Title

Scalable Emulation of Sign-Problem–Free Hamiltonians with Room-Temperature p-bits

Permalink
https://escholarship.org/uc/item/25z842hd

Journal
Physical Review Applied, 12(3)

ISSN
2331-7043

Authors
Camsari, Kerem Y
Chowdhury, Shuvro
Datta, Supriyo

Publication Date
2019-09-01

DOI
10.1103/physrevapplied.12.034061

Peer reviewed

Scalable Emulation of Sign-Problem−Free Hamiltonians with Room Temperature p-bits

Kerem Y. Camsari, Shuvro Chowdhury and Supriyo Datta[1]
[1] *School of Electrical and Computer Engineering, Purdue University, IN, 47907*

(Dated: October 23, 2019)

The growing field of quantum computing is based on the concept of a q-bit which is a delicate superposition of 0 and 1, requiring cryogenic temperatures for its physical realization along with challenging coherent coupling techniques for entangling them. By contrast, a probabilistic bit or a p-bit is a robust classical entity that fluctuates between 0 and 1, and can be implemented at room temperature using present-day technology. Here, we show that a probabilistic coprocessor built out of room temperature p-bits can be used to accelerate simulations of a special class of quantum many-body systems that are sign-problem−free or “stoquastic”, leveraging the well-known Suzuki-Trotter decomposition that maps a d-dimensional quantum many body Hamiltonian to a d+1-dimensional classical Hamiltonian. This mapping allows an efficient emulation of a quantum system by classical computers and is commonly used in software to perform Quantum Monte Carlo (QMC) algorithms. By contrast, we show that a compact, embedded MTJ-based coprocessor can serve as a highly efficient hardware-accelerator for such QMC algorithms providing several orders of magnitude improvement in speed compared to optimized CPU implementations. Using realistic device-level SPICE simulations we demonstrate that the correct quantum correlations can be obtained using a classical p-circuit built with existing technology and operating at room temperature. The proposed coprocessor can serve as a tool to study stoquastic quantum many-body systems, overcoming challenges associated with physical quantum annealers.

I. INTRODUCTION

The basic building block of conventional digital electronics is the CMOS (Complementary Metal Oxide Semiconductor) transistor that is used to represent deterministic bits, that are either 0 or 1. Quantum computing, on the other hand, is based on q-bits that are coherent, delicate superpositions of 0 and 1. It is possible to define an entity intermediate between bits and q-bits that are classical but probabilistic, which we call “p-bits” [1]. It has been argued that just as three-terminal transistors provide a building block for large functional circuits, a three terminal realization of the p-bit can provide a building block for p-circuits [2] reminiscent of the probabilistic computer described by Feynman in the same paper that helped launch the field of quantum computing [3].

Such p-circuits can perform useful functions broadly relevant in the context of quantum computing and machine learning [4]. For example, p-circuits can be used to perform classical annealing in hardware [5], perform integer factorization by operating multipliers in an invertible mode [1, 6], just like quantum annealers that have been used for similar applications [7, 8]. In the machine learning context, p-bits can function as hardware accelerators for binary stochastic neurons [9] that can be used to become efficient inference engines [10, 11], or they can be used in an efficient calculation of correlations to accelerate learning algorithms, an application area also discussed in the context of quantum computing [12–15].

Scope

In this paper, we introduce an application of p-circuits to accelerate Quantum Monte Carlo (QMC) simulations of quantum systems based on the well-known Suzuki-Trotter decomposition that maps a d-dimensional quantum many body Hamiltonian to a d+1-dimensional classical Hamiltonian. This allows a quantum system to be emulated by a number of classical replicas that are interacting with each other [16] (FIG. 1) and this approach is commonly used in software or high-level hardware simulations [17–21]. By contrast, we show that a compact, embedded MTJ-based coprocessor can speed up the simulation by several orders of magnitude.

For a class of quantum Hamiltonians generally referred to as stoquastic Hamiltonians [22] that avoid the sign problem [23] and are therefore amenable to efficient QMC simulation, it should be possible to build hardware accelerators using replicated p-bits to emulate the thermodynamics of q-bit networks. The number of p-bits required to emulate a given q-bit network is typically a factor of 25-100 larger, but this is offset by the relative ease of implementation. Three-terminal p-bits can be implemented at room temperature with Magnetoresistive Random Access Memory (MRAM) technology which is currently in production with hundreds of millions memory cells. Non-magnetic and completely digital implementations of p-bits are also possible [6, 24] though they would require much larger energy and area [25] and while they can provide speed up over CPU/GPU implementations they would not achieve the potential speed up that can be obtained with the MTJ-based implementation.

A highly efficient classical coprocessor made out of conventional p-bits could overcome fundamental difficulties

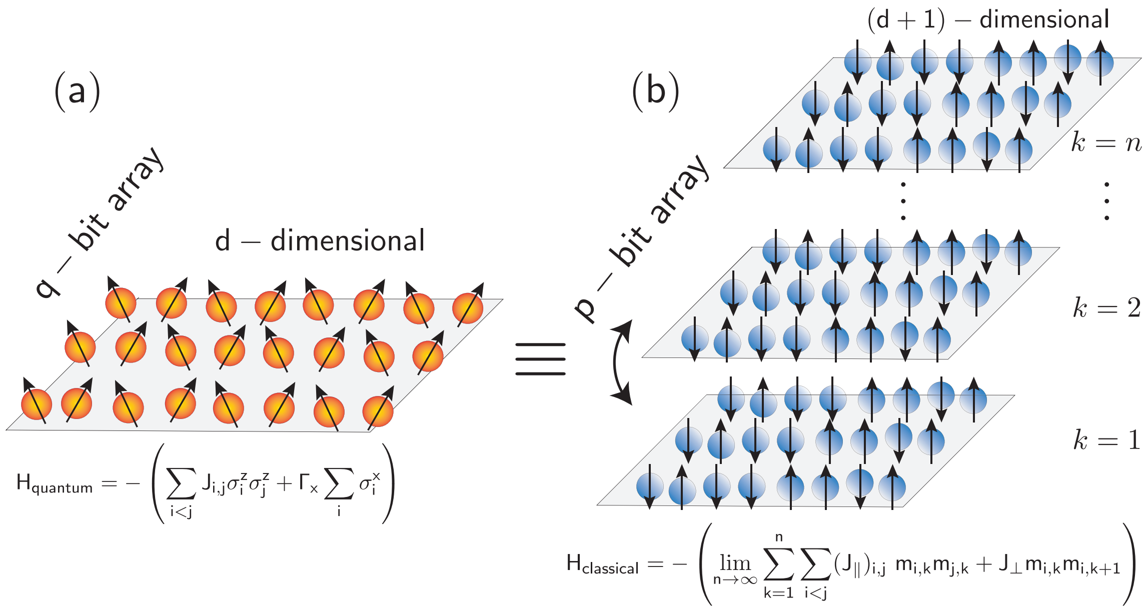
(a)
q − bit array
d − dimensional
$H_{quantum} = -\left(\sum_{i<j} J_{i,j}\sigma_i^z\sigma_j^z + \Gamma_x \sum_i \sigma_i^x\right)$
≡
(b)
p − bit array
(d + 1) − dimensional
$k = n$
$k = 2$
$k = 1$
$H_{classical} = -\left(\lim_{n\to\infty}\sum_{k=1}^{n}\sum_{i<j}(J_{\parallel})_{i,j}\, m_{i,k}m_{j,k} + J_{\perp}m_{i,k}m_{i,k+1}\right)$

FIG. 1. **q-bit to p-bit mapping:** (a) A d-dimensional q-bit array described by the Transverse Ising Hamiltonian can be mapped to a d+1-dimensional p-bit array with n-replicas that are coupled in the vertical direction by the Suzuki-Trotter decomposition (The case for $d = 2$ is illustrated). In this scheme, the replicas are always connected with periodic boundary conditions such that $m_{i,n+1} = m_{i,1}$. The many-body quantum and the corresponding classical Hamiltonian are shown where the operators σ^z, σ^x of the quantum system are replaced with binary p-bits in the classical system with $m_{i,j} \in \{-1, +1\}$. Corresponding coupling terms are $(J_{\parallel})_{i,j} = J_{i,j}/n$ and $J_{\perp} = -1/(2\beta)\log\tanh(\beta\Gamma_x/n)$.

associated with the low temperature operation of quantum annealers [26] while operating almost as fast as physical annealers. For example, it has recently been shown that an optimized CPU-based simulated quantum annealing (SQA) implementation was 10^8 times slower than a physical quantum annealer, even though it has shown similar algorithmic scaling on a model problem [19].

With appropriate magnet designs [27] individual p-bits can flip in a nanosecond or less so that with a million of them operating in parallel, we should have ∼ petaflips per second which is several orders of magnitude faster than existing digital implementations including parallelized GPU [28] and multi-core CPU implementations [29] that operate typically with ∼ 1-30 gigaflips per second.

It has also recently been suggested [30] that among quantum annealing options, SQA exhibits the best scaling properties, performing even better than experimental quantum annealers in some cases. As such, accelerating the software-based SQA with specialized hardware is a desirable goal and has led to recent interest in this area [21, 31].

Another advantage of p-bit networks is that unlike q-bit networks they can be interconnected using conventional electronic devices such as GPUs or FPGAs. This could allow all-to-all connectivity beyond nearest neighbor coupling without requiring any special encoding [32, 33]. Moreover, it should allow the implementation of arbitrary k-body interactions that are usually avoided by introducing ancillary bits to map them into 2-body interactions [34, 35].

Organization of the paper

We start in Section II, with a description of the mapping from the q-bit network to the p-bit network, along with the behavioral equations describing the dynamics of the latter. These behavioral equations for p-circuits are similar to those used for stochastic neural networks and are often implemented in software for machine learning applications. However, a hardware implementation can provide a significant speed-up especially because it can allow parallel asynchronous operation under the right conditions.

Next in Section III we consider a common example of a stoquastic Hamiltonian, namely the Transverse Ising Hamiltonian [36, 37], commonly employed by quantum annealers [38]. We compare the exact quantum results for the averages and correlations with the results obtained from the p-bit network demonstrating the impressive accuracy that can be achieved with a limited number of replicas. In Section IV we show another example, namely the ferromagnetic Heisenberg Model [16, 39], once again comparing the exact quantum results with probabilistic simulations of the p-bit network. In Section V, we show how Classical and Quantum Annealing can be performed using a network of p-bits. Finally in Section VI

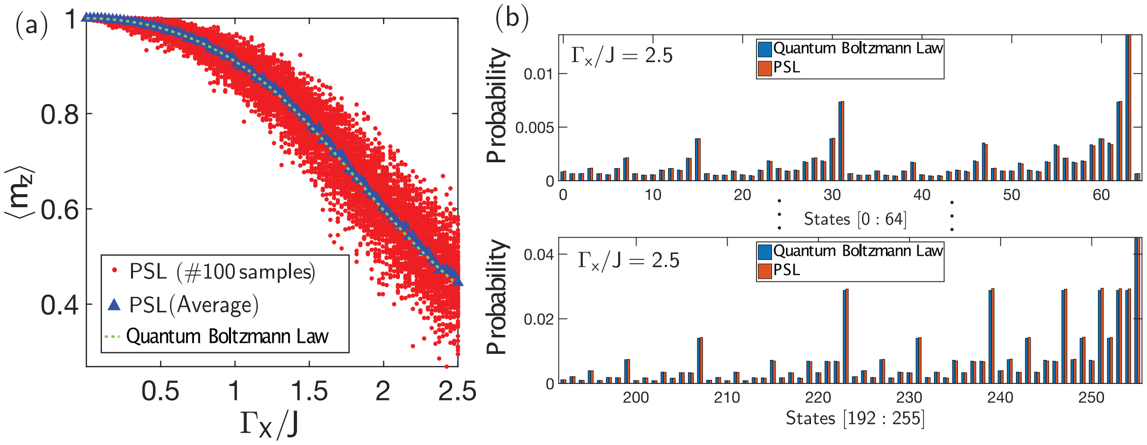

FIG. 2. **Exact quantum solution of a 1D Transverse Ising Hamiltonian vs Probabilistic Spin Logic (PSL):** (a) A 1D ferromagnetic linear chain ($J_{i,j} = +2$) with $M = 8$ spins (nearest neighbor with periodic boundary conditions) described by the quantum Transverse Ising Hamiltonian (Eq. 1) is solved exactly, as a function of the transverse magnetic field (Γ_x) at an inverse temperature of $\beta = 10$. A symmetry breaking magnetic field in the $+\hat{z}$ direction is used, $\Gamma_z = 1$, so that at $\Gamma_x = 0$, all spins are pointing in the $+\hat{z}$ direction. The green dashed line is obtained by evaluating Eq. 4 as a function of Γ_x. The red dots represent 100 different PSL runs obtained with different RNGs, each running for $t_f = 2000$ time steps. The blue triangles represent the average of PSL simulations and closely match the exact solution, establishing the accuracy of the quantum to classical mapping, with $n = 250$ replicas. (b) A probability histogram of correlations of the form $|\downarrow\downarrow\ldots\downarrow\rangle = 0, |\uparrow\uparrow\ldots\uparrow\rangle = 255$ are obtained from PSL and Quantum Boltzmann Law at $\Gamma_x/J = 2.5$ that corresponds to the last point of the x-axis in (a). Only a portion of the states are shown for clarity, states in between show essentially identical agreement.

we present SPICE simulations of actual hardware implementations that can be built with existing Embedded Magnetoresistive RAM (eMRAM) technology that has been under development by a number of foundries [40–42]. Unlike standard eMRAM where a non-volatile MTJ is carefully engineered with a large energy barrier ($E_B \approx$ 40-60 k_BT) so that the magnetization state is retained for a long time [43], the free layer of the MTJ for the p-bit is designed as a thermally unstable magnet ($E_B \approx 0\ k_BT$) whose magnetization rapidly fluctuates in time in the presence of thermal noise [44]. Using full device-level SPICE simulations corresponding to the p-bit and a resistive interconnection matrix, we demonstrate that the correct quantum correlations can be obtained using this classical p-circuit which can be built with existing technology at room temperature.

II. Q-BIT TO P-BIT

Since the seminal work of Suzuki [16], it is well-known that a d-dimensional quantum many-body Hamiltonian can be mapped to a d+1-dimensional classical Hamiltonian applying the so-called Suzuki-Trotter decomposition [16, 45], which is used as a basis for PIMC methods to simulate quantum annealing using classical computers [17]. This decomposition results in the quantum system being mapped to a classical system with n replicas that are coupled to each other. In this paper we consider two examples as described in the next two Sections, but the principles apply to stoquastic Hamiltonians in general.

Consider for example the Transverse Ising Hamiltonian in 1D written as [37]:

$$\mathcal{H}_Q = -\left(\sum_i^M J_{i,i+1}\sigma_i^z\sigma_{i+1}^z + \Gamma_x \sum_i^M \sigma_i^x + \Gamma_z \sum_i^M \sigma_i^z\right) \quad (1)$$

The Suzuki-Trotter mapping produces the following classical 2D Hamiltonian [17]:

$$\mathcal{H}_C = -\Bigg(\lim_{n\to\infty} \sum_{k=1}^{n}\sum_{i=1}^{M} (J_{\parallel})_{i,i+1}\ m_{i,k}m_{i+1,k} + \gamma_z m_{i,k} + J_{\perp}\ m_{i,k}m_{i,k+1} \Bigg) \quad (2)$$

where $(J_{\parallel})_{i,j} = J_{i,j}/n$, n being the number of replicas, $\gamma_z = \Gamma_z/n$ and the vertical coupling term is $J_\perp = -1/(2\beta)\log\tanh(\beta\Gamma_x/n)$ and $m_{i,j} \in \{-1,+1\}$. Note how the quantum mechanical operators in Eq. 1 have become classical spins in Eq. 2. The mapping of Eq. 2 becomes exact in the limit of infinite replicas ($n \to \infty$) however, for finite replicas the error scales as $O(1/n^2)$ [18] and can be made arbitarily small by choosing an appropriate number of replicas.

Behavioral model for p-bits

The classical system expressed by Eq. 2 can be represented by p-circuits that are built out of p-bits. There are two central equations that are used to describe p-bit networks [1]:

$$m_i(t+1) = \mathrm{sgn}\big[r + \tanh\ \beta I_i(t)\big] \quad (3a)$$

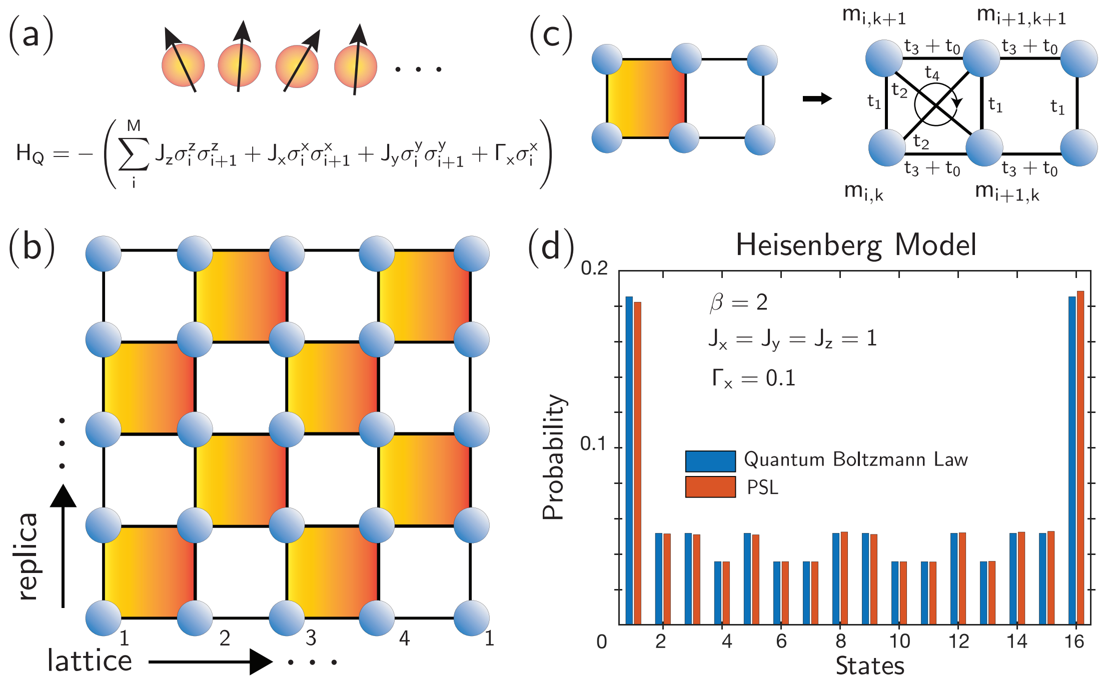

FIG. 3. **1D Heisenberg Model:** (a) A 1D Heisenberg model ($M = 4$ spins) with a transverse magnetic field is considered. (b) The chessboard lattice corresponding the 2D classical mapping of the model with periodic boundary conditions in the replica and lattice directions. (c) The interaction terms within the shaded and unshaded unit cells are shown. Within shaded cells, all neighbors are coupled vertically, diagonally and horizontally (t_1, t_2, t_3) in addition to a 4-body interaction energy term (t_4) that involves the product of all four spins. Terms that arise from the diagonal part of the Hamiltonian, H_0, result in additional horizontal interactions (t_0) in both the shaded and unshaded cells. (d) Probability histogram corresponding to a $M = 4$ spin ferromagnetic Heisenberg Model. The histogram is obtained by solving the Quantum Boltzmann Law (Eq. 4) with the parameters shown in the inset and by solving behavioral PSL equations but with a modified Eq. 3b to account for a 3-body term arising from the 4-body interaction. For the PSL simulation 20 replicas are used with $t_f = 10^7$ time steps. Samples taken from different replicas are considered independent and reduced to 16 probabilities to be compared with the original quantum system.

where t is dimensionless time that is incremented one at a time, r is a random number uniformly distributed between -1 and $+1$ and r at each time step is uncorrelated with the r chosen at the previous step. βI_i is the dimensionless current to each p-bit, where β is the inverse temperature. I_i in general, is calculated according to,

$$I_i(t) \equiv -\frac{\partial \mathcal{H}_C}{\partial m_i}$$

which in the present case, becomes:

$$I_i(t) = \left(b_i + \sum_j W_{ij} m_j(t)\right) \tag{3b}$$

where W_{ij} is the interconnection matrix and b_i is the bias term. We refer to Eq. 3 as Probabilistic Spin Logic (PSL) equations and note that these equations are essentially the same as those discussed in the context of stochastic neural networks such as Boltzmann Machines, developed by Hinton and colleagues [9].

It is important to note that while Eq. 3b is a linear synapse that typically arises from quadratic Hamiltonians with 2-body interactions, specially designed digital CMOS circuits can be used to implement more complicated interactions arising from cost functions such as generalized Hopfield models with k-body interactions [46, 47]. Such a flexibility of implementing complicated interactions could be a key advantage for hardware p-circuits.

PSL dynamics

PSL equations can be updated to approximate the steady state joint probability density for any W matrix, symmetric or asymmetric. For symmetric W matrices, the joint probability density is simply expressed by the classical Boltzmann Law, $\rho(\{m\}) \propto \exp[-\beta E(\{m\})]$, where E is the energy for a given configuration $\{m\}$, $E = 1/2\ m^T[W]m$. There are two important conditions regarding the updating of Eq. 3. First, Eq. 3b needs to be calculated much faster than Eq. 3a for proper convergence [6], a requirement particularly relevant for hardware implementations. Second, Eq. 3a needs to be updated sequentially, as in Gibbs sampling [48]. The requirement of sequential updating prohibits parallelization in software implementations, except in special cases such as restricted Boltzmann machines where the lack of intralayer connections between "visible" and "hidden" layers allows each layer to be updated in parallel [49]. For asynchronous hardware implementations, however, a clockless operation seems to satisfy the requirement of sequential updating naturally [5, 6].

III. TRANSVERSE ISING HAMILTONIAN

For the 1D Transverse Ising Hamiltonian (Eq. 1), we assume periodic boundary conditions such that $\sigma^z_{M+1} = \sigma^z_1$. Γ_x is the local transverse magnetic field and Γ_z is a local z-directed magnetic field. Eq. 1 can be constructed by first writing each term, σ_i, as a $2^M \times 2^M$ matrix followed by ordinary matrix multiplication for each product term. These terms are written in terms of 2×2 Pauli spin matrices ($\varsigma^{x,y,z}$) at the j^{th} lattice point as $\sigma_j = I \otimes I \otimes \ldots \otimes \varsigma \otimes \ldots \otimes I \otimes I$ where I is the 2×2 identity matrix and ς is the Pauli spin matrix at the j^{th} term in the product.

Quantum Boltzmann Law

In principle, Eq. 1 can be exactly solved for any quantity of interest as a function of temperature and all other parameters J and Γ, from the principles of quantum statistical mechanics [50]:

$$\langle S \rangle = \frac{\text{Tr.}\ [S_{op} \exp(-\beta \mathcal{H}_Q)]}{\text{Tr.}\ [\exp(-\beta \mathcal{H}_Q)]} \tag{4}$$

where $\beta \equiv 1/k_B T$ is the "inverse temperature" (as defined in Eq. 3a) and we have chosen to use a unit system in which $k_B = 1$. S is the quantity we wish to calculate with a corresponding operator S_{op}. In practice, directly solving Eq. 4 becomes intractable due to the exponential dependence of the Hamiltonian ($2^M \times 2^M$) to the size of the problem (M). Due to its similarity to the classical Boltzmann Law [51], we refer to Eq. 4 as the "Quantum Boltzmann Law" throughout this paper and solve it for small 1D systems. To obtain numerically stable results at low temperatures (high β), we first diagonalize the Hamiltonian and subtract the ground state energy from the diagonals, without changing any observable quantities.

Averages and correlations

In FIG. 2a we calculate the average z-magnetization of a 1D ferromagnetic ($J_{ij} = +2$) chain with $M = 8$ spins, as a function of a transverse magnetic field. The average z-magnetization, $\langle m_z \rangle$, is obtained by the operator $\sigma^z = \sum \sigma^z_j / M$ where σ^z_j provides the net z-spin, $|\uparrow\rangle - |\downarrow\rangle$, at site j. To break the symmetry of $m_z = \pm 1$ at low temperatures ($\beta = 10$) we introduce a $+\hat{z}$-directed magnetic field. As the transverse magnetic field increases, $\langle m_z \rangle$ gradually decreases, while $\langle m_x \rangle$ (not shown) increases, as spins become aligned with the transverse magnetic field. Incidentally, the reverse process, starting from a large Γ_x at a low temperature and slowly decreasing it to find the ground state of a complicated spin-glass, is commonly used in quantum annealing algorithms [18].

FIG. 2b shows the probabilities of correlated states at a given temperature and transverse field expressed as decimal numbers. This is done by first converting the states to binary numbers such that ↑ denotes +1 and ↓ denotes 0 and then converting the full state into a decimal number, for example the all down state $|\downarrow\downarrow\ldots\downarrow\rangle$ corresponds to 0, and the all up state $|\uparrow\uparrow\ldots\uparrow\rangle$ corresponds to 255 and so on. There are $2^8 = 256$ such states, each with a given probability obtained from Eq. 4. These correlated states are calculated by first constructing an operator for the probability of finding a $|\uparrow\rangle$ state at a given site, $P_j(|\uparrow\rangle) = (I + \sigma^z_j)/2$ where I is the $2^M \times 2^M$ identity matrix. Similarly, $P_j(|\downarrow\rangle) = (I - \sigma^z_j)/2$. Using these operators, any correlation of the form $|\downarrow\uparrow\ldots\uparrow\rangle$ can be calculated from the corresponding composite operator:

$$P(\downarrow\uparrow\ldots\uparrow) = P(\downarrow)P(\uparrow)\ldots P(\uparrow) = \prod_{k=1}^{M} P_k \tag{5}$$

There are 256 such operators and Eq. 4 can be used for each of them to obtain a probability for each state for any J, Γ, β. FIG. 2b shows these probabilities at a chosen parameter combination and they are in agreement with results obtained from a simulation of p-bits, as we next explain in Section II. Note that this joint probability density contains all statistical information in the system, as averages and other correlations of interest can be calculated from it, for example one can obtain $\langle m_z \rangle$ by weighting each state by the net z-spin they contribute to the average.

PSL vs Quantum Boltzmann Law

With this picture, the mapped classical Hamiltonian with n replicas described in Eq. 2 is used to obtain a consolidated $[W]$ matrix that is of size $(Mn) \times (Mn)$ to be used in Eq. 3. FIG. 2 shows the equivalence of the PSL implementation of the Transverse Ising Hamiltonian to the exact quantum many-body description for a 1D-chain with $M = 8$ spins. Note that the p-bit mapping can be applied to much larger spin systems, but an exact solution by Eq. 4 quickly becomes intractable. We investigate the average z-spin of this ferromagnetic chain at a constant temperature ($\beta = 10$) as a function of the transverse magnetic field, Γ_x. A symmetry breaking field (to favor a +1 order) of $\Gamma_z = 1$ is applied. As expected, the exact result shows how the average z-spin becomes disordered. The PSL results for a $n = 250$ replica system reproduce this behavior. The z-spin average is obtained by taking an average over the length of the chain, as well as over each replica. The final average (for a given red dot) is recorded at the end of $t_f = 2000$ dimensionless time steps. Since a single stochastic point is recorded at the end of t_f, for each Γ_x point, we observe a variance in the final results, however averaging over 100 different simulations for the same system, we get a very close match to the exact solution.

In FIG. 2b, the full joint probability density for the classical system is obtained from a PSL simulation that is run for $t_f = 10^5$ dimensionless time steps. The state of each replica with 8-spins is converted into a binary number at each time step, as in the exact solution, and then collected over all replicas. The striking agreement with PSL and the Quantum Boltzmann Law in FIG. 2 establishes the faithful mapping of the quantum system to the classical system, from the behavioral PSL equations Eq. 3.

IV. FERROMAGNETIC HEISENBERG MODEL

Before proceeding to a hardware implementation showing how replicated networks of p-bits can be built by existing nanodevices, we show another example of a stoquastic Hamiltonian that can be represented by p-bits. The Heisenberg Hamiltonian in 1D in the presence of a transverse magnetic field can be written as:

$$\mathcal{H}_Q = -\left(\sum_i^M J_z \sigma_i^z \sigma_{i+1}^z + J_x \sigma_i^x \sigma_{i+1}^x + J_y \sigma_i^y \sigma_{i+1}^y + \Gamma_x \sigma_i^x \right) \quad (6)$$

Following [16, 39, 52], we apply the Suzuki-Trotter transformation to this system and obtain the chessboard lattice that is shown in Fig. 3b with shaded and unshaded unit cells. The interactions terms for this hardware neural network are shown in Fig. 3c. For the shaded unit cells all two-body interactions $(t_1, t_2, t_3,)$ exist in addition to a 4-body interaction (t_4) that involves the product of each spin. The two-body interactions can be implemented using a linear synapse of the form of Eq. 3b but the 4-body interaction requires a non-linear synapse that computes the input terms that are products of three neighboring spins. The interaction terms t_0 arise due to the diagonal parts of the quantum system, as in the case of the Transverse Ising Hamiltonian, and exists for both the shaded and unshaded unit cells shown in Fig. 3c. The detailed derivations of all interaction terms are shown in Appendix A.

In Fig. 3d, we show a simulation of the classical system using behavioral PSL equations and compare this to the exact solution as before. We choose a set of parameters, $J_x = J_y = J_z = 1$ that corresponds to the ferromagnetic Heisenberg Model with a small transverse magnetic field in the x-direction, such that all off-diagonal terms in the $\exp(-\beta\mathcal{H}_Q)$ are *positive*, hence making this system *stoquastic* [52]. In this small example with $M = 4$ spins, we observe good agreement between the mapped system and the exact solution.

V. FACTORIZATION AS INVERSE MULTIPLICATION

So far we have shown probabilistic emulation of quantum systems in equilibrium without performing annealing. In Fig. 4, we show how classical and quantum annealing can be performed by the probabilistic coprocessor using Eq. 3a and Eq. 3b. We choose the problem of integer factorization by expressing a p-circuit that performs binary multiplication using Full Adders and AND gates. This structure is similar to the factorization descriptions in [6, 53–55]. We improve our previous design [6, 53] by eliminating nodes that are connected to each other, for example if a Full Adder carry out is connected to the carry in of another Full Adder, these two nodes are combined into a single node so that the p-bit in this node receives the sum of the inputs that each node receives. This allows us to reduce the problem size and exhibits better scaling for the inverse multiplier.

Fig. 4 compares Classical Annealing (CA) with simulated Quantum Annealing (SQA) with p-circuits. For classical annealing the temperature β^{-1} is linearly decreased from 1 to 0.1, while for simulated quantum annealing the inverse temperature is fixed at $\beta = 10$ but the transverse magnetic field is linearly reduced from 3 to 0.1. For these parameters, we observe that for the 8 bit multiplier, the SQA seems to perform better than CA as SQA finds the correct factors with 100% probability with 58% probability of finding (11,13) and 42% probability of finding (13,11) while CA finds the correct factors with 77.8% probability out of 100 samples. While we note that SQA seems to work better for this particular set of parameters, we have not attempted to optimize the parameters for CA or SQA. In SQA m-replicas of the original system is needed to map the Ising Hamiltonian to the corresponding Transverse Ising Hamiltonian. To make a fair comparison between SQA and CA in terms of the statistical samples being used, we performed CA with the same number of replicas but they are not interacting with each other. With a sophisticated synapse design that would allow replica swapping, replicas in the CA mode can be held at different temperatures for parallel tempering algorithms [56] but we do not explore this further. Our purpose has been to show that an autonomously operating p-circuit can be used to perform CA and SQA with *physical replicas* to speed up these algorithms as we show in the next section.

VI. P-BIT TO STOCHASTIC MRAM

We now show how the behavioral p-bit model can be represented by a stochastic neural network in hardware (FIG. 5a). Each replica in the classical system consists

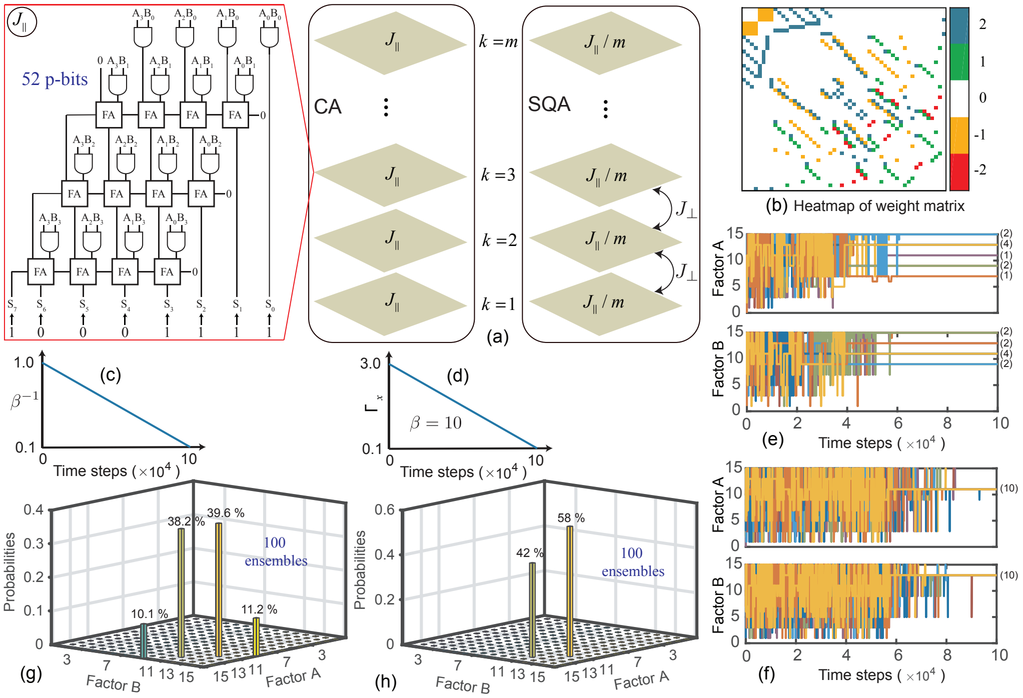

FIG. 4. Classical versus quantum annealing with p-circuits: (a) An 8-bit binary multiplier designed as a p-circuit. When operated in invertible mode this circuit functions as a factorizer. In the classical annealing (CA) scheme, this circuit is replicated m times ($m = 10$) without any interaction between replicas but to collect equal statistics as simulated quantum annealing (SQA). For SQA, the weight matrix is obtained according to the Transverse Ising mapping as discussed. (b) The heatmap of the weight matrix $J_{\|}$ shows that the connections among p-bits in the invertible multiplier circuit are sparse and discrete. (c) Classical annealing schedule: β^{-1} is linearly decreased from 1 to 0.1. (d) In SQA, the transverse field (Γ_x) is linearly reduced from 3 to 1 while $\beta = 10$. (e) The time evolution of CA: The numbers inside parenthesis on the right shows the multiplicity of replicas at a particular value. (f) Time evolution for SQA: All replicas find the right factors at the end annealing. (g-h) Histograms are calculated by averaging the results over 100 ensembles.

of p-bits that are interconnected to each other with a resistive network (synapse), a typical architecture often used in many hardware neural networks [57, 58] though for more complicated systems involving k-body interactions ($k > 2$), standard electronic devices such as FPGA's could also be used for this purpose, for example as in Ref. [59]. The extra dimension added by the Suzuki-Trotter transformation would increase the synaptic complexity but for sparse quantum networks, this transformation would only slightly increase the fan-in of each p-bit in the classical network.

We assume that the weighted summation is carried out by ideal operational amplifiers. The replicas are also connected in the vertical direction (not shown in FIG. 5) with nearest neighbor coupling according to the coupling coefficient $J_{\perp}$.

In the case of quantum annealing, the vertical resistors need to be reconfigurable, therefore they need to be designed differently compared to the fixed resistors that represent the transverse coupling $(J_{\|})_{i,j}$. In our device level examples, we use fixed resistors in order to establish the equivalence between the classical and quantum systems and have not performed annealing.

Network parameters

The device equations for the synapse and the p-bit shown in FIG. 5 are given as [10, 44]:

$$V_{\mathrm{OUT}_j}/(V_{\mathrm{DD}}/2) = \mathrm{sgn}[r + \tanh(V_{\mathrm{IN}_j}/V_0)] \qquad (7)$$

$$V_{\mathrm{IN}_j} = \sum_i \frac{R_{\mathrm{ref}}}{R_{ji}} \overline{V}_{\mathrm{OUT}_i} \qquad (8)$$

Eq. 7 and Eq. 8 are combined with the PSL equations, Eq. 3, to obtain the following equations that map the behavioral PSL equations to physical parameters:

$$m_j = \frac{V_{\mathrm{OUT}_j}}{(V_{\mathrm{DD}}/2)}, \quad W_{ji} = \frac{R_0}{R_{ji}}, \quad \beta = \frac{V_{\mathrm{DD}} R_{\mathrm{ref}}}{2 V_0 R_0} \qquad (9)$$

where R_0 is a unit resistor that is used to electrically change the inverse temperature β, and V_0 is a transistor dependent parameter ($\approx$ 40 mV) that defines the stochastic window of the p-bit (FIG. 5c). Depending on the sign of the interconnection, W_{ij}, the non-inverted output V_{OUT_j} or the inverted output $\overline{V}_{\mathrm{OUT}_j}$ is used for the synaptic connections.

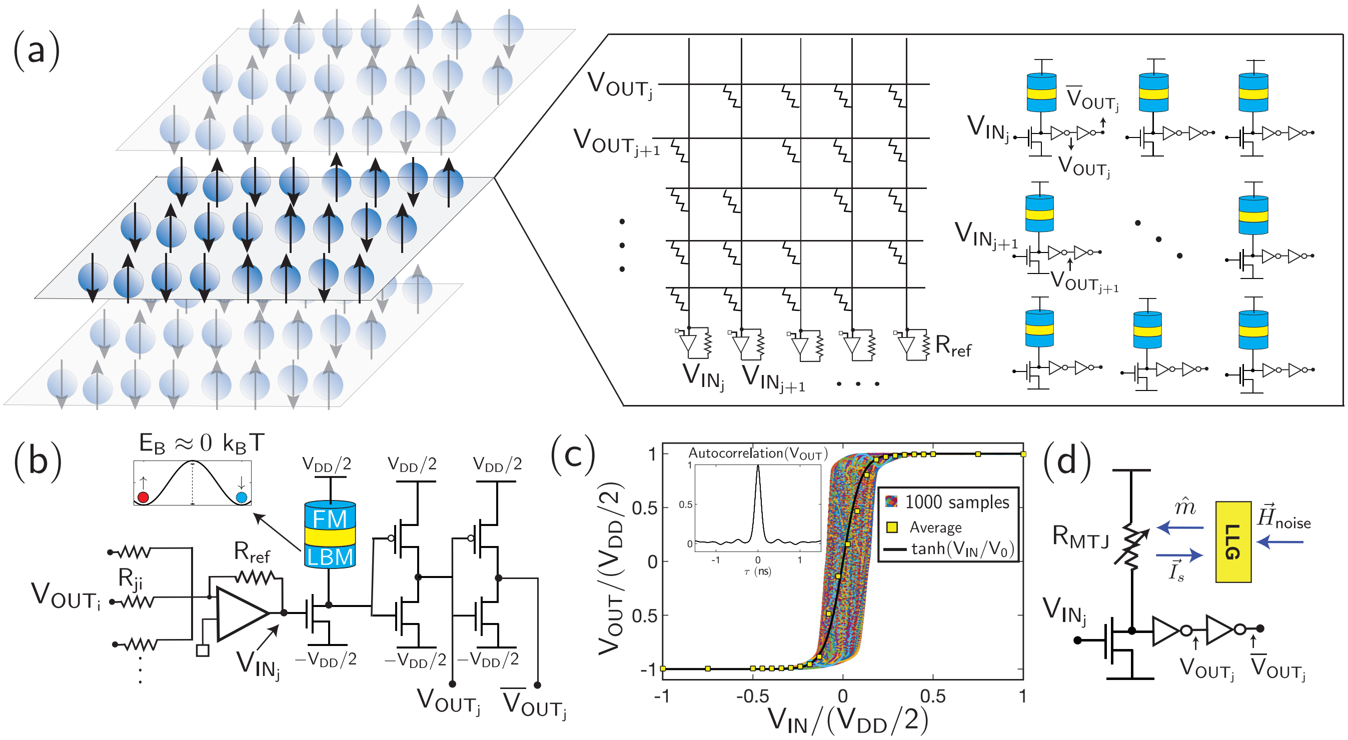

FIG. 5. **p-bit to Stochastic eMRAM:** (a) Each replica in the classical system is represented by a hardware neural network involving p-bits (neurons), interconnected by a resistive network (synapse). The outputs of the p-bits are weighted by the resistive network to become inputs to each other. Bias terms are added as fixed external voltage sources. (b) Detailed circuit schematics of a given p-bit and synapse following Ref. [44]: The outputs are collected by a fast operational amplifier (assumed ideal in circuit simulations). Fixed layer ferromagnet (FM) is a stable magnet with a large energy barrier, while the free layer is a circular low-barrier magnet (LBM) ($E_B \approx 0\ k_BT$) whose magnetization fluctuates in the presence of thermal noise. (c) SPICE simulations for the input-output characteristics of the p-bit: The results for 1000 p-bits where the input voltage is swept from $-V_{\mathrm{DD}}/2$ to $+V_{\mathrm{DD}}/2$, in $t_{sim} = 1$ ns (Inset shows the autocorrelation of the p-bit at $V_{\mathrm{IN}} = 0$). Each p-bit has a randomized resistance due to the random magnetization of the free layer, showing a range of outputs bounded by the parallel and anti-parallel resistance of the MTJ. Ensemble averaged output for 1000 samples at a given input voltage ($t_{\mathrm{sim}} = 2$ ns for each sample) shows a $\tanh(V_{\mathrm{IN}}/V_0)$ behavior. (d) The circuit model that self-consistently solves the stochastic LLG equation with the MTJ and transistor models. $\vec{I}_s$ is the spin-current exerted on the free layer due to the current polarized by the fixed layer, $\hat{m}$ is the instantaneous magnetization and $\vec{H}_{\mathrm{noise}}$ is the thermal noise field.

Device models

The 1T/1MTJ p-bit is modeled by combining a 14 nm-High Performance FinFET model from the open source Predictive Technology Models (PTM) [60] with a stochastic Landau-Lifshitz-Gilbert (sLLG) solver implemented in SPICE [61], following the design described in [44] (FIG. 5d). The MTJ is modeled as a simple conductor whose conductance depends on the instantaneous magnetization $m_z(t)$, provided by the sLLG such that

$$G_{\mathrm{MTJ}}(t) = G_0\left[1 + m_z(t)\frac{R_{AP} - R_P}{R_{AP} + R_P}\right] \qquad (10)$$

where R_P and R_{AP} are the parallel and antiparallel resistance of the MTJ and $G_0 = (R_{AP}^{-1} + R_P^{-1})/2$. We use an experimentally measured value for the tunneling magnetoresistance (TMR) $= (R_{AP} - R_P)/R_P = 110\%$ after Ref. [40]. G_0 is set equal to the transistor resistance at $V_{\mathrm{IN}} = 0$ to produce a symmetric sigmoid with no offsets, in this case $G_0^{-1} = 23.4$ kΩ. The free layer is assumed to be a circular low barrier nanomagnet [62, 63] with a diameter of 22 nm and thickness of 2 nm and a saturation magnetization of $M_s = 1100$ emu/cc, with a damping coefficient $\alpha = 0.01$, typical parameters for CoFeB [64].

The time dependent magnetization is obtained by solving sLLG equation in the monodomain approximation [65]:

$$(1+\alpha^2)\frac{d\hat{m}}{dt} = -|\gamma|\left(\hat{m}\times\vec{H}\right) - \alpha|\gamma|\left(\hat{m}\times\hat{m}\times\vec{H}\right) + \frac{1}{qN}\left(\hat{m}\times\vec{I}_S\times\hat{m}\right) + \frac{\alpha}{qN}\left(\hat{m}\times\vec{I}_S\right) \qquad (11a)$$

γ is the electron gyromagnetic ratio, q is electron charge and N is the number of Bohr magnetons (μ_B) in the volume of the magnet, $N = M_s\mathrm{Vol.}/\mu_B$. $\vec{H}$ contains the external magnetic and internal anisotropy fields of the magnet as well as the noise field. In the case of a circular nanomagnet without an easy-axis anisotropy,

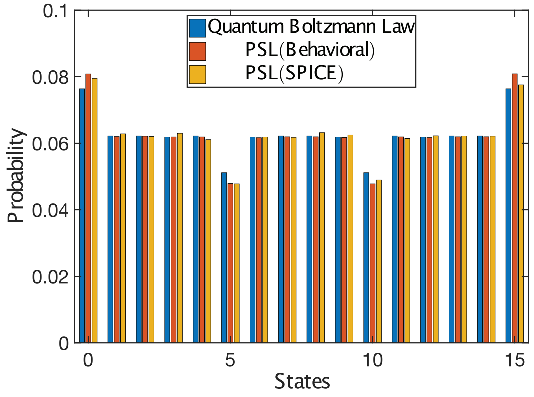

FIG. 6. Full circuit simulation of a 4-spin chain with 10 replicas: 1D classical Ising chain of $M = 4$ spins is simulated in SPICE with full device models for the stochastic MRAM-based p-bit and a resistive interconnection matrix and compared with PSL and Quantum Boltzmann Equation. $\beta = 0.5$, $\Gamma_x = 10$ and $J_{i,j} = +1$. The joint probability density is expressed in decimal numbers similar to the previous examples.

the total internal magnetic anisotropy becomes $\vec{H_m} = -4\pi M_s m_x \hat{x}$, where z-y is the easy plane of the magnet. The thermal noise is added in three directions ($\hat{x}$, $\hat{y}$ and $\hat{z}$) with zero mean and $\langle H^2_{\text{noise}} \rangle = 2\alpha k_B T / [(\gamma M_s \text{Vol.})]$ in units [Oe^2/s] [66].

Device operation

The p-bit shown in FIG. 5c is a series-resistance controlled device where the transistor resistance can be made much smaller or much larger compared to the fluctuating MTJ resistance. Therefore, the operation of the p-bit does not require manipulating the magnetization of the free layer unlike in standard spin-transfer-torque MRAM cells. However, the current flowing through the fixed layer of the MTJ produces a spin-polarized spin current that can unintentionally torque the magnet. We assume that this current is given as $\vec{I_s} = P I_{\text{MTJ}} \hat{z}$, where $\hat{z}$ is the fixed layer orientation and P is an interface polarization that can be related to TMR [67]. This spin-current is fed back to the sLLG solver and fully accounted for in the calculation of magnetization in our simulations, however for the circular LBM with a large demagnetization field used here, its effects are negligible [68]. Using these models, FIG. 5c shows transient SPICE simulations of a single p-bit output, V_{OUT} for 1000 samples where V_{IN} is rapidly swept in 2 ns. The range of stochastic outputs is bounded by a distribution of resistances ranging from R_P to R_{AP}. The ensemble average shows an approximate hyperbolic tangent behavior that allows the mapping shown in Eq. 7.

The inset of FIG. 5d shows the autocorrelation time of the circular in-plane magnet with a lifetime of ≈ 100 ps. The fluctuations for a circular magnet is expected to be faster compared to a magnet with perpendicular anisotropy due to the strong demagnetizing field that keeps the magnetization vector in the easy plane of the magnet [69]. The very short lifetime of such a circular low barrier magnet could allow very fast and efficient sampling times, as long as the interconnection network operates faster than these timescales. In present simulations, the resistive network operates instantaneously with an ideal operational amplifier therefore this requirement is met naturally, however in real implementations the synapse needs to be designed carefully.

The second requirement, the need for sequential updating of each p-bit is met naturally since the probability of simultaneous flips among p-bits is extremely unlikely, therefore hardware p-bits evolve autonomously without a synchronizing clock, effectively going through a random update order that does not affect their final distribution.

Stochastic MRAM-based p-bit vs Quantum Boltzmann Law

In FIG. 6, using full SPICE simulations for a 40 p-bit network we compute the joint probability density of a $M = 4$ spin ferromagnetic chain ($J_{i,j} = +1$) using 10 replicas, with $\beta = 0.5$ and $\Gamma_x = 10$. Unlike FIG. 2, no symmetry breaking field is applied and the network is asynchronously operated for $t_{sim} = 250$ ns, with a time step of 1 ps. All analog voltage values at the end of the SPICE simulation are thresholded (> 0 V $\equiv 1$, < 0 V $\equiv -1$) and a time-average is obtained similar to the PSL averaging after converting the state of each p-bit to binary and then to decimal. The results from the full device models seem to be in good agreement with the exact solution obtained from Eq. 4 and the behavioral PSL equations that are included for reference. Note the suppression of states $5 = (0101)_2$ and $10 = (1010)_2$ that correspond to the energetically unfavorable antiferromagnetic configurations $|{\downarrow}{\uparrow}{\downarrow}{\uparrow}\rangle$ and $|{\uparrow}{\downarrow}{\uparrow}{\downarrow}\rangle$, respectively. The agreement between the full SPICE models with the behavioral and exact solutions establishes the feasibility of the proposed quantum circuit emulator.

Projected performance improvement

In annealing algorithms, a key parameter is the time-to-solution (TTS) that is defined as the total time a solver requires to reach the desired answer of a problem with a predefined accuracy [30]. TTS clearly depends on the intrinsic hardware substrate that is used to implement the algorithm but also on the type of the problem and the required accuracy in the solution. Since the type of problem can have varied scaling properties with no generic answers [29, 30, 70], here we attempt to define a basic hardware unit, which is the time to provide a spin-flip attempt as defined by Eq. 3a. As shown in the inset of Fig. 5, the correlation time of an in-plane circular magnet can be as low as about 100 ps due to a precessional fluctuation mechanism found in such low barrier magnets [27]. Assuming a nearest neighbor 3D classical network that is mapped from a 2D quantum network, we assume that the synapse delay due to a crossbar structure can be much faster than magnetic fluctuations, ensuring all spin flip updates use up-to-date information and are useful. In such a scenario having $N = 10^6$ spins that are operating autonomously with a 0.1 to 1 ns correlation times, we project that the spin-flip rate can be 0.1 to 1 petaflips per second, which is orders of magnitude faster than present day CPU implementations [29] as well as parallelized GPU implementations [28]. A detailed projection with power estimates can be found in [71]. Finally, we note that such GHz rate fluctuations of low barrier magnets have been not only theoretically predicted [27] but also experimentally observed in in-plane magnets [72].

VII. CONCLUSION

We have presented a scalable, room-temperature quantum emulator using stochastic p-bits that can be built by a simple modification of the existing 1T/1MTJ cell of the eMRAM technology. The proposed emulator uses physical replicas for repeated Trotter slices used in software Quantum Monte Carlo methods. Having physical replicas for each slice could enable better scaling properties for quantum annealing compared to classical annealing as discussed in [17], since choosing the optimal number of replicas or probing each replica separately to find better energy minima is possible in a physically engineered design, unlike in real quantum systems [18]. The electrical control of annealing parameters, inverse temperature (β) and transverse field (Γ_x), could allow a very large number of q-bits to be reliably emulated with room temperature p-bits. Using conventional electronic devices such as GPU's or FPGA's to implement the synapses, it should be possible to engineer complicated interactions that extend beyond nearest neighbors and/or involve k-body interactions ($k > 2$). We note that even though the "sign problem" limits the universal use of our p-computer, a large number of practically relevant quantum systems could be efficiently emulated by it, considering a large number of optimization problems have been mapped on to the Transverse Ising Hamiltonian [73]. Our results provide a method of emulating quantum systems with probabilistic hardware in advance of a scalable universal quantum computer.

ACKNOWLEDGMENTS

KYC is grateful to Brian M. Sutton for helpful discussions. This work was supported in part by ASCENT, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

APPENDIX A : MAPPING QUANTUM HEISENBERG MODEL TO A CLASSICAL SYSTEM

The Heisenberg Hamiltonian emulated in Section IV is,

$$\mathcal{H}_Q = -\left(\sum_i^M J_z\sigma_i^z\sigma_{i+1}^z + J_x\sigma_i^x\sigma_{i+1}^x + J_y\sigma_i^y\sigma_{i+1}^y + \Gamma_x\sigma_i^x\right) \quad (12)$$

Following [16, 39], we divide this Hamiltonian into three non-commuting parts, i.e., $\mathcal{H}_Q = \mathcal{H}_0 + \mathcal{H}_1 + \mathcal{H}_2$, where

$$\mathcal{H}_0 = -\sum_{i=1}^M J_z\sigma_i^z\sigma_{i+1}^z \quad (13)$$

$$\mathcal{H}_1 = -\sum_{i=1,3,\cdots}^M \left(J_x\sigma_i^x\sigma_{i+1}^x + J_y\sigma_i^y\sigma_{i+1}^y\right) - \frac{1}{2}\sum_i^M \Gamma_x\sigma_i^x \quad (14)$$

$$\mathcal{H}_2 = -\sum_{i=2,4,\cdots}^M \left(J_x\sigma_i^x\sigma_{i+1}^x + J_y\sigma_i^y\sigma_{i+1}^y\right) - \frac{1}{2}\sum_i^M \Gamma_x\sigma_i^x \quad (15)$$

The n-th approximant of the Suzuki-Trotter transformation for this Hamiltonian is then given by

$$\begin{aligned} Z_Q^{(n)} = \sum_{\alpha_1,\alpha_2,\cdots,\alpha_{2n}} \Bigg[&\prod_{k=1}^{2n} e^{-\beta\mathcal{H}_0(\alpha_k)/2n} \\ &\times \prod_{k=1,3,\cdots}^{2n-1} \langle\alpha_k|e^{-\beta\mathcal{H}_1/n}|\alpha_{k+1}\rangle \\ &\times \prod_{k=2,4,\cdots}^{2n} \langle\alpha_k|e^{-\beta\mathcal{H}_2/n}|\alpha_{k+1}\rangle \Bigg] \end{aligned}$$

and the classical system becomes,

$$\begin{aligned} \mathcal{H}_{d+1} = &\sum_{k=1,2,3,\cdots}^{2n} \frac{1}{2n}\mathcal{H}_0(\alpha_k) \\ &- \frac{1}{\beta}\sum_{k=1,3,\cdots}^{2n-1} \ln\langle\alpha_k|e^{-\beta\mathcal{H}_1/n}|\alpha_{k+1}\rangle \\ &- \frac{1}{\beta}\sum_{k=2,4,\cdots}^{2n} \ln\langle\alpha_k|e^{-\beta\mathcal{H}_2/n}|\alpha_{k+1}\rangle \end{aligned} \quad (16)$$

with periodicity along $(d+1)^{\text{th}}$ dimension such that $|\alpha_{2n+1}\rangle = |\alpha_1\rangle$. Also notice that any k-th replica can also be written more explicitly using Dirac's bra-ket notation in terms of the constituent spins of that replica as $|\alpha_k\rangle = |m_{1,k}m_{2,k}\cdots m_{M,k}\rangle$ which is actually a $2^M \times 1$ column vector and $m_{i,j}$ denotes i-th spin of j-th replica.

Then the first summation on the right hand side of Eq. (16) can be written as

$$\sum_{k=1,2,3,\cdots}^{2n} \frac{1}{2n}\mathcal{H}_0(\alpha_k) = \sum_{k=1}^{2n}\sum_{i=1}^{M} \frac{J_z}{2n} m_{i,k}m_{i+1,k} + \frac{\Gamma_z}{2n} m_{i,k} \quad (17)$$

In order to evaluate $\ln\langle\alpha_k|e^{-\beta\mathcal{H}_1/n}|\alpha_{k+1}\rangle$ and $\ln\langle\alpha_k|e^{-\beta\mathcal{H}_2/n}|\alpha_{k+1}\rangle$, we start by repeatedly using the following identity of the Kronecker product:

$$e^{\mathbf{A}\otimes\mathbf{I}_B+\mathbf{I}_A\otimes\mathbf{B}} = e^{\mathbf{A}} \otimes e^{\mathbf{B}} \tag{18}$$

to write the following:

$$e^{-\beta\mathcal{H}_1/n} = e^{-\frac{\beta}{n}\zeta} \otimes e^{-\frac{\beta}{n}\zeta} \otimes \cdots \otimes e^{-\frac{\beta}{n}\zeta} \tag{19}$$

where

$$\begin{aligned}\zeta =& -J_x\left(\sigma^x\otimes\mathbf{I}_2\right)\left(\mathbf{I}_2\otimes\sigma^x\right) - J_y\left(\sigma^y\otimes\mathbf{I}_2\right)\left(\mathbf{I}_2\otimes\sigma^y\right)\\ &-\frac{1}{2}\Gamma_x\left(\sigma^x\otimes\mathbf{I}_2+\mathbf{I}_2\otimes\sigma^x\right)\end{aligned} \tag{20}$$

and is a 4 × 4 matrix representing a two-body Hamiltonian.

Here we note that $|\alpha_k\rangle$ can also be partitioned in terms of Kronecker products of two spin systems as

$$|\alpha_k\rangle = |m_{1,k}m_{2,k}\rangle \otimes |m_{3,k}m_{4,k}\rangle \otimes \cdots \otimes |m_{M-1,k}m_{M,k}\rangle. \tag{21}$$

With the definition of $|\alpha_k\rangle$ above in mind, we also repeatedly use another Kronecker product identity:

$$\left(\mathbf{A}\otimes\mathbf{B}\right)\left(\mathbf{C}\otimes\mathbf{D}\right) = \left(\mathbf{AC}\right)\otimes\left(\mathbf{BD}\right) \tag{22}$$

to write

$$\begin{aligned}&\langle\alpha_k|e^{-\beta\mathcal{H}_1/n}|\alpha_{k+1}\rangle\\ &=\prod_{i=1,3,\cdots}^{M}\langle m_{i,k}m_{i+1,k}|e^{-\beta\zeta/n}|m_{i,k+1}m_{i+1,k+1}\rangle\end{aligned} \tag{23}$$

where we have also made use of Eq. (19). Taking e-base logarithm on both sides, we finally get

$$\begin{aligned}&\ln\langle\alpha_k|e^{-\beta\mathcal{H}_1/n}|\alpha_{k+1}\rangle\\ &=\sum_{i=1,3,\cdots}^{M}\ln\langle m_{i,k}m_{i+1,k}|e^{-\beta\zeta/n}|m_{i,k+1}m_{i+1,k+1}\rangle.\end{aligned} \tag{24}$$

In a similar manner, we can also write,

$$\begin{aligned}&\ln\langle\alpha_k|e^{-\beta\mathcal{H}_2/m}|\alpha_{k+1}\rangle\\ &=\sum_{i=2,4,\cdots}^{M}\ln\langle m_{i,k}m_{i+1,k}|e^{-\beta\zeta/n}|m_{i,k+1}m_{i+1,k+1}\rangle.\end{aligned} \tag{25}$$

Next, we evaluate the 4×4 density matrix:

$$e^{-\beta\varsigma/n} = \begin{bmatrix} X_1 & X_5 & X_5 & X_2\\ X_5 & X_3 & X_4 & X_5\\ X_5 & X_4 & X_3 & X_5\\ X_2 & X_5 & X_5 & X_1\end{bmatrix} \tag{26}$$

The corresponding X_i are given by:

$$X = \sqrt{{\Gamma_x}^2 + {J_y}^2}$$

$$\begin{aligned}X_1 &= \frac{1}{2}e^{\frac{\beta J_x}{n}}\left[\cosh\left(\frac{\beta}{n}X\right)-\left(\frac{J_y}{X}\right)\sinh\left(\frac{\beta}{n}X\right)\right]+\frac{1}{2}e^{\frac{-\beta(J_x-J_y)}{n}}\\ X_2 &= \frac{1}{2}e^{\frac{\beta J_x}{n}}\left[\cosh\left(\frac{\beta}{n}X\right)-\left(\frac{J_y}{X}\right)\sinh\left(\frac{\beta}{n}X\right)\right]-\frac{1}{2}e^{\frac{-\beta(J_x-J_y)}{n}}\\ X_3 &= \frac{1}{2}e^{\frac{\beta J_x}{n}}\left[\cosh\left(\frac{\beta}{n}X\right)+\left(\frac{J_y}{X}\right)\sinh\left(\frac{\beta}{n}X\right)\right]+\frac{1}{2}e^{\frac{-\beta(J_x+J_y)}{n}}\\ X_4 &= \frac{1}{2}e^{\frac{\beta J_x}{n}}\left[\cosh\left(\frac{\beta}{n}X\right)+\left(\frac{J_y}{X}\right)\sinh\left(\frac{\beta}{n}X\right)\right]-\frac{1}{2}e^{\frac{-\beta(J_x+J_y)}{n}}\\ X_5 &= \frac{1}{2}e^{\frac{\beta J_x}{n}}\frac{\Gamma_x}{X}\sinh\left(\frac{\beta}{n}X\right).\end{aligned}$$

We then use the following energy relation (a justification of the energy model will be presented in Appendix B):

$$\begin{aligned}&-\frac{1}{\beta}\ln\langle m_{i,k}m_{i+1,k}|e^{-\beta\varsigma/n}|m_{i,k+1}m_{i+1,k+1}\rangle\\ &=2\epsilon - t_1\left(m_{i,k}m_{i,k+1}+m_{i+1,k}m_{i+1,k+1}\right)\\ &-t_2\left(m_{i,k}m_{i+1,k+1}+m_{i+1,k}m_{i,k+1}\right)\\ &-t_3\left(m_{i,k}m_{i+1,k}+m_{i,k+1}m_{i+1,k+1}\right)\\ &-t_4 m_{i,k}m_{i,k+1}m_{i+1,k}m_{i+1,k+1}\end{aligned} \tag{27}$$

where ϵ is a constant that we ignore and

$$\begin{aligned}t_0 &= \frac{1}{2n}J_z\\ t_1 &= \frac{1}{8\beta}\left(\ln X_1 - \ln X_2 + \ln X_3 - \ln X_4\right)\\ t_2 &= \frac{1}{8\beta}\left(\ln X_1 - \ln X_2 - \ln X_3 + \ln X_4\right)\\ t_3 &= \frac{1}{8\beta}\left(\ln X_1 + \ln X_2 - \ln X_3 - \ln X_4\right)\\ t_4 &= \frac{1}{8\beta}\left(\ln X_1 + \ln X_2 + \ln X_3 + \ln X_4 - 4\ln X_5\right)\end{aligned}$$

This corresponds to the energy model for the Heisenberg Hamiltonian as shown in Fig. 3.

APPENDIX B : JUSTIFICATION OF USING FORM OF THE ENERGY MODEL IN APPENDIX A

We start by simplifying the notation such that $m_{i,k}\equiv m_1$, $m_{i+1,k}\equiv m_2$, $m_{i,k+1}\equiv m_3$, and $m_{i+1,k+1}\equiv m_4$ and put different I_1 values for different configurations of $\{m_2, m_3, m_4\}$ into a truth table as shown in Table I with following definitions:

$$f_1 = \frac{1}{2}\ln\left(\frac{X_1}{X_5}\right) \tag{28}$$

$$f_2 = \frac{1}{2}\ln\left(\frac{X_5}{X_2}\right) \tag{29}$$

$$f_3 = \frac{1}{2}\ln\left(\frac{X_5}{X_4}\right) \tag{30}$$

$$f_4 = \frac{1}{2}\ln\left(\frac{X_3}{X_5}\right) \tag{31}$$

TABLE I. Truth table for I_1.

s_3	s_4	s_2	$I_1 = \dfrac{1}{2}\ln\left(\dfrac{P\left(m_1=+1\vert m_3,m_4,m_2\right)}{P\left(m_1=-1\vert m_3,m_4,m_2\right)}\right)$
1	1	1	f_1
1	1	0	f_2
1	0	1	f_3
1	0	0	f_4
0	1	1	$-f_4$
0	1	0	$-f_3$
0	0	1	$-f_2$
0	0	0	$-f_1$

We have also used the notation that $s_i \in \{0,1\}, (i \in \{1,2,3,4\})$ is the binary counterpart of the bipolar spin $m_i \in \{-1,1\}$.

In the binary representation, we can cast I_1 into the following form:

$$\begin{aligned}I_1 &= f_1 s_3 s_4 s_2 - f_1 \bar{s}_3 \bar{s}_4 \bar{s}_2 + f_2 s_3 s_4 \bar{s}_2 - f_2 \bar{s}_3 \bar{s}_4 s_2 \\ &\quad + f_3 s_3 \bar{s}_4 s_2 - f_3 \bar{s}_3 s_4 \bar{s}_2 + f_4 s_3 \bar{s}_4 \bar{s}_2 - f_4 \bar{s}_3 s_4 s_2\end{aligned} \tag{32}$$

We use the following two transformations to switch from s_i to m_i:

$$m_i \to \frac{1+s_i}{2} \tag{33}$$

$$\bar{m}_i \to \frac{1-s_i}{2}, \tag{34}$$

we can recast Eq.(32) into its bipolar form as

$$\begin{aligned}I_1 &= \frac{f_1}{8}\left[(1+m_3)(1+m_4)(1+m_2)-(1-m_3)(1-m_4)(1-m_2)\right] \\ &+ \frac{f_2}{8}\left[(1+m_3)(1+m_4)(1-m_2)-(1-m_3)(1-m_4)(1+m_2)\right] \\ &+ \frac{f_3}{8}\left[(1+m_3)(1-m_4)(1+m_2)-(1-m_3)(1+m_4)(1-m_2)\right] \\ &+ \frac{f_4}{8}\left[(1+m_3)(1-m_4)(1-m_2)(1-m_3)(1+m_4)(1+m_2)\right]\end{aligned} \tag{35}$$

Upon simplification and re-arrangement of the terms we get,

$$\begin{aligned}I_1 &= \left(\frac{f_1}{4}-\frac{f_2}{4}+\frac{f_3}{4}-\frac{f_4}{4}\right)m_2 \\ &+ \left(\frac{f_1}{4}+\frac{f_2}{4}+\frac{f_3}{4}+\frac{f_4}{4}\right)m_3 \\ &+ \left(\frac{f_1}{4}+\frac{f_2}{4}-\frac{f_3}{4}-\frac{f_4}{4}\right)m_4 \\ &+ \left(\frac{f_1}{4}-\frac{f_2}{4}-\frac{f_3}{4}+\frac{f_4}{4}\right)m_2 m_3 m_4\end{aligned} \tag{36}$$

Integrating Eq.(36) with respect to m_1 and multiplying by $\left(-\dfrac{1}{\beta}\right)$, we partially get the energy model

$$\begin{aligned}E = &-\frac{1}{\beta}\left(\frac{f_1}{4}-\frac{f_2}{4}+\frac{f_3}{4}-\frac{f_4}{4}\right)m_1 m_2 \\ &-\frac{1}{\beta}\left(\frac{f_1}{4}+\frac{f_2}{4}+\frac{f_3}{4}+\frac{f_4}{4}\right)m_1 m_3 \\ &-\frac{1}{\beta}\left(\frac{f_1}{4}+\frac{f_2}{4}-\frac{f_3}{4}-\frac{f_4}{4}\right)m_1 m_4 \\ &-\frac{1}{\beta}\left(\frac{f_1}{4}-\frac{f_2}{4}-\frac{f_3}{4}+\frac{f_4}{4}\right)m_1 m_2 m_3 m_4 \\ &+ K_1\left(m_2,m_3,m_4\right)\end{aligned} \tag{37}$$

where K_1 is a function of m_2, m_3 and m_4 but independent of m_1.

Defining,

$$\begin{aligned}t_3 &= \frac{1}{4\beta}\left(f_1-f_2+f_3-f_4\right) \\ &= \frac{1}{8\beta}\left(\ln X_1+\ln X_2-\ln X_3-\ln X_4\right)\end{aligned} \tag{38}$$

$$\begin{aligned}t_1 &= \frac{1}{4\beta}\left(f_1+f_2+f_3+f_4\right) \\ &= \frac{1}{8\beta}\left(\ln X_1-\ln X_2+\ln X_3-\ln X_4\right)\end{aligned} \tag{39}$$

$$\begin{aligned}t_2 &= \frac{1}{4\beta}\left(f_1+f_2-f_3-f_4\right) \\ &= \frac{1}{8\beta}\left(\ln X_1-\ln X_2-\ln X_3+\ln X_4\right)\end{aligned} \tag{40}$$

$$\begin{aligned}t_4 &= \frac{1}{4\beta}\left(f_1-f_2-f_3+f_4\right) \\ &= \frac{1}{8\beta}\left(\ln X_1+\ln X_2+\ln X_3+\ln X_4-4\ln X_5\right)\end{aligned} \tag{41}$$

we get the simplified energy expression:

$$\begin{aligned}E\left(m_1,m_2,m_3,m_4\right)\Big|_{m_1} &= -t_1 m_1 m_3 - t_2 m_1 m_4 \\ - t_3 m_1 m_2 - t_4 m_1 m_2 m_3 m_4 &+ K_1\left(m_2,m_3,m_4\right)\end{aligned} \tag{42}$$

Repeating the whole procedure for I_2, I_3 and I_4 separately, give us

$$\begin{aligned}K_1\left(m_2,m_3,m_4\right) &= -t_1 m_2 m_4 - t_2 m_2 m_3 \\ &\quad + K_2\left(m_3,m_4\right)\end{aligned} \tag{43}$$

$$K_2\left(m_3,m_4\right) = -t_3 m_3 m_4 + K_3\left(m_4\right) \tag{44}$$

$$K_3\left(m_4\right) = 0. \tag{45}$$

Putting Eqs.(42-45) together, gives us the energy model used in Eq.(27).

[1] Kerem Yunus Camsari, Rafatul Faria, Brian M Sutton, and Supriyo Datta, "Stochastic p-bits for invertible logic," Physical Review X **7**, 031014 (2017).

[2] Behtash Behin-Aein, Vinh Diep, and Supriyo Datta, "A building block for hardware belief networks," Scientific reports **6**, 29893 (2016).

[3] Richard P Feynman, "Simulating physics with computers," International journal of theoretical physics **21**, 467–488 (1982).

[4] Kerem Y Camsari, Brian M Sutton, and Supriyo Datta, "p-bits for probabilistic spin logic," arXiv preprint arXiv:1809.04028 (2018).

[5] Brian Sutton, Kerem Yunus Camsari, Behtash Behin-Aein, and Supriyo Datta, "Intrinsic optimization using stochastic nanomagnets," Scientific Reports **7**, 44370 (2017).

[6] Ahmed Zeeshan Pervaiz, Lakshmi Anirudh Ghantasala, Kerem Yunus Camsari, and Supriyo Datta, "Hardware emulation of stochastic p-bits for invertible logic," Scientific reports **7**, 10994 (2017).

[7] Roman Martoňák, Giuseppe E Santoro, and Erio Tosatti, "Quantum annealing of the traveling-salesman problem," Physical Review E **70**, 057701 (2004).

[8] Xinhua Peng, Zeyang Liao, Nanyang Xu, Gan Qin, Xianyi Zhou, Dieter Suter, and Jiangfeng Du, "Quantum adiabatic algorithm for factorization and its experimental implementation," Physical review letters **101**, 220405 (2008).

[9] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski, "A Learning Algorithm for Boltzmann Machines*," Cognitive Science **9**, 147–169 (1985).

[10] Rafatul Faria, Kerem Y Camsari, and Supriyo Datta, "Implementing bayesian networks with embedded stochastic mram," AIP Advances **8**, 045101 (2018).

[11] Ramtin Zand, Kerem Yunus Camsari, Steven D Pyle, Ibrahim Ahmed, Chris H Kim, and Ronald F DeMara, "Low-energy deep belief networks using intrinsic sigmoidal spintronic-based probabilistic neurons," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI* (ACM, 2018) pp. 15–20.

[12] Zhengbing Bian, Fabian Chudak, William G Macready, and Geordie Rose, "The ising model: teaching an old problem new tricks," D-wave systems **2** (2010).

[13] Steven H Adachi and Maxwell P Henderson, "Application of quantum annealing to training of deep neural networks," arXiv preprint arXiv:1510.06356 (2015).

[14] Jeremy Liu, Federico M Spedalieri, Ke-Thia Yao, Thomas E Potok, Catherine Schuman, Steven Young, Robert Patton, Garrett S Rose, and Gangotree Chamka, "Adiabatic quantum computation applied to deep learning networks," Entropy **20**, 380 (2018).

[15] Mohammad H Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko, "Quantum boltzmann machine," Physical Review X **8**, 021050 (2018).

[16] Masuo Suzuki, "Relationship between d-dimensional quantal spin systems and (d+ 1)-dimensional ising systems: Equivalence, critical exponents and systematic approximants of the partition function and spin correlations," Progress of theoretical physics **56**, 1454–1469 (1976).

[17] Giuseppe E Santoro, Roman Martoňák, Erio Tosatti, and Roberto Car, "Theory of quantum annealing of an ising spin glass," Science **295**, 2427–2430 (2002).

[18] Bettina Heim, Troels F Rønnow, Sergei V Isakov, and Matthias Troyer, "Quantum versus classical annealing of ising spin glasses," Science **348**, 215–217 (2015).

[19] Vasil S Denchev, Sergio Boixo, Sergei V Isakov, Nan Ding, Ryan Babbush, Vadim Smelyanskiy, John Martinis, and Hartmut Neven, "What is the computational value of finite-range tunneling?" Physical Review X **6**, 031015 (2016).

[20] Carlo Baldassi and Riccardo Zecchina, "Efficiency of quantum vs. classical annealing in nonconvex learning problems," Proceedings of the National Academy of Sciences , 201711456 (2018).

[21] T. Okuyama, M. Hayashi, and M. Yamaoka, "An ising computer based on simulated quantum annealing by path integral monte carlo method," in *2017 IEEE International Conference on Rebooting Computing (ICRC)* (2017) pp. 1–6.

[22] Tameem Albash and Daniel A. Lidar, "Adiabatic quantum computation," Rev. Mod. Phys. **90**, 015002 (2018).

[23] Matthias Troyer and Uwe-Jens Wiese, "Computational complexity and fundamental limitations to fermionic quantum monte carlo simulations," Physical review letters **94**, 170201 (2005).

[24] Ahmed Zeeshan Pervaiz, Brian M. Sutton, Lakshmi Anirudh Ghantasala, and Kerem Y. Camsari, "Weighted p-bits for FPGA implementation of probabilistic circuits," arXiv:1712.04166 [cs] (2017), arXiv: 1712.04166.

[25] Ramtin Zand, Kerem Y Camsari, Supriyo Datta, and Ronald F DeMara, "Composable probabilistic inference networks using mram-based stochastic neurons," ACM Journal on Emerging Technologies in Computing Systems (JETC) **15**, 17 (2019).

[26] Tameem Albash, Victor Martin-Mayor, and Itay Hen, "Temperature scaling law for quantum annealing optimizers," Physical review letters **119**, 110502 (2017).

[27] Orchi Hassan, Rafatul Faria, Kerem Yunus Camsari, Jonathan Z Sun, and Supriyo Datta, "Low barrier magnet design for efficient hardware binary stochastic neurons," IEEE Magnetics Letters (2019).

[28] Ye Fang, Sheng Feng, Ka-Ming Tam, Zhifeng Yun, Juana Moreno, Jagannathan Ramanujam, and Mark Jarrell, "Parallel tempering simulation of the three-dimensional edwards–anderson model with compact asynchronous multispin coding on gpu," Computer Physics Communications **185**, 2467–2478 (2014).

[29] Sergio Boixo, Troels F. Rønnow, Sergei V. Isakov, Zhihui Wang, David Wecker, Daniel A. Lidar, John M. Martinis, and Matthias Troyer, "Evidence for quantum annealing with more than one hundred qubits," Nature Physics **10**, 218–224 (2014).

[30] Tameem Albash and Daniel A. Lidar, "Demonstration of a scaling advantage for a quantum annealer over simulated annealing," Phys. Rev. X **8**, 031016 (2018).

[31] Hasitha Muthumala Waidyasooriya, Masanori Hariyama, Masamichi J Miyama, and Masayuki Ohzeki, "Opencl-based design of an fpga accelerator for quantum annealing simulation," The Journal of Supercomputing , 1–21

(2019).

[32] Wolfgang Lechner, Philipp Hauke, and Peter Zoller, "A quantum annealing architecture with all-to-all connectivity from local interactions," Science advances **1**, e1500838 (2015).

[33] Gemma De las Cuevas and Toby S Cubitt, "Simple universal models capture all classical spin physics," Science **351**, 1180–1183 (2016).

[34] JD Biamonte, "Nonperturbative k-body to two-body commuting conversion hamiltonians and embedding problem instances into ising spins," Physical Review A **77**, 052331 (2008).

[35] Shuxian Jiang, Keith A Britt, Travis S Humble, and Sabre Kais, "Quantum annealing for prime factorization," arXiv preprint arXiv:1804.02733 (2018).

[36] Tadashi Kadowaki and Hidetoshi Nishimori, "Quantum annealing in the transverse ising model," Phys. Rev. E **58**, 5355–5363 (1998).

[37] Pierre Pfeuty, "The one-dimensional ising model with a transverse field," ANNALS of Physics **57**, 79–90 (1970).

[38] Mark W Johnson, Mohammad HS Amin, Suzanne Gildert, Trevor Lanting, Firas Hamze, Neil Dickson, R Harris, Andrew J Berkley, Jan Johansson, Paul Bunyk, *et al.*, "Quantum annealing with manufactured spins," Nature **473**, 194 (2011).

[39] Mustansir Barma and B Sriram Shastry, "Classical equivalents of one-dimensional quantum-mechanical systems," Physical Review B **18**, 3351 (1978).

[40] CJ Lin, SH Kang, YJ Wang, K Lee, X Zhu, WC Chen, X Li, WN Hsu, YC Kao, MT Liu, *et al.*, "45nm low power cmos logic compatible embedded stt mram utilizing a reverse-connection 1t/1mtj cell," in *Electron Devices Meeting (IEDM), 2009 IEEE International* (IEEE, 2009) pp. 1–4.

[41] YJ Song, JH Lee, HC Shin, KH Lee, K Suh, JR Kang, SS Pyo, HT Jung, SH Hwang, GH Koh, *et al.*, "Highly functional and reliable 8mb stt-mram embedded in 28nm logic," in *Electron Devices Meeting (IEDM), 2016 IEEE International* (IEEE, 2016) pp. 27–2.

[42] D. Shum, D. Houssameddine, S. T. Woo, Y. S. You, J. Wong, K. W. Wong, C. C. Wang, K. H. Lee, K. Yamane, V. B. Naik, C. S. Seet, T. Tahmasebi, C. Hai, H. W. Yang, N. Thiyagarajah, R. Chao, J. W. Ting, N. L. Chung, T. Ling, T. H. Chan, S. Y. Siah, R. Nair, S. Deshpande, R. Whig, K. Nagel, S. Aggarwal, M. DeHerrera, J. Janesky, M. Lin, H. J. Chia, M. Hossain, H. Lu, S. Ikegawa, F. B. Mancoff, G. Shimon, J. M. Slaughter, J. J. Sun, M. Tran, S. M. Alam, and T. Andre, "Cmos-embedded stt-mram arrays in 2x nm nodes for gp-mcu applications," in *2017 Symposium on VLSI Technology* (2017) pp. T208–T209.

[43] Sabpreet Bhatti, Rachid Sbiaa, Atsufumi Hirohata, Hideo Ohno, Shunsuke Fukami, and SN Piramanayagam, "Spintronics based random access memory: A review," Materials Today (2017).

[44] Kerem Yunus Camsari, Sayeef Salahuddin, and Supriyo Datta, "Implementing p-bits with embedded mtj," IEEE Electron Device Letters **38**, 1767–1770 (2017).

[45] Hale F Trotter, "On the product of semi-groups of operators," Proceedings of the American Mathematical Society **10**, 545–551 (1959).

[46] E Gardner, "Multiconnected neural network models," Journal of Physics A: Mathematical and General **20**, 3453 (1987).

[47] Yuya Seki and Hidetoshi Nishimori, "Quantum annealing with antiferromagnetic transverse interactions for the hopfield model," Journal of Physics A: Mathematical and Theoretical **48**, 335301 (2015).

[48] Stuart Geman and Donald Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," IEEE Transactions on pattern analysis and machine intelligence , 721–741 (1984).

[49] Geoffrey E Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade* (Springer, 2012) pp. 599–619.

[50] Leo P Kadanoff and Gordon A Baym, "Quantum statistical mechanics: Green's function methods in equilibrium and nonequilibirum problems," (1962).

[51] Richard P Feynman, Robert B Leighton, and Matthew Sands, *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, Vol. 1 (Basic books, 2011).

[52] Sergey Bravyi, "Monte carlo simulation of stoquastic hamiltonians," arXiv preprint arXiv:1402.2295 (2014).

[53] Kerem Yunus Camsari, Rafatul Faria, Brian M. Sutton, and Supriyo Datta, "Stochastic p -Bits for Invertible Logic," Physical Review X **7** (2017), 10.1103/PhysRevX.7.031014.

[54] Fabio L Traversa and Massimiliano Di Ventra, "Polynomial-time solution of prime factorization and np-complete problems with digital memcomputing machines," Chaos: An Interdisciplinary Journal of Nonlinear Science **27**, 023107 (2017).

[55] Sean C Smithson, Naoya Onizawa, Brett H Meyer, Warren J Gross, and Takahiro Hanyu, "Efficient cmos invertible logic using stochastic computing," IEEE Transactions on Circuits and Systems I: Regular Papers **66**, 2263–2274 (2019).

[56] Zheng Zhu, Andrew J Ochoa, and Helmut G Katzgraber, "Efficient cluster algorithm for spin glasses in any space dimension," Physical review letters **115**, 077201 (2015).

[57] Shimeng Yu, Yi Wu, Rakesh Jeyasingh, Duygu Kuzum, and H-S Philip Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," IEEE Transactions on Electron Devices **58**, 2729–2737 (2011).

[58] Miao Hu, John Paul Strachan, Zhiyong Li, Emmanuelle M Grafals, Noraica Davila, Catherine Graves, Sity Lam, Ning Ge, Jianhua Joshua Yang, and R Stanley Williams, "Dot-product engine for neuromorphic computing: programming 1t1m crossbar to accelerate matrix-vector multiplication," in *Proceedings of the 53rd annual design automation conference* (ACM, 2016) p. 19.

[59] Peter L McMahon, Alireza Marandi, Yoshitaka Haribara, Ryan Hamerly, Carsten Langrock, Shuhei Tamate, Takahiro Inagaki, Hiroki Takesue, Shoko Utsunomiya, Kazuyuki Aihara, *et al.*, "A fully programmable 100-spin coherent ising machine with all-to-all connections," Science **354**, 614–617 (2016).

[60] "Predictive Technology Model (PTM) (http://ptm.asu.edu/),".

[61] M. M. Torunbalci, P. Upadhyaya, S. A. Bhave, and K. Y. Camsari, "Modular compact modeling of mtj devices," IEEE Transactions on Electron Devices , 1–7 (2018).

[62] R. P. Cowburn, D. K. Koltsov, A. O. Adeyeye, M. E. Welland, and D. M. Tricker, "Single-domain circular nanomagnets," Physical Review Letters **83**, 1042 (1999).

[63] Punyashloka Debashis, Rafatul Faria, Kerem Yunus

Camsari, and Zhihong Chen, "Designing stochastic nanomagnets for probabilistic spin logic," IEEE Magnetics Letters (2018).

[64] Jack C Sankey, Yong-Tao Cui, Jonathan Z Sun, John C Slonczewski, Robert A Buhrman, and Daniel C Ralph, "Measurement of the spin-transfer-torque vector in magnetic tunnel junctions," Nature Physics **4**, 67 (2008).

[65] Jonathan Z Sun, "Spin-current interaction with a monodomain magnetic body: A model study," Physical Review B **62**, 570 (2000).

[66] Z Li and S Zhang, "Thermally assisted magnetization reversal in the presence of a spin-transfer torque," Physical Review B **69**, 134416 (2004).

[67] Deepanjan Datta, Behtash Behin-Aein, Supriyo Datta, and Sayeef Salahuddin, "Voltage asymmetry of spin-transfer torques," IEEE Transactions on Nanotechnology **11**, 261–272 (2012).

[68] Rafatul Faria, Kerem Yunus Camsari, and Supriyo Datta, "Low-barrier nanomagnets as p-bits for spin logic," IEEE Magnetics Letters **8**, 1–5 (2017).

[69] L Lopez-Diaz, L Torres, and E Moro, "Transition from ferromagnetism to superparamagnetism on the nanosecond time scale," Physical Review B **65**, 224406 (2002).

[70] Ravi Montenegro, Prasad Tetali, *et al.*, "Mathematical aspects of mixing times in markov chains," Foundations and Trends® in Theoretical Computer Science **1**, 237–354 (2006).

[71] Brian Sutton, Rafatul Faria, Lakshmi A Ghantasala, Kerem Y Camsari, and Supriyo Datta, "Autonomous probabilistic coprocessing with petaflips per second," arXiv preprint arXiv:1907.09664 (2019).

[72] Matthew R Pufall, William H Rippard, Shehzaad Kaka, Steven E Russek, Thomas J Silva, Jordan Katine, and Matt Carey, "Large-angle, gigahertz-rate random telegraph switching induced by spin-momentum transfer," Physical Review B **69**, 214409 (2004).

[73] Andrew Lucas, "Ising formulations of many np problems," Frontiers in Physics **2**, 5 (2014).