flakes on the 50 nm Al_2O_3 substrate. The source/drain contacts (10 nm Pt/30 nm Au) are defined using e-beam lithography as discussed above with a channel length of 1 μm. Following that, in order to fabricate the p-i-p structure, spin coating the channel with PMMA and subsequent e-beam exposure is used to expose 250 nm of the channel near the source and drain contact, leaving the middle 500 nm covered with PMMA. The WSe_2 FET is further doped with O_2 plasma using a Tepla M4L plasma etch tool. The WSe_2 FET is exposed to O_2 plasma with a power of 100 W for 300 s. O_2 and He gas flow rates of 150 $sccm$ and 50 $sccm$ with a chamber pressure of 500 mT is used for O_2 plasma doping. Finally, the photoresist is rinsed away by lift off process using acetone and IPA.

Electrical characterization: Lake Shore CRX-VF probe station and Keysight B1500A parameter analyzer were used to perform the electrical characterization at room temperature in high vacuum ($\approx 10^{-6}$ Torr). The measurements with the resistor and capacitor modules were performed outside the probe station on a bread board.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Code availability

The codes used for plotting the data are available from the corresponding authors on reasonable request.

References

[1] E. L. Lawler, *The Travelling Salesman Problem: A Guided Tour of Combinatorial Optimization*: John Wiley & Sons, 1985.

[2] D. Bookstaber, "Simulated Annealing for Traveling Salesman Problem," 1999.

[3] A. Singh and A. S. Baghel, "A new grouping genetic algorithm approach to the multiple traveling salesperson problem," *Soft Computing,* vol. 13, pp. 95-101, 2008.

[4] S. Nallaperuma, M. Wagner, and F. Neumann, "Analyzing the Effects of Instance Features and Algorithm Parameters for Max–Min Ant System and the Traveling Salesperson Problem," *Frontiers in Robotics and AI,* vol. 2, 2015.

[5] S. Gavrilov, D. Zheleznikov, V. Khvatov, and R. Chochaev, "Clustering optimization based on simulated annealing algorithm for reconfigurable systems-on-chip," pp. 1492-1495, 2018.

[6] Z. Wang, Y. Zhao, Y. Liu, and C. Lv, "A speculative parallel simulated annealing algorithm based on Apache Spark," *Concurrency and Computation: Practice and Experience,* vol. 30, p. e4429, 2018.

[7] X. Xiao, Y. Liu, H. Song, and T. Kikkawa, "Optimal microwave breast imaging using quality metrics and simulated annealing algorithm," *International Journal of RF and Microwave Computer-Aided Engineering,* vol. 30, 2020.

[8] L. M. R. Rere, M. I. Fanany, and A. M. Arymurthy, "Simulated Annealing Algorithm for Deep Learning," *Procedia Computer Science,* vol. 72, pp. 137-144, 2015.

[9] A. Biswas and T. Acharya, "A very fast simulated annealing method for inversion of magnetic anomaly over semi-infinite vertical rod-type structure," *Modeling Earth Systems and Environment,* vol. 2, pp. 1-10, 2016.

[10] F. Neri, "Case Study on Modeling the Silver and Nasdaq Financial Time Series with Simulated Annealing," vol. 746, pp. 755-763, 2018.

[11] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, and M. Troyer, "Optimised simulated annealing for Ising spin glasses," *Computer Physics Communications,* vol. 192, pp. 265-271, 2015.

[12] T. Leleu, Y. Yamamoto, S. Utsunomiya, and K. Aihara, "Combinatorial optimization using dynamical phase transitions in driven-dissipative systems," *Phys Rev E,* vol. 95, p. 022118, Feb 2017.

[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics,* vol. 21, pp. 1087-1092, 1953.

[14] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science,* vol. 220, pp. 671-80, May 13 1983.

[15] A. K. Peprah, S. K. Appiah, and S. K. Amponsah, "An Optimal Cooling Schedule Using a Simulated Annealing Based Approach," *Applied Mathematics,* vol. 08, pp. 1195-1210, 2017.

[16] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans Pattern Anal Mach Intell,* vol. 6, pp. 721-41, Jun 1984.

[17] G. Dueck and T. Scheuer, "Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing," *Journal of computational physics,* vol. 90, pp. 161-175, 1990.

[18] J. Niittylahti, H. Raittinen, and K. Kaski, "General purpose simulated annealing on hardware," pp. 5.2_3.1-5.2_3.6, 1993.

[19] D. Abramson, "A very high speed architecture for simulated annealing," *Computer,* vol. 25, pp. 27-36, 1992.

[20] A. M. Ferreiro, J. A. García, J. G. López-Salas, and C. Vázquez, "An efficient implementation of parallel simulated annealing algorithm in GPUs," *Journal of Global Optimization,* vol. 57, pp. 863-890, 2012.

[21] C. Cook, H. Zhao, T. Sato, M. Hiromoto, and S. X. D. Tan, "GPU-based Ising computing for solving max-cut combinatorial optimization problems," *Integration,* vol. 69, pp. 335-344, 2019.
[22] C. Yoshimura, M. Hayashi, T. Okuyama, and M. Yamaoka, "FPGA-based Annealing Processor for Ising Model," pp. 436-442, 2016.
[23] A. HajiRassouliha, A. J. Taberner, M. P. Nash, and P. M. F. Nielsen, "Suitability of recent hardware accelerators (DSPs, FPGAs, and GPUs) for computer vision and image processing algorithms," *Signal Processing: Image Communication,* vol. 68, pp. 101-119, 2018.
[24] H. Ushijima-Mwesigwa, C. F. A. Negre, and S. M. Mniszewski, "Graph Partitioning using Quantum Annealing on the D-Wave System," pp. 22-29, 2017.
[25] J. H. Shin, Y. J. Jeong, M. A. Zidan, Q. Wang, and W. D. Lu, "Hardware Acceleration of Simulated Annealing of Spin Glass by RRAM Crossbar Array," pp. 3.3.1-3.3.4, 2018.
[26] K. Yang, Q. Duan, Y. Wang, T. Zhang, Y. Yang, and R. Huang, "Transiently chaotic simulated annealing based on intrinsic nonlinearity of memristors for efficient solution of optimization problems," *Sci Adv,* vol. 6, p. eaba9901, Aug 2020.
[27] F. Cai, S. Kumar, T. Van Vaerenbergh, X. Sheng, R. Liu, C. Li*, et al.*, "Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks," *Nature Electronics,* vol. 3, pp. 409-418, 2020.
[28] M. R. Mahmoodi, H. Kim, Z. Fahimi, H. Nili, L. Sedov, V. Polishchuk*, et al.*, "An Analog Neuro-Optimizer with Adaptable Annealing Based on 64×64 0T1R Crossbar Circuit," pp. 14.7.1-14.7.4, 2019.
[29] Q. Smets, G. Arutchelvan, J. Jussot, D. Verreck, I. Asselberghs, A. N. Mehta*, et al.*, "Ultra-scaled MOCVD MoS 2 MOSFETs with 42nm contact pitch and 250μA/μm drain current," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 23.2. 1-23.2. 4.
[30] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, and S. Das, "Benchmarking monolayer MoS2 and WS2 field-effect transistors," *Nature Communications,* vol. 12, p. 693, 2021/01/29 2021.
[31] S. Wachter, D. K. Polyushkin, O. Bethge, and T. Mueller, "A microprocessor based on a two-dimensional semiconductor," *Nature communications,* vol. 8, p. 14948, 2017.
[32] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing*, et al.*, "A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector," *Nature Electronics,* vol. 3, pp. 646-655, 2020/10/01 2020.
[33] H. Jang, C. Liu, H. Hinton, M. H. Lee, H. Kim, M. Seol*, et al.*, "An Atomically Thin Optoelectronic Machine Vision Processor," *Adv Mater,* vol. 32, p. e2002431, Sep 2020.
[34] G. Migliato Marega, Y. Zhao, A. Avsar, Z. Wang, M. Tripathi, A. Radenovic*, et al.*, "Logic-in-memory based on an atomically thin semiconductor," *Nature,* vol. 587, pp. 72-77, 2020/11/01 2020.
[35] H. Maletta and W. Felsch, "Insulating spin-glass systemEuxSr1−xS," *Physical Review B,* vol. 20, pp. 1245-1260, 1979.
[36] M. K. Singh, W. Prellier, M. P. Singh, R. S. Katiyar, and J. F. Scott, "Spin-glass transition in single-crystalBiFeO3," *Physical Review B,* vol. 77, 2008.
[37] J. Vannimenus and G. Toulouse, "Theory of the frustration effect. II. Ising spins on a square lattice," *Journal of Physics C: Solid State Physics,* vol. 10, pp. L537-L542, 1977.
[38] S. F. Edwards and P. W. Anderson, "Theory of spin glasses," *Journal of Physics F: Metal Physics,* vol. 5, pp. 965-974, 1975.
[39] S. Das and J. Appenzeller, "WSe2 field effect transistors with enhanced ambipolar characteristics," *Applied Physics Letters,* vol. 103, Sep 2 2013.
[40] A. J. Arnold, D. S. Schulman, and S. Das, "Thickness Trends of Electron and Hole Conduction and Contact Carrier Injection in Surface Charge Transfer Doped 2D Field Effect Transistors," *ACS Nano,* vol. 14, pp. 13557-13568, 2020/10/27 2020.

[41] D. S. Schulman, A. J. Arnold, and S. Das, "Contact engineering for 2D materials and devices," *Chem Soc Rev,* Mar 2 2018.
[42] A. Wali, S. Kundu, A. J. Arnold, G. Zhao, K. Basu, and S. Das, "Satisfiability Attack-Resistant Camouflaged Two-Dimensional Heterostructure Devices," *ACS Nano,* Jan 28 2021.
[43] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, and S. Das, "Benchmarking monolayer MoS2 and WS2 field-effect transistors," *Nat Commun,* vol. 12, p. 693, Jan 29 2021.
[44] A. Dodda, A. Oberoi, A. Sebastian, T. H. Choudhury, J. M. Redwing, and S. Das, "Stochastic resonance in MoS2 photodetector," *Nature Communications,* vol. 11, p. 4406, 2020/09/02 2020.
[45] S. Das, H. Y. Chen, A. V. Penumatcha, and J. Appenzeller, "High performance multilayer MoS2 transistors with scandium contacts," *Nano Lett,* vol. 13, pp. 100-5, Jan 09 2013.
[46] A. Sebastian, A. Pannone, S. S. Radhakrishnan, and S. Das, "Gaussian synapses for probabilistic neural networks," *Nature communications,* vol. 10, pp. 1-11, 2019.
[47] S. Das, A. Dodda, and S. Das, "A biomimetic 2D transistor for audiomorphic computing," *Nature Communications,* vol. 10, p. 3450, 2019/08/01 2019.

AUTHOR INFORMATION

Corresponding Author

sud70@psu.edu, das.sapt@gmail.com

Author Contributions

A.S and Saptarshi Das conceived the idea and designed the experiments. A.S, Saptarshi Das, and Sarbashis Das performed the experiments, analyzed the data, discussed the results, agreed on their implications. All authors contributed to the preparation of the manuscript.

Competing Interest

The authors declare no competing interests

Acknowledgement

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. Authors also acknowledge Mr. Andrew J Arnold for help with WSe_2 device fabrication. Authors also acknowledge the materials support from the National Science Foundation (NSF) through the Pennsylvania State University 2D Crystal Consortium–Materials Innovation Platform (2DCCMIP) under NSF cooperative agreement DMR-1539916.

Figure Captions

Figure 1. Simulated annealing (SA) and Ising spin glass system. a) Illustration of the basic principle of SA used for finding the ground state(s) or lowest energy state(s) of a system in a large search space with multiple local minima. Unlike many other optimization methods, SA accepts transitions increasing H ("hill-climbing") based on the annealing temperature (T). b) A 4×4 Ising spin glass system with 16 randomly oriented spins in up (white) or down (black) direction. The neighboring spin interactions can be either ferromagnetic (green) or antiferromagnetic (purple). c) The corresponding coupling strength matrix $[CS]$. d) Free energy (H) corresponding to the 2^{16} possible states for the spin glass system in (b). The energy landscape shows multiple local minima with ground state degeneracy of 4. The number of brute force trials increases exponentially as the size of the spin lattice increases, qualifying the spin glass system as a challenging combinatorial optimization problem, where SA can accelerate the search. e) Flowchart of the SA algorithm for a spin glass system. A predetermined cooling schedule is used for T. At each T, the algorithm runs for a predetermined number of iterations (I_{max}). During each iteration, a random spin is flipped and ΔH associated with the spin flip is evaluated. For $\Delta H \leq 0$, the spin flip is always accepted. For $\Delta H > 0$, the spin flip is accepted if the cost is less than a predetermined value, P. For a sufficiently large number of iterations, the system converges to one of the f) 4 ground states for the spin glass systems in (b).

Figure 2. Analog in-memory complementary 2D field effect transistors (FETs). a) Schematic and b) optical image of a back-gated p-type WSe_2 FET. The channel is selectively exposed to mild O_2 plasma to form the p-i-p structure with the length of the intrinsic region as 500 nm. Corresponding c) transfer and d) output characteristics for WSe_2 FET. e) Schematic and f) optical

image of back-gated n-type MoS_2 FET with channel length as 500 nm. g) Transfer and h) output characteristics for MoS_2 FET. The p^{++}-Si/TiN/Pt/Al_2O_3 stack offers analog and non-volatile memory where the threshold voltage of the FETs can be adjusted by applying a programming pulse to the back gate. i) Transfer characteristics of post-programmed MoS_2 FET by applying negative programming voltages of different amplitudes for t_{P} = 100 ms. j) Retention characteristics, i.e., post-programmed I_{DS} *versus* time measured at V_{BG} = 0 V. i) Transfer and k) corresponding retention characteristics of post-erased MoS_2 FET after applying positive erase voltages of different amplitudes for t_{E} = 100 ms.

Figure 3. Circuit modules for hardware acceleration of SA. a) The multiplier module (M1) has a *p*-type WSe_2 FET (T1) and an *n*-type MoS_2 FET (T2), connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, $V_{\mathrm{in}-1}$ and $V_{\mathrm{in}-2}$. V_{in-1} is applied to the common-gate terminal, V_{in-2} is applied to the drain terminal of T1, and $-V_{in-2}$ is applied to the drain terminal of T2. b) Transfer characteristic of T1 and T2. c) Transfer characteristics of M1 i.e., output voltage, V_{out} *versus* $\mathrm{V_{in-1}}$ for $\mathrm{V_{in-2}}$ = ±0.1 V. Using M1 the product between the i^{th} elements of $[\sigma]$ and $[CS]_j$ is obtained at the i^{th} time step ($i\tau_p$) as $V_{out}(i\tau_p)$ by applying, $V_{in-1}(i\tau_p) = V_1\sigma_i$ and $V_{in-2}(i\tau_p) = V_2 CS_{ij}$. Note that i = 1:1:K^2, K = 4. We have used τ_p =60 ms, V_1 = 1 V, and V_2 = 0.1 V resulting in $V_{out}(i\tau_p)$ = 0.1×$CS_{ij}\sigma_i$. d) The voltage to current converter module (M2) transforms V_{out} from M1 into current, I_{out} following $I_{out} = GV_{out}$ with $G \approx 0.5$ µS. e) The integrator module (M3), a capacitor (C_I = 20 nF), sums I_{out} from M2 over K^2 time steps into voltage, $V_{\Delta H}$. f) V_{in-1}, V_{in-2}, $-V_{in-2}$, V_{out}, I_{out}, and the output from M3, i.e., V_I for representative ferromagnetic, antiferromagnetic, and a spin glass system during a given iteration of SA. $V_{\Delta H}$ and σ_j are multiplied to obtain ΔH. g) Schematic and h) transfer characteristics

of a programmable MoS_2 FET used for evaluating the cost associated with the state transition as well as for realizing cooling schedule in hardware. The subthreshold conduction governed by Boltzmann statistics is exploited to evaluate the cost of "hill-climbing" by applying $V_{\mathrm{BG}} = \Delta H$ and the spin flip is accepted if $I_{\Delta \mathrm{H}} < I_{\mathrm{cost}}$ = 100 pA. The cooling schedule is implemented by shifting the threshold voltage of the FET through back-gate programming.

Figure 4. Experimental demonstration of SA. Randomly initiated a) ferromagnetic, b) antiferromagnetic, and c) a spin glass system converge to their ground states as successive iterations (I) are performed using SA in hardware. Compared to an exhaustive search using BFT that requires a maximum of 2^{K^2} (= 65536 for $K = 4$) spin flips, SA accelerates the search by ~ 700X. Evolution of free energy (H) with I for d) ferromagnetic, e) antiferromagnetic and f) spin glass systems for different ground states. The signature of SA can be seen in the energy landscapes. At higher "temperature", more "hill-climbing" is allowed, whereas at the lowest "temperature", the free energy decreases monotonically without any "hill-climbing".

Figure 5. Hardware-realistic simulation of SA using virtual source model. Convergence accuracy of 1000 randomly initiated 4×4 a) ferromagnetic, b) antiferromagnetic, and a c) spin glass system subjected to SA for different total number of iterations (I). d) Convergence accuracy for 100 randomly oriented K× K ferromagnetic spin systems as a function of total number of spins (K^2) and total number of SA iterations. e) Acceleration factor i.e., the ratio of maximum number of spin flips using brute force trail to the maximum number of SA spin flips for 100% convergence accuracy as a function of K^2.

Figures

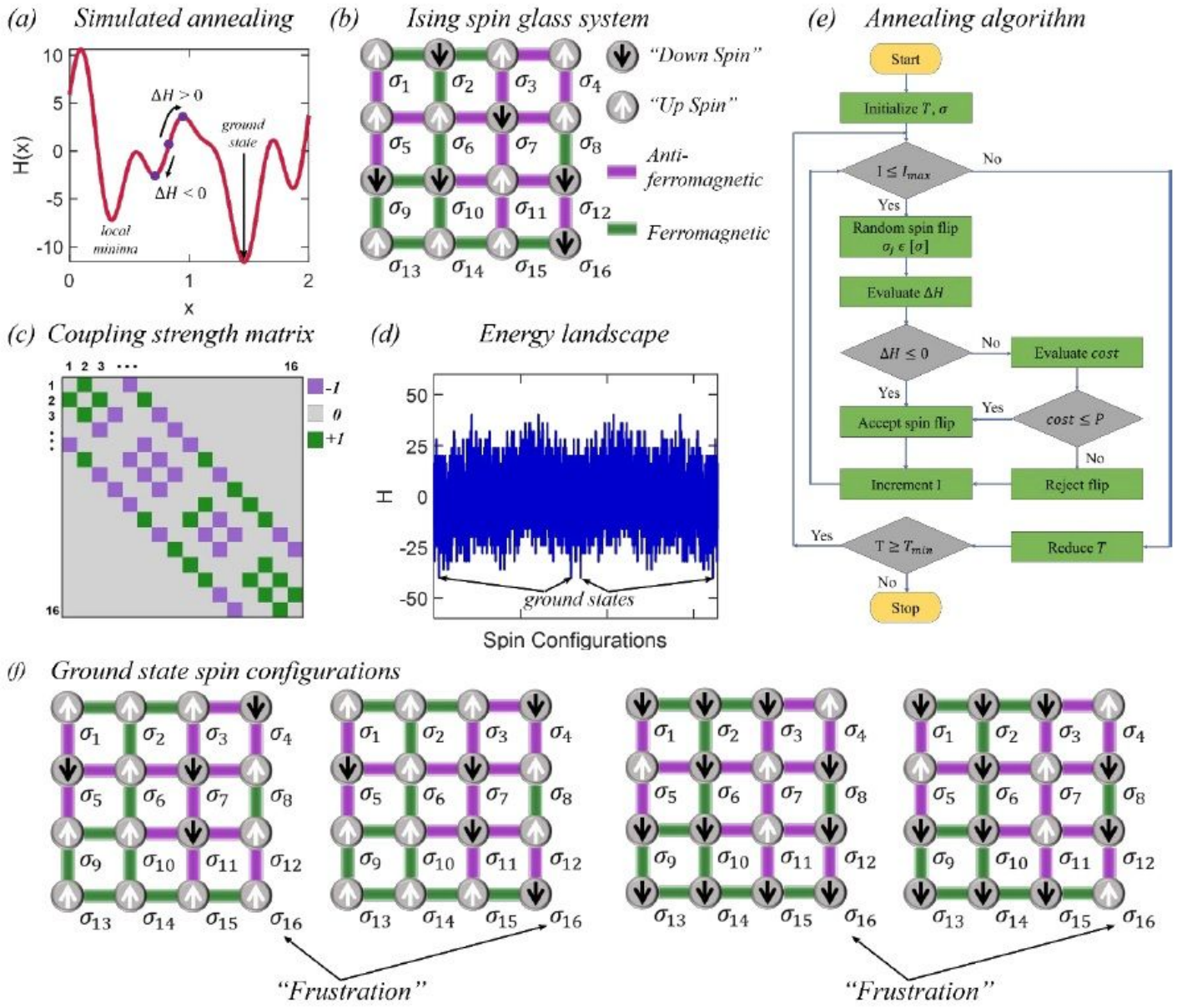

Figure 1

Simulated annealing (SA) and Ising spin glass system. a) Illustration of the basic principle of SA used for finding the ground state(s) or lowest energy state(s) of a system in a large search space with multiple local minima. Unlike many other optimization methods, SA accepts transitions increasing ⍰⍰ ("hill-climbing") based on the annealing temperature (⍰⍰). b) A 4×4 Ising spin glass system with 16 randomly oriented spins in up (white) or down (black) direction. The neighboring spin interactions can be either ferromagnetic (green) or antiferromagnetic (purple). c) The corresponding coupling strength matrix [⍰⍰⍰⍰]. d) Free energy (⍰⍰) corresponding to the 216 possible states for the spin glass system in (b). The energy landscape shows multiple local minima with ground state degeneracy of 4. The number of brute force

trials increases exponentially as the size of the spin lattice increases, qualifying the spin glass system as a challenging combinatorial optimization problem, where SA can accelerate the search. e) Flowchart of the SA algorithm for a spin glass system. A predetermined cooling schedule is used for ⬚. At each ⬚, the algorithm runs for a predetermined number of iterations (⬚⬚⬚⬚⬚⬚⬚⬚). During each iteration, a random spin is flipped and Δ⬚⬚ associated with the spin flip is evaluated. For Δ⬚⬚≤0, the spin flip is always accepted. For Δ⬚⬚>0, the spin flip is accepted if the cost is less than a predetermined value, ⬚. For a sufficiently large number of iterations, the system converges to one of the f) 4 ground states for the spin glass systems in (b).

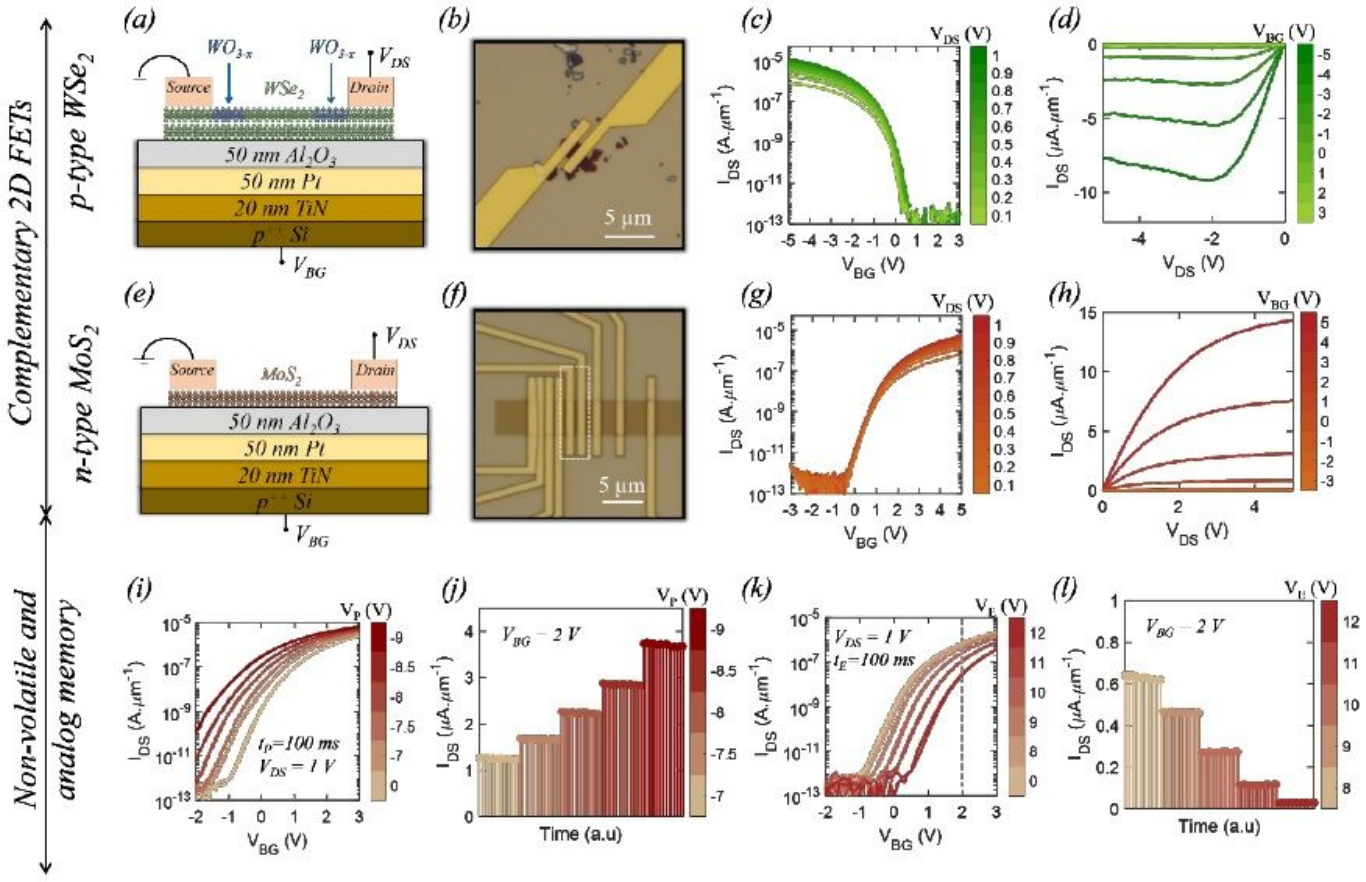

Figure 2

Analog in-memory complementary 2D field effect transistors (FETs). a) Schematic and b) optical image of a back-gated p-type WSe2 FET. The channel is selectively exposed to mild O2 plasma to form the p-i-p structure with the length of the intrinsic region as 500 ⬚⬚⬚. Corresponding c) transfer and d) output characteristics for WSe2 FET. e) Schematic and f) optical image of back-gated n-type MoS2 FET with channel length as 500 ⬚⬚⬚. g) Transfer and h) output characteristics for MoS2 FET. The p++-Si/TiN/Pt/Al2O3 stack offers analog and non-volatile memory where the threshold voltage of the FETs can be adjusted by applying a programming pulse to the back gate. i) Transfer characteristics of post-programmed MoS2 FET by applying negative programming voltages of different amplitudes for ⬚⬚⬚ = 100 ⬚ . j) Retention characteristics, i.e., post-programmed ⬚⬚⬚⬚ versus time measured at ⬚⬚⬚⬚ = 0 ⬚. i)

Transfer and k) corresponding retention characteristics of post-erased MoS2 FET after applying positive erase voltages of different amplitudes for □□□□ = 100 □ .

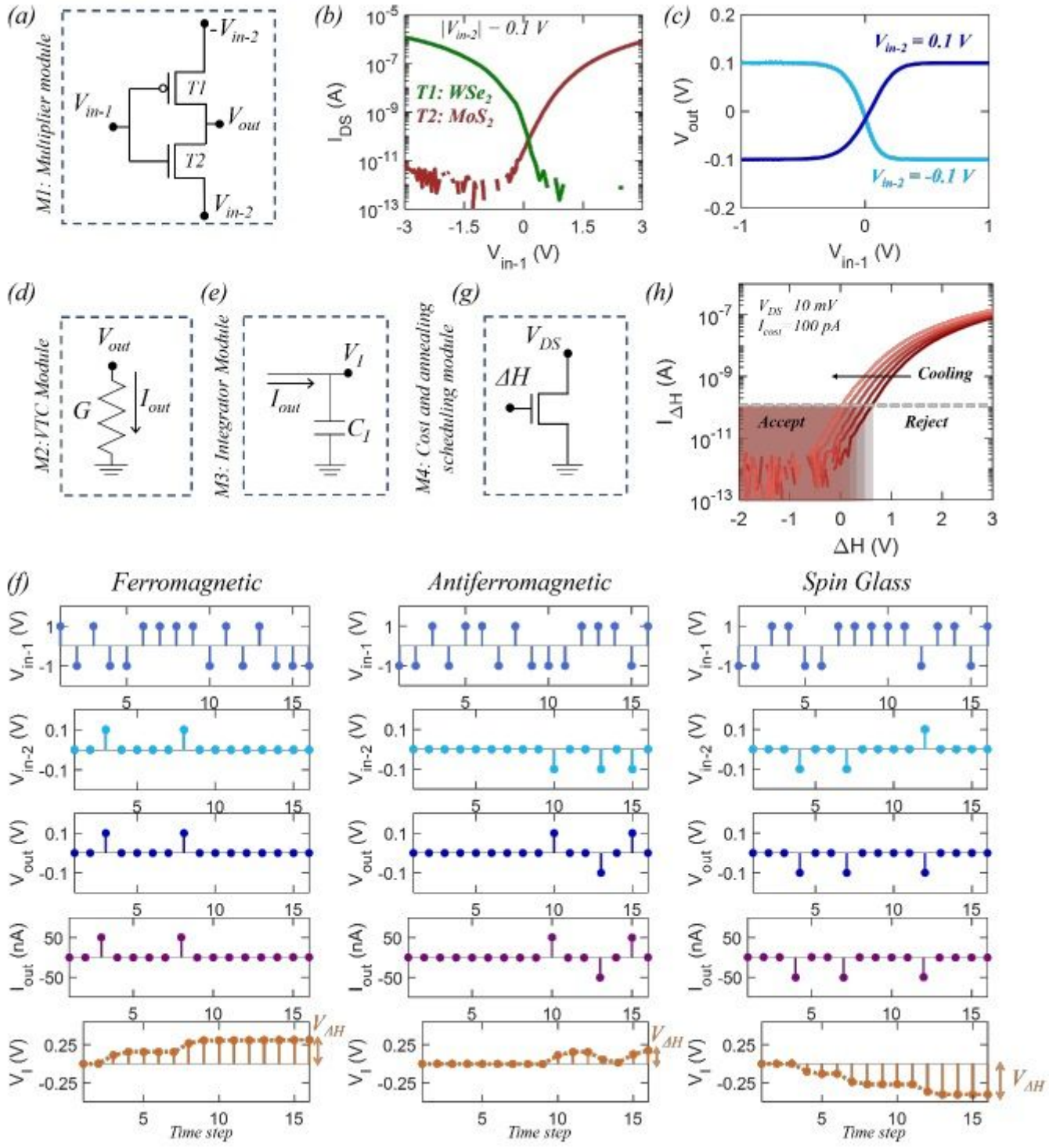

Figure 3

Circuit modules for hardware acceleration of SA. a) The multiplier module (M1) has a p-type WSe2 FET (T1) and an n-type MoS2 FET (T2), connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, □□□□□□−1 and □□□□□□−2. □□□□□□−1 is applied to the common-gate terminal, □□□□□□−2 is applied to the drain terminal of T1, and −□□□□□□−2 is applied to the

drain terminal of T2. b) Transfer characteristic of T1 and T2. c) Transfer characteristics of M1 i.e., output voltage, □□□□□□□ versus □□□□□□−1 for □□□□□□−2 = ±0.1 V. Using M1 the product between the □□□□□ elements of [□□] and [□□□□]□ is obtained at the □□□□□ time step (□□□□□□) as □□□□□□□ (□□□ □□) by applying, □□□□□□−1(□□□ □□)=□□1□□□□ and □□□□□□− Note that □□ = 1:1:□□2, □□ = 4. We have used □ □□ =60 ms, □□1 = 1 V, and □□2 = 0.1 V resulting in □□□□□□□ (□□□ □□) = 0.1×□□□□□□□□□□□□□. d) The voltage to current converter module (M2) transforms □□□□□□□ from M1 into current, □□□□□□□ following □□□□□□□ = □□□□□□□□□ with □□ ≈ 0.5 μS. e) The integrator module (M3), a capacitor (□□□□ = 20 nF), sums □□□□□□□ from M2 over □□2 time steps into voltage, VΔH. f) □□□□□□−1, □□□□□□−2, −□□□□□□−2, □□□□□□□ , □□□□□□□ , and the output from M3, i.e., □□□□ for representative ferromagnetic, antiferromagnetic, and a spin glass system during a given iteration of SA. □□Δ□□ and □□□ are multiplied to obtain □□□ . g) Schematic and h) transfer characteristics of a programmable MoS2 FET used for evaluating the cost associated with the state transition as well as for realizing cooling schedule in hardware. The subthreshold conduction governed by Boltzmann statistics is exploited to evaluate the cost of “hill-climbing” by applying □□□□□□ = □□□ and the spin flip is accepted if □□Δ□□ < □□□□□ □□□ = 100 □□□ . The cooling schedule is implemented by shifting the threshold voltage of the FET through back-gate programming.

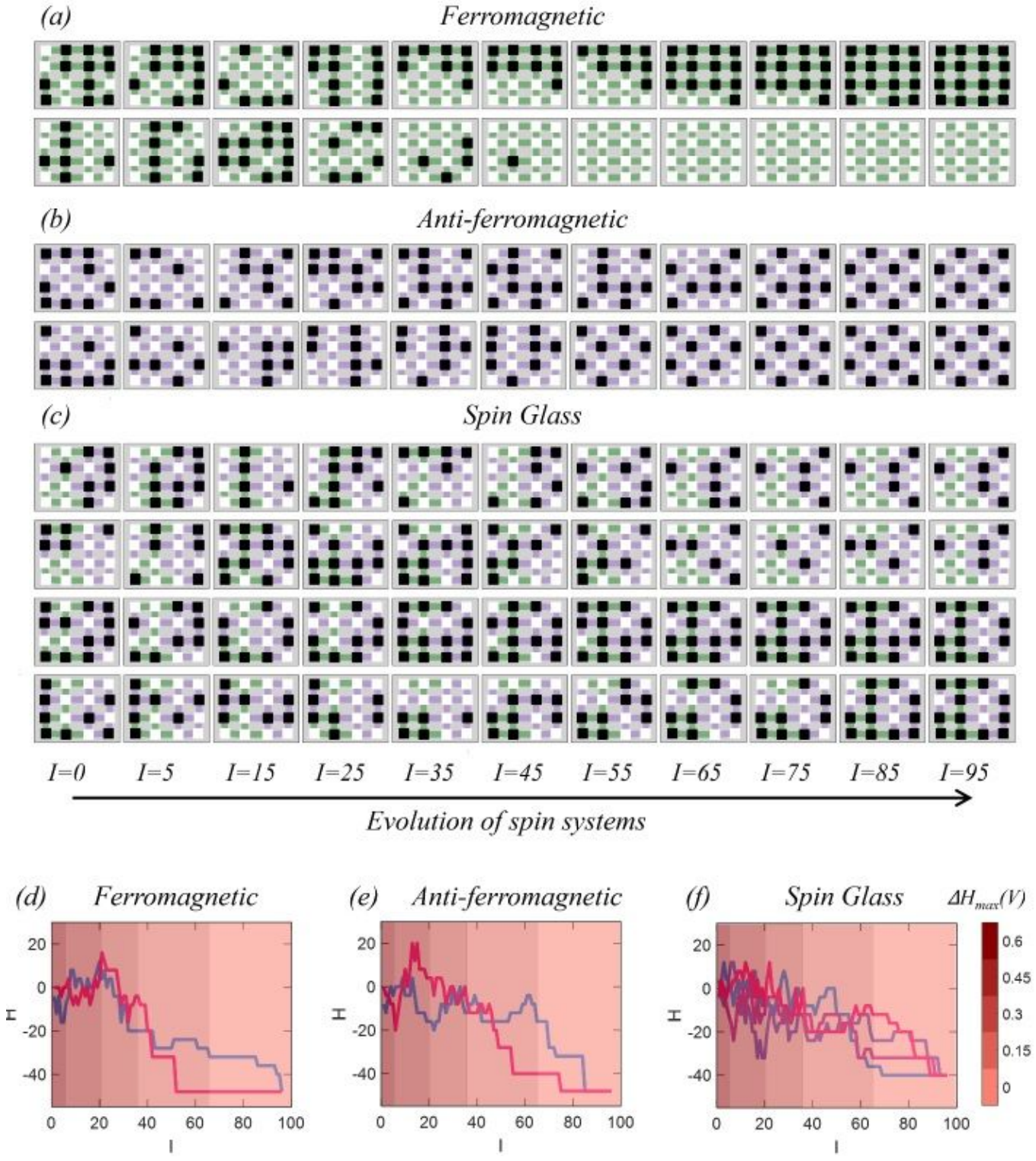

Figure 4

Experimental demonstration of SA. Randomly initiated a) ferromagnetic, b) antiferromagnetic, and c) a spin glass system converge to their ground states as successive iterations (I) are performed using SA in hardware. Compared to an exhaustive search using BFT that requires a maximum of 2□□2 (= 65536 for □□=4) spin flips, SA accelerates the search by > 800 X. Evolution of free energy (□□) with I for d) ferromagnetic, e) antiferromagnetic and f) spin glass systems for different ground states. The signature

of SA can be seen in the energy landscapes. At higher “temperature”, more “hill-climbing” is allowed, whereas at the lowest “temperature”, the free energy decreases monotonically without any “hill-climbing”.

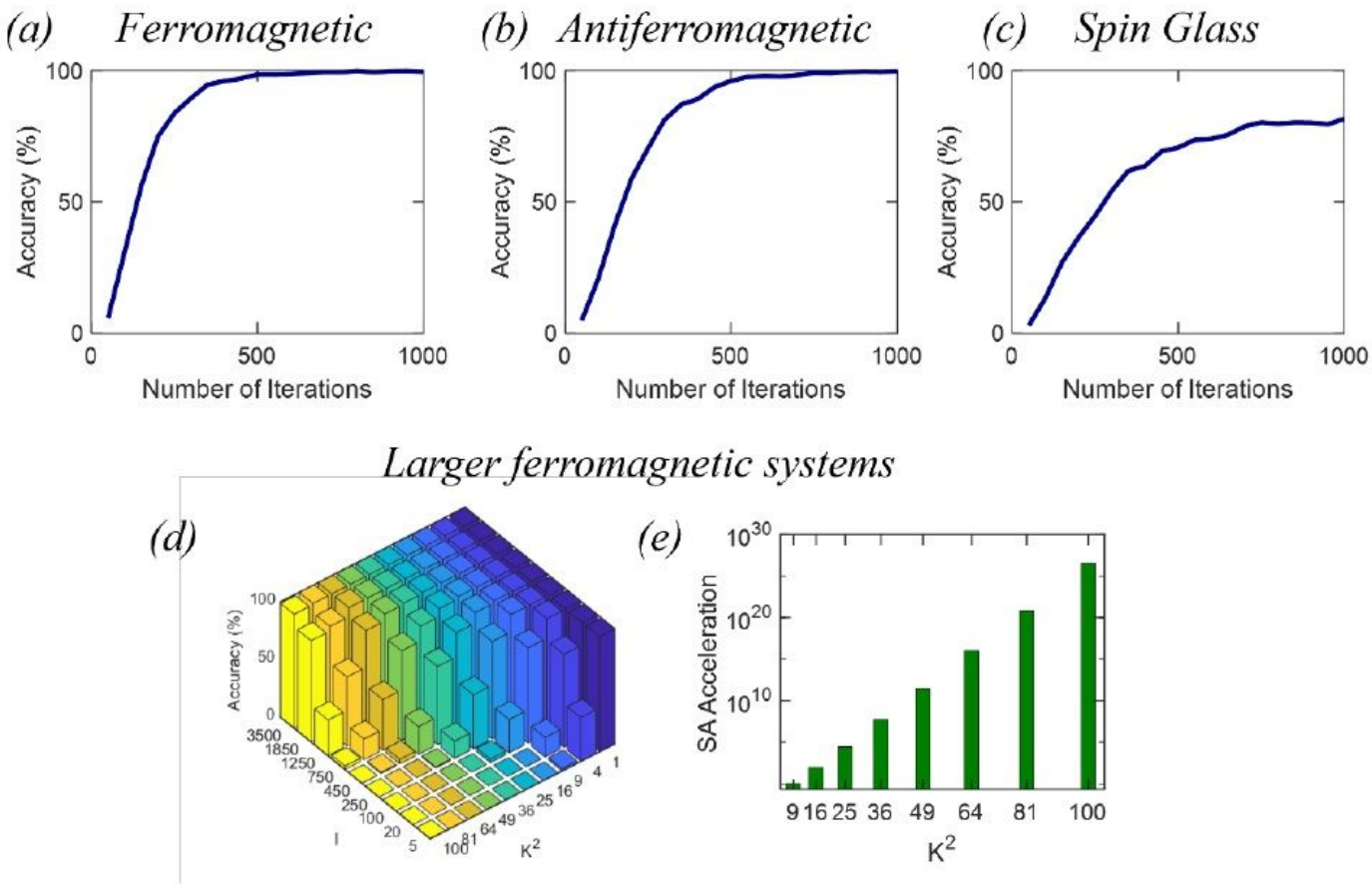

Figure 5

Hardware-realistic simulation of SA using virtual source model. Convergence accuracy of 1000 randomly initiated 4×4 a) ferromagnetic, b) antiferromagnetic, and a c) spin glass system subjected to SA for different total number of iterations (□□). d) Convergence accuracy for 100 randomly oriented □□× □□ ferromagnetic spin systems as a function of total number of spins (□□2) and total number of SA iterations. e) Acceleration factor i.e., the ratio of maximum number of spin flips using brute force trail to the maximum number of SA spin flips for 100% convergence accuracy as a function of □□2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryVideo1.mp4
- SupplementaryVideo2.mp4
- SupplementaryVideo3.mp4
- SupplementaryVideo4.mp4

- SupplementaryInformation.pdf