An Annealing Accelerator for Ising Spin Systems Based on In-Memory Complementary 2D FETs

Amritanand Sebastian, Sarbashis Das, and Saptarshi Das*

Metaheuristic algorithms such as simulated annealing (SA) are often implemented for optimization in combinatorial problems, especially for discrete 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 a high number of metastable states and ground-state degeneracy. In this work, subthreshold Boltzmann transport is exploited in complementary 2D field-effect transistors (p-type WSe2 and n-type MoS2) integrated with an analog, nonvolatile, and programmable floating-gate memory stack to develop in-memory computing primitives necessary for energy- and area-efficient hardware acceleration of SA for Ising spin systems. Search acceleration of >800× is demonstrated for 4×4 ferromagnetic, antiferromagnetic, and spin glass systems using SA compared to an exhaustive search using a brute force trial at miniscule total energy expenditure of ≈120 nJ. The hardware-realistic numerical simulations further highlight the astounding benefits of SA in accelerating the search for larger spin lattices.

1. 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, and industrial resource allocation to artificial intelligence, applied mathematics, and theoretical computer science. The traveling salesman problem (TSP) is a representative optimization problem where the shortest route connecting all of a given number of cities must be found.[3] In TSP, the time complexity is on 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) method, wherein every combination is evaluated. However, BFT becomes impractical for large n. For example, 40 cities would require 10^69 trials! While 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 the Ising model of magnetism.[11]

SA draws inspiration from physical annealing, in which a material is heated above its recrystallization temperature to allow atoms to rearrange and is then 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 also 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 (Figure 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, P, following Equation (1)

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

(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 simple structure.[17] The basic principle of SA is illustrated in Figure 1a. To minimize the free energy function $H(x)$ corresponding to the argument set $x = \{x_1,x_2,x_3,...,x_N\}$, where $N$ is a large number, a random $x_i$ is initialized. At each iteration, a new random point $x_i$ is chosen and the $\Delta H$ associated with the transition is evaluated. If $\Delta H < 0$, the transition is automatically accepted, whereas if $\Delta H > 0$, the transition is accepted conditionally, i.e., “hill-climbing” is allowed following Equation (1). The SA algorithm starts at a high $T$, where significant “hill-climbing” is allowed, and $T$ is progressively reduced following an annealing schedule. At a sufficient lower $T$ the system finds the global minima or arrives close to the same.

Most of the work on optimization algorithms such as SA is done in software,[11,13–16] with the few hardware demonstrations[18,19] based on von-Neumann architecture rendering...
In the Ising spin model, \( \sigma \) and \( -\sigma \) denote that only the nearest
neighbor interactions are taken into account, i.e.,
\( \sigma_{ij} \) denotes the nature of the interaction between
\( \sigma_i \) and \( \sigma_j \) in a 1T1R structure and also require peripheral
circuits based on silicon complementary–metal–oxide semiconductor (CMOS) technology.

In this work, we explore SA using in-memory complementary
field-effect transistors (FETs) based on ultrathin-body 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
development of 2D FET-based in-memory computing platforms. 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.

2. Spin Glass System

Since the early years of SA, the Ising spin glass problem has
been extensively studied since it offers a combinatorially 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
ordered.\(^{35,36}\) 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 stemming from frozen spin states and interactions. Spin glasses demonstrate interesting properties such as frustration. At low temperatures, a spin glass system has a roughly equal number of ferromagnetic and antiferromagnetic inter-
actions, 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 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 spec-
ic 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 Figure 1b. The natures of the various 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 \( [\sigma] = [\sigma_1, \sigma_2, \ldots, \sigma_{16}] \). 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 Edwards–Anderson (EA)
model described in Equation (2)\(^{38}\)

\[
H = -\sum_{ij} J_{ij} \sigma_i \sigma_j - \frac{1}{2} \sum_i \sigma_i N_i \sigma_i = -\frac{1}{2} \sum_i \sigma_i |CS_i| \sigma_i \\
= -\frac{1}{2} [\sigma|CS|\sigma]^T
\]  

Here, \( \sigma_i \) and \( \sigma_j \) are the ith and jth elements of \( \sigma \), \( J_{ij} \) denotes
the nature of the interaction between \( \sigma_i \) and \( \sigma_j \), i.e., \( J_{ij} = +1 \) for
ferromagnetic interactions and \( J_{ij} = -1 \) for antiferromagnetic
interactions. The operator \( \sigma \) denotes that only the nearest
neighbor interactions are taken into account, i.e., \( N_i = +1 \) if
\( \sigma_i \) and \( \sigma_j \) are immediate neighbors and \( N_i = 0 \) otherwise.\(^{25}\)
Note that both the \( J \) 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] (Figure 1c). The green squares represent +1, the purple
squares represent –1, and the gray squares represent 0. Note
that the \( K \times K \) spin glass system with \( K^2 \) spins, each with two
possible orientations ("up"/"down"), has \( 2^K \) possible states. Also
note that, by changing \( N \), long-range spin interactions can be
incorporated into the model. The free energy \( (H) \) values corre-
sponding to each of these \( 2^K \) states are shown in Figure 1d
for the spin glass system shown in Figure 1b. Clearly, the energy
landscape is non-monotonic with multiple local minima and
degenerate global minima, or ground states. Furthermore, as
the size of the spin lattice \( (K) \) increases, the search for spin
configurations that satisfy all interactions becomes unfeasible
using BFTs, thus qualifying the spin glass system as a chal-
lenging combinatorial optimization problem where SA can
speed-up the search process.

3. Simulated Annealing for Spin Glass System

The SA algorithm to find the ground state of a spin glass
system is shown in Figure 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 \( \Delta H \) value is
evaluated using Equation (2). If \( \Delta H < 0 \), 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 Equation (1) is lower than
the predetermined value, $P$. If both conditions are not satisfied, then the spin flip is rejected. At a given $T$, a predetermined number of iterations ($I_{\text{max}}$) 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 (Figure 1$^f$). Video S1 in the Supporting Information shows ground-state degeneracy. See Figure S1a–d in the Supporting Information for the spin increased as the ground state is unable to satisfy all interactions in opposite directions (Figure S1d, Supporting Information).

Since all interactions are satisfied if adjacent spins are oriented in the same direction, the system converges to its ground state (Figure 1$^f$). Note that the spin glass systems typically exhibit ground-state degeneracy.

In a ferromagnetic system, the ground state degeneracy is 2 since all interactions are satisfied if all spins are pointing "up" or "down" (Figure S1c, Supporting Information). An antiferromagnetic system also possesses ground state degeneracy of 2 since all interactions are satisfied if adjacent spins are oriented in opposite directions (Figure S1d, Supporting Information). 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 $\sigma_{\text{fs}}$ for the spin glass system to be in its ground state (Figure 1$^f$). Video S1 in the Supporting Information shows SA for $4 \times 4$ ferromagnetic, antiferromagnetic, and "frustrated" spin glass systems.

4. 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 ($\Delta H$) of the spin system associated with the random spin flip following Equations (2 and 3), 3) a computational unit to determine the cost of "hill-climbing" if $\Delta H > 0$ to accept or reject the spin flip following Equation (1), and, finally, 4) a hardware mechanism equivalent to the annealing/cooling schedule 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 $\Delta H$. Similarly, co-location of analog memory and analog computing enabled by nonvolatile and programmable 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 the same 50 nm Al$_2$O$_3$ gate dielectric and with a channel length of 500 nm are shown in Figure 2$^e,f$, respectively. We have used monolayer MoS$_2$ grown using the same metal–organic chemical vapor deposition (MOCVD) technique as described in our earlier reports[42,43] (see the Experimental 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[42,43]. The corresponding transfer characteristics and output characteristics are shown in Figure 2$g,h$, respectively. For monolayer MoS$_2$ FET, we extracted current on-off ratio of $\approx 10^8$, subthreshold slope, $SS = 310$ mV dec$^{-1}$, electron mobility, $\mu_e = 15$ cm$^2$ V$^{-1}$ s$^{-1}$, and on-current, $I_{\text{ON}} = 6.7$ $\mu$A $\mu$m$^{-1}$ for an inversion carrier density of $3.7 \times 10^{12}$ cm$^{-2}$.

The schematic and optical image of a back-gated p-type MoS$_2$ FET demonstrates unipolar p-type conduction with current on-off ratio of $\approx 10^8$, subthreshold slope, $SS = 310$ mV dec$^{-1}$, electron mobility, $\mu_e = 15$ cm$^2$ V$^{-1}$ s$^{-1}$, and on-current, $I_{\text{ON}} = 6.7$ $\mu$A $\mu$m$^{-1}$ for an inversion carrier density of $3.7 \times 10^{12}$ cm$^{-2}$.

The above demonstration of p-type WSe$_2$ FET and the back-gate electrode for a programming pulse time ($t_p$). The programmability is shown in Figure 2$\ell$ using the transfer characteristics of a MoS$_2$ FET at $V_{\text{DS}} = 1$ V after the application of negative programming voltages of different amplitudes for $t_p = 100$ ms. Figure 2$\ell$ shows the retention characteristics, i.e., post-programmed $I_{\text{DS}}$ vs time, measured at $V_{\text{BG}} = 2$ V, confirming nonvolatile and analog programmability of the MoS$_2$ FET. Similarily, 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 may be shifted in the opposite direction as shown in Figure 2$k$. The corresponding nonvolatile retention characteristics are shown in Figure 2$l$. Note that programming (erase) operations can also be achieved by varying $t_p$ ($t_E$) for a fixed $V_P$ ($V_E$), as shown in Figure 2$a,b$ in the Supporting Information. The working principle of the analog and nonvolatile back-gate memory stack has been described in detail in our earlier report[32].

The above demonstration of p-type WSe$_2$ FETs, n-type MoS$_2$ FETs, and the analog compute and analog nonvolatile storage capability allow us to design the computational primitives necessary for the hardware realization of SA. Note that $\Delta H$ due to random spin flip events can be computed from the difference in the free energy of the spin glass system before ($H_f$) and after ($H_i$) the spin flip. According to Equation (2), evaluation of $H_f$ and $H_i$ requires 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...
are difficult to realize using conductance states and require additional circuitry (note that earlier demonstration of SA\textsuperscript{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 \( \Delta H \) following Equation (3)

\[
\Delta H = H_I - H_I = \left( -\frac{1}{2} \sum_{i,j} C_{ij} \sigma_i \sigma_j \right) - \left( -\frac{1}{2} \sum_{i,j} C_{ij} \sigma_i \sigma_j \right) = \left( -\frac{1}{2} \Delta \sigma_j \right) \left[ \sum_{i=1}^{n} C_{ij} \sigma_i \right]
\]

Equation (3)

Here, we assume that the \( j \)th spin is flipped. Note that the first term inside the square bracket, i.e., \( -\frac{1}{2} \Delta \sigma_j \), is computationally equivalent to the initial spin state of \( \sigma_j \). For example, if \( \sigma_j \) flips from +1 to -1, \( \Delta \sigma_j = -2 \) and hence \( -\frac{1}{2} \Delta \sigma_j = +1 \) (Equation 3).

Therefore, following Equation (3), \( \Delta H \) can be obtained just by finding the dot product of the spin vector \( \sigma \) with the \( j \)th row of the [CS] matrix, i.e., \( [C_S]_j \), and by multiplying the result with \( \sigma_j \). Figure 3 shows the circuit modules used for computing \( \Delta H \). The multiplication module (M1) in Figure 3a is composed of a WSe\textsubscript{2} FET (T1) and MoS\textsubscript{2} FET (T2), connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, \( V_{in} \). \( V_{in} \) is applied to the common-gate terminal, \( V_{in} \) is applied to the drain terminal of T1, and \( -V_{in} \) 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 Figure 3b. The transfer characteristics of M1, i.e., output voltage, \( V_{out} \) vs \( V_{in} \), for \( V_{in} = \pm 1 \text{ V} \), are shown in Figure 3c. For \( V_{in} = 1 \text{ V} \), \( V_{out} = V_{in} \), whereas for \( V_{in} = -1 \text{ V} \), \( V_{out} = -V_{in} \). This is expected since for \( V_{in} = 1 \text{ V} \), T2 is more conductive than T1, thus allowing for \( V_{out} \) to follow \( V_{in} \) and vice versa for \( V_{in} = -1 \text{ V} \) (Figure 3b). The nonvolatile memory of the MoS\textsubscript{2} FET is used here to ensure that the cross-point between the transfer characteristics of T1 and T2 occurs in the subthreshold

Figure 2. Analog in-memory complementary 2D field-effect transistors (FETs). a) Schematic and b) optical image of a back-gated p-type WSe\textsubscript{2} FET. The channel is selectively exposed to mild \( \text{O}_2 \) plasma to form a p-i-p structure with the length of the intrinsic region as 500 nm. c,d) Corresponding transfer (c) and output (d) characteristics for the WSe\textsubscript{2} FET. e) Schematic and f) optical image of a back-gated n-type MoS\textsubscript{2} FET with a channel length of 500 nm. g) Transfer and h) output characteristics for the MoS\textsubscript{2} FET. The p++/Si/TiN/Al\textsubscript{2}O\textsubscript{3} stack offers analog and nonvolatile memory where the threshold voltage of the FETs can be adjusted by applying a programming pulse to the back gate. i) Transfer characteristics of the post-programmed MoS\textsubscript{2} FET after applying negative programming voltages of different amplitudes for \( t_P = 100 \text{ ms} \). j) Retention characteristics, i.e., post-programmed \( I_{DS} \) vs time, measured at \( V_{BG} = 2 \text{ V} \). k) Transfer and l) corresponding retention characteristics of the post-erased MoS\textsubscript{2} FET after applying positive erase voltages of different amplitudes for \( t_E = 100 \text{ ms} \).
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, \( V_{\text{in}-1} \) and \( V_{\text{in}-2} \). \( V_{\text{in}-1} \) is applied to the common-gate terminal, \( V_{\text{in}-2} \) is applied to the drain terminal of T1, and \(-V_{\text{in}-2}\) is applied to the drain terminal of T2. b) Transfer characteristics of T1 and T2. c) Transfer characteristics of M1, i.e., output voltage, \( V_{\text{out}} \) vs \( V_{\text{in}-1} \), for \( V_{\text{in}-2} = \pm 0.1 \text{ V} \). Using M1, the product between the \( i \)th elements of \([\sigma] \) and \([CS] \) is obtained at the \( i \)th time step (\( i\tau_p \)) as \( V_{\text{out}}(i\tau_p) \) by applying \( V_{\text{in}-1}(i\tau_p) = V_1\sigma_i \) and \( V_{\text{in}-2}(i\tau_p) = V_2CS_{ij} \). Note that \( i = 1:1:K_2 \) and \( K_2 = 4 \). We have used \( \tau_p = 60 \text{ ms}, V_1 = 1 \text{ V}, \text{ and } V_2 = 0.1 \text{ V, resulting in } V_{\text{out}}(i\tau_p) = 0.1 \times CS_{ij} \). d) The voltage-to-current converter module (M2) transforms \( V_{\text{out}} \) from M1 into current, \( I_{\text{out}} \), following \( I_{\text{out}} = GV_{\text{out}} \) with \( G \approx 0.5 \mu\text{S} \). e) The integrator module (M3), a capacitor (\( C_I = 20 \text{ nF} \)), sums \( I_{\text{out}} \) from M2 over \( K_2 \) time steps into voltage, \( V_{\Delta H} \). f) \( V_{\text{in}-1}, V_{\text{in}-2}, -V_{\text{in}-2}, V_{\text{out}}, I_{\text{out}}, \) and the output from M3, i.e., \( V_{\Delta H} \) for representative ferromagnetic, antiferromagnetic, and spin glass systems during a given iteration of SA. \( V_{\Delta H} \) and \( \sigma_j \) are multiplied to obtain \( \Delta H \). 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 \( V_{BG} = \Delta H \) and the spin flip is accepted if \( I_{\Delta H} < I_{\text{cost}} = 100 \text{ pA} \). The cooling schedule is implemented by shifting the threshold voltage of the FET through back-gate programming.
region, as shown in Figure 3b, and not in the OFF region. Having the cross-point in the OFF-state can lead to erroneous \( V_{\text{out}} \) values. Additionally, it is used to ensure that at \( V_{\text{in}-1} \) of \( \pm 1 \), \( V_{\text{out}} \) reaches \( \pm V_{\text{in}-2} \) (Figure 3c). This necessitates at least 3–4 orders of magnitude difference in resistance between T1 and T2 at \( V_{\text{in}-1} = \pm 1 \), as the multiplier module is basically a voltage divider circuit and, thus, having comparable resistances for T1 and T2 can lead to the \( V_{\text{out}} \) being lower than \( \pm V_{\text{in}-2} \). Using M1, the product between the ith elements of \( \sigma \) and \( [C_{\text{S}}] \) is obtained at the ith time step (\( t_{\text{p}} \)) as \( V_{\text{out}}(i) = V_{\text{in}-1}(i) \) by applying \( V_{\text{in}-1}(i) = V_{\text{in}}(i) \) and \( V_{\text{in}-2}(i) = V_{\text{in}}(i) \). We have used \( t_{\text{p}} = 60 \) ms, \( V_{1} = 1 \) V, and \( V_{2} = 0.1 \) V, resulting in \( V_{\text{out}}(i) = 0.1 \times C_{\text{S}i}\sigma_{i} \). Next, \( V_{\text{out}} \) is converted to \( I_{\text{out}} \) using the voltage-to-current converter module, M2 (Figure 3d), composed of a resistor, following \( I_{\text{out}} = G V_{\text{out}} \) with \( G = 0.5 \) \( \mu \)S, where \( G \) is the conductance. Finally, the contribution due to all spins are summed over \( K \) time steps using an integrator module, M3 (Figure 3e), composed of a capacitor \( (C_{1} = 20 \) nF), resulting in \( V_{\text{SH}} \) given by Equation (4)

\[
V_{\text{SH}} = \frac{r_{p}}{C_{1}} \sum_{i=1}^{K} I_{\text{out}}(i) = \frac{r_{p}}{C_{1}} \sum_{i=1}^{K} CV_{\text{in}}(i) C_{\text{S}i}\sigma_{i}.
\]

Figure 3f shows \( V_{\text{in}-1} \), \( V_{\text{in}-2} \), \( -V_{\text{in}-2} \), \( V_{\text{out}} \), \( I_{\text{out}} \), and the output of the integrator, \( V_{1} \), for representative ferromagnetic, antiferromagnetic, and spin glass systems during a given iteration. Finally, \( V_{\text{SH}} \) and \( \sigma \) are multiplied in MATLAB to obtain \( \Delta H \).

Following the evaluation of \( \Delta H \), it must be determined if the spin flip is accepted. This is done using two separate steps: one to determine the sign of \( \Delta H \) and another one to determine the cost of “hill-climbing” if \( \Delta H \) is positive. In hardware, both steps are combined in M4 (Figure 3g) by exploiting the subthreshold conduction in a FET where the carrier injection from the source contact into the semiconductor channel is given by Boltzmann statistics. The cost of “hill-climbing” is, therefore, obtained by applying \( V_{\text{BC}} = \Delta H \) and the spin flip is accepted if \( I_{\text{SH}} < I_{\text{cost}} = 100 \) pA (Figure 3h) following Equation (5)

\[
I_{\text{SH}} = I_{0} \exp\left( -\frac{q\Delta H}{mk_{B}T} \right).
\]

Here, \( q \) is the electronic charge, \( I_{0} \) is the static leakage current, and \( m \) is the body factor, which determines the SS following Equation (6)

\[
SS = \frac{mk_{B}T}{q} \ln(10); \quad m = \left( 1 + \frac{C_{S}}{C_{\text{OX}}} + \frac{C_{\text{IT}}}{C_{\text{OX}}} \right).
\]

Here, \( C_{S} \) is the semiconductor capacitance, \( C_{\text{IT}} \) is the interface trap capacitance, and \( C_{\text{OX}} \) is the oxide capacitance. For fully depleted ultrathin-body FETs, \( C_{S} = 0 \). We extracted a SS of 430 mV dec\(^{-1}\) and hence \( m \) of 71. Note that in Figure 3i the transfer characteristics are represented as a plot of \( I_{\text{SH}} \) versus \( \Delta H \). Also note that all negative \( \Delta H \) (low-energy transition) naturally leads to \( I_{\text{SH}} < I_{\text{cost}} \), whereas positive \( \Delta H \) up to \( \Delta H_{\text{max}} \) satisfies the “hill-climbing” criterion of \( I_{\text{SH}} < I_{\text{cost}} \).

Next, we implement cooling schedule in hardware. While temperature can be used to change the current \( I_{\text{SH}} \) for the same \( \Delta H \) following Equation (5), physically changing the temperature of a system requires excessive energy. Alternatively, \( \Delta H_{\text{max}} \) can be modulated by programming the MoS\(_2\) FET in different states as shown in Figure 3h. As the annealing temperature is lowered, the transfer characteristics are shifted towards the left. This ensures acceptance of more positive \( \Delta H \) at higher temperatures and no “hill climbing” at \( T = 0 \) K, i.e., \( \Delta H_{\text{max}} = 0 \) V. As shown in Figure 3h, our annealing schedule involved five different \( \Delta H_{\text{max}} \), i.e., 0.6, 0.45, 0.3, 0.15, and 0 V, analogous to temperature in metallurgical annealing with maximum number of iterations, \( I_{\text{max}} = 5, 15, 15, 30, \) and 30, at the respective “temperatures”.

The flow chart for the experimental demonstration of SA using our hardware-software approach is summarized in Figure S3 in the Supporting Information. The operations performed in hardware are highlighted in red. First, the spin system, i.e., \([\sigma] \) and \([C_{\text{S}}]\), and parameters of the simulated annealing algorithm such as \( I_{\text{SH}} \) and \( \Delta H_{\text{max}} \) are initiated in MATLAB. Using the random number generator in MATLAB, a random spin is flipped. \( \Delta H \) due to this random spin flip is computed in hardware, using M1, M2 and M3. By applying \( \Delta H \) to M4, \( I_{\text{SH}} \) is evaluated to check if the spin flip is accepted. After \( I_{\text{max}} \) iterations, \( \Delta H_{\text{max}} \) is reduced according to the cooling schedule using M4 and the process is repeated for \( \Delta H_{\text{max}} \) up to 0.

Figure 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 Figures S4–S6 and Videos S2–S4 in the Supporting Information 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 \( 2^{K} \) (i.e., \( 65 \times 536 \times 4 \)) 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 \( \approx 1365 \times \), \( \approx 1260 \times \), and \( \approx 1310 \times \), respectively, whereas, the average acceleration for the systems that converged to their ground states were \( \approx 850 \times \), \( \approx 900 \times \), and \( \approx 810 \times \), respectively. Evolution of \( H \) for representative ferromagnetic, antiferromagnetic, and spin glass systems are shown in Figure 4d–f, respectively (see Figure S7 in the Supporting Information 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.

To obtain further insight, hardware-realistic simulation of SA is performed using the virtual source (VS) model developed in our earlier work to capture 2D FET characteristics.\(^{46,47}\)
1000 randomly initiated $4 \times 4$ ferromagnetic, antiferromagnetic, and spin glass systems subjected to SA for different total numbers of iterations ($I$). All ferromagnetic and antiferromagnetic systems converged after $\approx 600$ and $\approx 750$ iterations, resulting in average accelerations of $\approx 110\times$ and $\approx 90\times$, 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). 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. Figure 5d shows the 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. Even for $K = 10$, SA requires only ~3500 iterations, or maximum number of spin flips, in comparison to an $10^{30}$ maximum number of spin flips required for an exhaustive search, demonstrating the tremendous improvement in acceleration ($10^{30} \times$). Figure 5e shows the acceleration for ~100% convergence accuracy as a function of $K^2$. Clearly, as $K$ increases the benefits of SA become even more astounding. Additionally, note that hardware modules don’t need to scale with $K$. The same modules can be implemented for larger spin systems. The only difference is that, in the case of a $K \times K$ system, the length of the vectors such as $V_{in-2}$, $V_{in-3-2}$, $V_{out}$, and $I_{out}$ that are passing through various modules would be $1 \times K^2$.

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

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

$$E_{SA} = \sum_{i=1}^{K^2} \left[ (I_{T1} + I_{T2})V_{in-2} \tau_p \right]_{M1} + \left[ I_{out} V_{out} \tau_p \right]_{M2} + \left[ 1/2 C(\Delta V)^2 \right]_{M3} + \left[ I_{in} V_{in} \tau_p \right]_{M4}$$

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. Note that this calculation does not take parasitic capacitances in our FETs into account. Figure S8a in the Supporting Information shows $E_{M1}$, $E_{M2}$, $E_{M3}$, and $E_{M4}$ averaged over all 60 spin systems as a function of $I$ for a $4 \times 4$ spin lattice. The average energy expenditure for the hardware module was found to be a minuscule $\approx 1.3$ nJ per iteration, which corresponds to a maximum total energy expenditure of $\approx 120$ nJ for finding the ground state of any $4 \times 4$ spin system. Note that due to the limitations imposed by our measurement instruments, we have used $\tau_p = 60$ ms. However, it is possible to scale $\tau_p$ 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.

We also attempt to estimate the energy consumption for larger spin systems. Since only the neighbor interactions are considered, out of these $K^2$ timesteps for a $K \times K$ system, ideally, there would be a maximum of 4 spikes and the rest would be zero. However, for the hardware implementation a nonideality is seen. When $V_{in-2} = 0$, $V_{out}$ must be zero for the multiplier module. However, in practice, when $V_{in-2} = 0$, $V_{out} = 10^{-4}$. This non ideality can carry into the other modules and lead to higher power consumption for larger systems. From Figure S8a in the Supporting Information, M2 and M3, i.e., the VTC module and the integrator module, dominate energy consumption. Hence, we estimate the worst-case energy consumption for spin systems with $K$ ranging from 1 to 1000, for both modules M3 and M4, as shown in Figure S8b in the Supporting Information. Figure S8b in the Supporting Information shows that the increase in energy consumption due to this non ideality is minimal even for very large spin systems, ensuring low energy consumption for larger spin systems. Figure S8b in the Supporting Information represents the energy consumption per iteration. Hence, note that there would be an obvious increase in energy associated with the larger number of iterations needed for larger spin systems.
5. 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 nonvolatile floating-gate memory stack. By designing in-memory computing primitives and annealing schedule equivalent of cooling, we were able to achieve >800× acceleration for 4 × 4 ferromagnetic, antiferromagnetic, and spin glass systems experimentally and at a frugal average energy expenditure of ≈120 nJ. Our numerical simulations show more striking benefits of SA for search acceleration of larger spin lattices.

6. Experimental Section

Device Fabrication: Back-gated MoS2 and WSe2 FETs were fabricated using e-beam lithography. MOCVD grown MoS2 was transferred onto 50 nm Al2O3 substrate using a PMMA (poly(methyl methacrylate)) assisted wet-transfer process. The substrate was spin-coated with PMMA and baked at 180 °C for 90 s to define the channel region. The PMMA photoresist was then exposed to an e-beam and developed using a 1:3 4-methyl-2-pentanone (MIBK) and 2-propanol (IPA) mixture. Using sulfur hexafluoride (SF6) at 5 °C for 30 s, the monolayer MoS2 film was subsequently etched. Next, the sample was rinsed in acetone and IPA to remove the photoresist. In order to fabricate the source/drain contacts, the substrate was spin-coated with MMA and PMMA and again patterned using e-beam lithography and developed using MIBK and IPA. E-beam evaporation was then used to deposit a 40 nm Ni/30 nm Au stack to serve as the contacts. Finally, the photoresist was rinsed away in a lift-off process using acetone and IPA.

For WSe2, micromechanical exfoliation was performed to obtain optimally thin WSe2 flakes on the 50 nm Al2O3 substrate. The source/drain contacts (10 nm Pt/30 nm Au) were 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, the channel was spin-coated with PMMA and a subsequent e-beam exposure was used to expose 250 nm of the channel near the source and drain contact, leaving the middle 500 nm covered with PMMA. The WSe2 FET was further doped with O2 plasma using a Tepla M4L plasma etch tool at a power of 100 W for 300 s. O2 and He gas flow rates of 150 sccm and 50 sccm with a chamber pressure of 500 mT were used for the O2 plasma doping. Finally, the photoresist was rinsed away in a lift-off process using acetone and IPA.

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

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The authors also acknowledge Andrew J Arnold for help with WSe2 device fabrication. The 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.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

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

Data Availability Statement

Research data are not shared.

Keywords

2D materials, field-effect transistors, Ising spin lattices, simulated annealing

Received: September 7, 2021
Revised: October 20, 2021
Published online: December 10, 2021

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The authors also acknowledge Andrew J Arnold for help with WSe2 device fabrication. The 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.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

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

Data Availability Statement

Research data are not shared.

Keywords

2D materials, field-effect transistors, Ising spin lattices, simulated annealing

Received: September 7, 2021
Revised: October 20, 2021
Published online: December 10, 2021

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The authors also acknowledge Andrew J Arnold for help with WSe2 device fabrication. The 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.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

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

Data Availability Statement

Research data are not shared.

Keywords

2D materials, field-effect transistors, Ising spin lattices, simulated annealing

Received: September 7, 2021
Revised: October 20, 2021
Published online: December 10, 2021

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The authors also acknowledge Andrew J Arnold for help with WSe2 device fabrication. The 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.
[16] S. Geman, D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. 1984, 6, 721.
[17] G. Dueck, T. Scheuer, J. Comput. Phys. 1990, 90, 161.
[18] J. Niittylahti, H. Raittinen, K. Kaski, in Proc. IEEE Winter Workshop on Nonlinear Digital Signal Processing, IEEE, Piscataway, NJ, USA 1993, https://doi.org/10.1109/NDSP.1993.767727.
[19] D. Abramson, Computer 1992, 25, 27.
[20] A. M. Ferreiro, J. A. García, J. G. López-Salas, C. Vázquez, J. Glob. Optim. 2012, 57, 863.
[21] C. Cook, H. Zhao, T. Sato, M. Hiromoto, S. X. D. Tan, Integration 2019, 69, 335.
[22] C. Yoshimura, M. Hayashi, T. Okuyama, M. Yamaoka, in 2016 Fourth Int. Symp. on Computing and Networking (CANDAR), IEEE, Piscataway, NJ, USA 2016, pp. 436–442, https://doi.org/10.1109/CANDAR.2016.0081.
[23] A. HajiRassouliha, A. J. Taberner, M. P. Nash, P. M. F. Nielsen, Signal Proc.: Image Commun. 2018, 68, 101.
[24] H. Ushijima-Mwesigwa, C. F. A. Negre, S. M. Mniszewski, in Int. Conf. for High Performance Computing, Networking, Storage and Analysis (PMES), ACM, New York, 2017, pp. 22–29, https://doi.org/10.1145/3149526.3149531.
[25] J. H. Shin, Y. J. Jeong, M. A. Zidan, Q. Wang, W. D. Lu, in 2018 IEEE Int. Electron Devices Meeting (IEDM), IEEE, Piscataway, NJ, USA 2018, https://doi.org/10.1109/IEDM.2018.8614698.
[26] K. Yang, Q. Duan, Y. Wang, T. Zhang, Y. Yang, R. Huang, Sci. Adv. 2020, 6, eaba9901.
[27] F. Cai, S. Kumar, T. Van Vaerenbergh, X. Sheng, R. Liu, C. Li, Z. Liu, M. Foltin, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. D. Lu, J. P. Strachan, Nat. Electron. 2020, 3, 409.
[28] M. R. Mahmoodi, H. Kim, Z. Fahimi, H. Nili, L. Sedov, V. Polishchuk, D. B. Strukov, in 2019 IEEE Int. Electron Devices Meeting (IEDM), IEEE, Piscataway, NJ, USA 2019, https://doi.org/10.1109/IEDM19573.2019.8993442.
[29] Q. Smets, G. Arutchelvan, J. Jusso, D. Verreck, I. Asselberghs, A. N. Mehta, in 2019 IEEE Int. Electron Devices Meeting (IEDM),