

CMOS + stochastic nanomagnets: heterogeneous computers for probabilistic inference and learning

Nihal Sanjay Singh,^{1,*} Keito Kobayashi,^{1,2,3,*} Qixuan Cao,^{1,*} Kemal Selcuk,¹ Tianrui Hu,¹ Shaila Niazi,¹ Navid Anjum Aadit,¹ Shun Kanai,^{2,3,4,5,6,7,9} Hideo Ohno,^{2,4,5,8} Shunsuke Fukami,^{2,3,4,5,8,10,†} and Kerem Y. Camsari^{1,‡}

¹*Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, 93106, CA, USA*

²*Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan*

³*Graduate School of Engineering, Tohoku University, 6-6 Aramaki Aza Aoba, Aoba-ku, Sendai 980-0845, Japan*

⁴*WPI Advanced Institute for Materials Research (WPI-AIMR),*

Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

⁵*Center for Science and Innovation in Spintronics (CSIS),*

Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

⁶*PRESTO, Japan Science and Technology Agency (JST), Kawaguchi 332-0012, Japan*

⁷*Division for the Establishment of Frontier Sciences of Organization for Advanced Studies at Tohoku University, Tohoku University, Sendai 980-8577, Japan*

⁸*Center for Innovative Integrated Electronic Systems (CIES), Tohoku University, 468-1 Aramaki Aza Aoba, Aoba-ku, Sendai 980-0845, Japan*

⁹*National Institutes for Quantum Science and Technology, Takasaki 370-1207, Japan*

¹⁰*Inamori Research Institute of Science (InaRIS), Kyoto 600-8411, Japan*

Extending Moore’s law by augmenting complementary-metal-oxide semiconductor (CMOS) transistors with emerging nanotechnologies (X) has become increasingly important. One important class of problems involve sampling-based Monte Carlo algorithms used in probabilistic machine learning, optimization, and quantum simulation. Here, we combine stochastic magnetic tunnel junction (sMTJ)-based probabilistic bits (p-bits) with Field Programmable Gate Arrays (FPGA) to create an energy-efficient CMOS + X (X = sMTJ) prototype. This setup shows how asynchronously driven CMOS circuits controlled by sMTJs can perform probabilistic inference and learning by leveraging the algorithmic update-order-invariance of Gibbs sampling. We show how the stochasticity of sMTJs can augment low-quality random number generators (RNG). Detailed transistor-level comparisons reveal that sMTJ-based p-bits can replace up to 10,000 CMOS transistors while dissipating two orders of magnitude less energy. Integrated versions of our approach can advance probabilistic computing involving deep Boltzmann machines and other energy-based learning algorithms with extremely high throughput and energy efficiency.

With the slowing down of Moore’s Law [1], there has been a growing interest in domain-specific hardware and architectures to address emerging computational challenges and energy-efficiency, particularly borne out of machine learning and AI applications. One promising approach is the co-integration of traditional complementary metal-oxide semiconductor (CMOS) technology with emerging nanotechnologies (X), resulting in CMOS + X architectures. The primary objective of this approach is to augment existing CMOS technology with novel functionalities, by enabling the development of physics-inspired hardware systems that realize energy-efficiency, massive parallelism, and asynchronous dynamics, and apply them to a wide range of problems across various domains.

Being named one of the top 10 algorithms of the 20th century [2], Monte Carlo methods have been one of the most effective approaches in computing to solve computationally hard problems in a wide range of applications, from probabilistic machine learning, optimization to quantum

simulation. Probabilistic computing with p-bits [3] has emerged as a powerful platform for executing these Monte Carlo algorithms in massively parallel [4, 5] and energy-efficient architectures. p-bits have been shown to be applicable to a large domain of computational problems from combinatorial optimization to probabilistic machine learning and quantum simulation [6–8].

Several p-bit implementations that use the inherent stochasticity in different materials and devices have been proposed, based on diffusive memristors [9], resistive RAM [10], perovskite nickelates [11], ferroelectric transistors [12], single photon avalanche diodes [13], optical parametric oscillators [14] and others. Among alternatives sMTJs built out of low-barrier nanomagnets have demonstrated significant potential due to their ability to amplify noise, converting millivolts of fluctuations to hundreds of millivolts over resistive networks [15], unlike alternative approaches with amplifiers [16]. Another advantage of sMTJ-based p-bits is the continuous generation of truly random bitstreams without the need to be reset in synchronous pulse-based designs [17, 18]. The possibility of designing energy-efficient p-bits using low-barrier nanomagnets has stimulated renewed interest in material and device research with several exciting demonstrations from nanosecond fluctuations [19–21] to better theoretical understanding of nanomagnet physics [22–25] and novel magnetic tunnel

* These authors contributed equally

† shunsuke.fukami.c8@tohoku.ac.jp

‡ camsari@ece.ucsb.edu

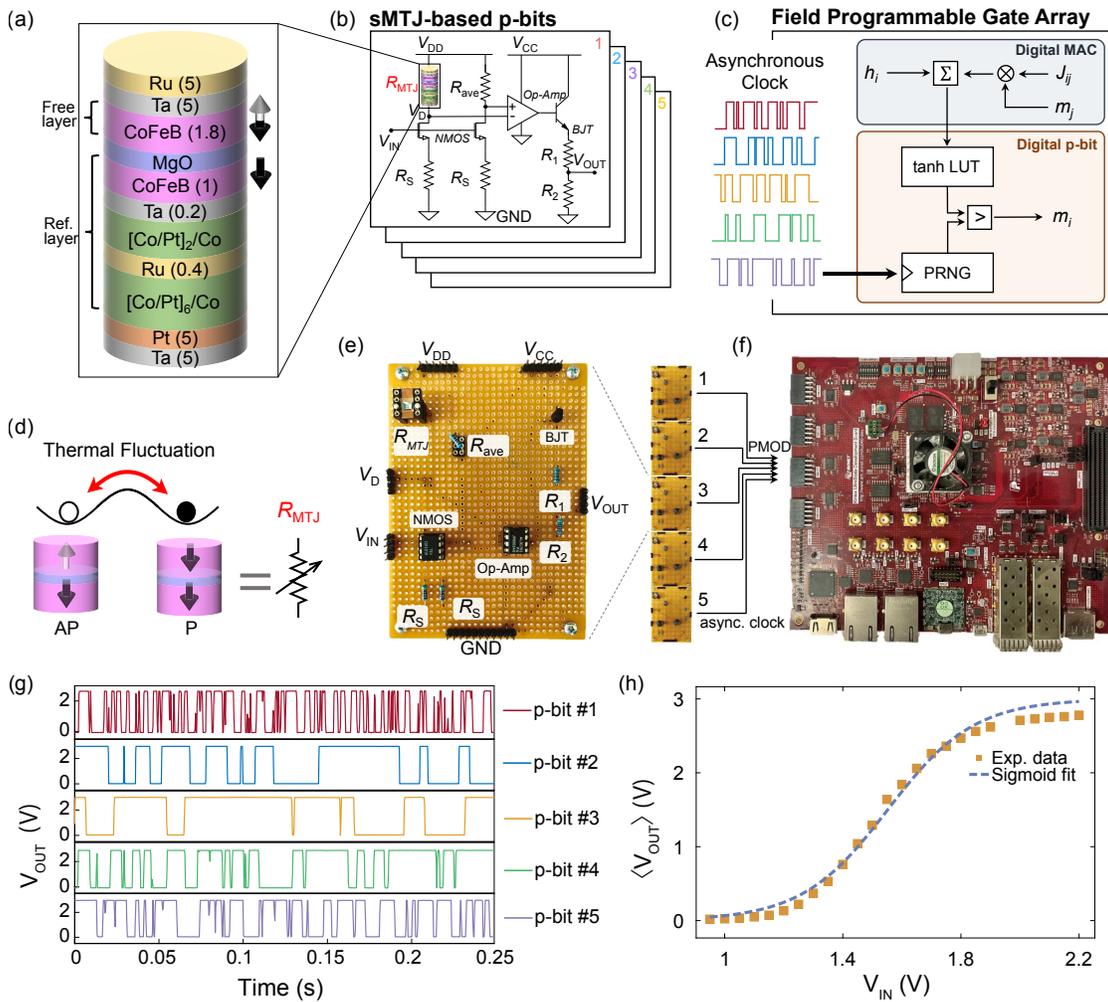


Fig. 1: Experimental setup for the CMOS + SMTJ probabilistic computer. (a) Stack structure of the stochastic magnetic tunnel junction (SMTJ). (b) The proposed SMTJ-based p-bit circuit with two branches whose outputs are provided to an operational amplifier. R_{ave} is the average resistance of R_P and R_{AP} of the SMTJ. 5 SMTJ-based p-bits provide tunable, truly random and asynchronous clocks to a digital field programmable gate array (FPGA). (c) Digital p-bits in the FPGA use lookup tables (LUT), comparators, synaptic weights, and pseudorandom number generators (PRNG). The clocks of the PRNG are driven by the truly random asynchronous outputs coming from the analog p-bits. (d) Pictorial representation of perpendicular SMTJ. (e) Image of a single p-bit circuit. (f) Image of the FPGA. The asynchronous clocks are input through the peripheral module (PMOD) pins. (g) Typical output of p-bits #1 to 5 using 5 SMTJs obtained from the p-bit circuit (see Supplementary Section III), showing variations in fluctuations. (h) Experimentally measured $\langle V_{OUT} \rangle$ the time average (over a period of 3 minutes) of the p-bit circuit output, as a function of DC input voltage V_{IN} . The yellow squares are experimental data, and the blue dashed line is a fit of the form $\langle V_{OUT} \rangle = 1/2 V_{CC}' [\tanh[\beta(V_{IN} - V_0)] + 1]$, where $V_0 = 1.55$ V, $\beta = 3.43$ V $^{-1}$, $V_{CC}' = 3$ V is a reduced voltage from $V_{CC} = 5$ V (see Supplementary Section IV).

junction designs [26, 27].

Despite promising progress with hardware prototypes [28–32], large-scale probabilistic computing using stochastic nanodevices remains elusive. As we will establish in this paper, designing purely CMOS-based high-performance probabilistic computers suited to sampling and optimization problems is prohibitive beyond a certain scale (>1 M p-bits) due to the large area and energy costs of pseudorandom number generators. As such, any large-scale integration of probabilistic computing will involve strong integration with

CMOS technology in the form of CMOS+X architectures. Given the unavoidable device-to-device variability, the interplay between continuously fluctuating stochastic nanodevices (e.g., SMTJs) with deterministic CMOS circuits and possible applications of such hybrid circuits remain unclear.

In this paper, we first introduce the notion of a heterogeneous CMOS+SMTJ system where the asynchronous dynamics of SMTJs control digital circuits in a standard CMOS Field Programmable Gate Array (FPGA). We view

the FPGA as a “drop-in replacement” for eventual integrated circuits where sMTJs could be situated on top of CMOS. Unlike earlier implementations where sMTJs were primarily used to implement neurons and CMOS or analog components circuits for synapses [28, 29], we design hybrid circuits where sMTJ-based p-bits control a large number of digital circuits residing in the FPGA without dividing the system into neurons (sMTJ) and synapses (CMOS). We show how the true randomness injected into deterministic CMOS circuits augment low-quality random number generators based on linear feedback shift registers (LFSR). This result represents an example of how sMTJs could be used to reduce footprint and energy consumption in the CMOS underlayer. In this work, we present a small example of a CMOS + sMTJ system, however, similar systems can be scaled up to much bigger densities, leveraging the proven manufacturability of magnetic memory at gigabit densities. Our results will help lay the groundwork for larger implementations in the presence of unavoidable device-to-device variations. We also focus beyond the common use case of combinatorial optimization of similar physical computers [33], considering probabilistic inference and learning in deep energy-based models.

Specifically, we use our system to train 3-hidden 1-visible layer deep and unrestricted Boltzmann machines that entirely rely on the asynchronous dynamics of the stochastic MTJs. Second, we evaluate the quality of randomness directly at the application level through probabilistic inference and deep Boltzmann learning. This approach contrasts with the majority of related work, which typically conducts statistical tests at the single device level to evaluate the quality of randomness [21, 34–38] (see Supplementary Sections VIII, XI, and XII for more randomness experiments). As an important new result, we find that the quality of randomness matters in machine learning tasks as opposed to optimization tasks that have been explored previously. And finally, we conduct a comprehensive benchmark using an experimentally calibrated 7-nm CMOS PDK and find that when the quality of randomness is accounted for, the sMTJ-based p-bits are about 4 orders of magnitude smaller in area and they dissipate 2 orders of magnitude less energy, compared to CMOS p-bits. We envision that large-scale CMOS+X p-computers ($\gg 10^5$) can be a reality in scaled up versions of the CMOS + sMTJ type computers we discuss in this work.

Constructing the heterogeneous p-computer

FIG. 1 shows a broad overview of our sMTJ-FPGA setup along with device and circuit characterization of sMTJ p-bits. Unlike earlier p-bit demonstrations with sMTJs as standalone stochastic binary neurons, in this work, we use sMTJ-based p-bits to generate asynchronous and truly random clock sources to drive digital p-bits in the FPGA (FIG. 1a,b,c).

The conductance of the sMTJ depends on the relative angle θ between the free and the fixed layers, $G_{\text{MTJ}} \propto [1 + P^2 \cos(\theta)]$, where P is the interfacial spin polarization. When the free layer is made out of a low barrier

nanomagnet θ becomes a random variable in the presence of thermal noise, causing conductance fluctuations between the parallel (P) and the antiparallel (AP) states (FIG. 1d).

The five sMTJs used in the experiment are designed with a diameter of 50 nm and have a relaxation time of about 1 to 20 ms, with energy barriers of $\approx 14\text{-}17 k_B T$, assuming an attempt time of 1 ns [40] (see Supplementary Section II). In order to convert these conductance fluctuations into voltages, we design a new p-bit circuit (FIG. 1b,e). This circuit creates a voltage comparison between two branches controlled by two transistors, fed to an operational amplifier. As we discuss in Supplementary Section III, the main difference of this circuit compared to the earlier 3 transistor/1MTJ design used in earlier demonstrations [28, 29] is in its ability to provide a larger stochastic window to tune the p-bit (FIG. 1h) with more variation tolerance (see Supplementary Section IV).

FIG. 1f,g show how the asynchronous clocks obtained from p-bits with 50/50 fluctuations are fed to the FPGA. Inside the FPGA, we design a digital probabilistic computer where a p-bit includes a lookup table (LUT) for the hyperbolic tangent function, a pseudorandom number generator (PRNG) and a digital comparator (see Supplementary Section V).

The crucial link between analog p-bits and the digital FPGA is established through the clock of the PRNG used in the FPGA, where a multitude of digital p-bits can be asynchronously driven by analog p-bits. As we discuss in Sections 3-4, depending on the quality of the chosen PRNG, the injection of additional entropy through the clocks has a considerable impact on inference and learning tasks. The potential for enhancing low-quality PRNGs using compact and scalable nanotechnologies, such as sMTJs, which can be integrated as a BEOL (Back-End-Of-Line) process on top of the CMOS logic, holds significant promise for future CMOS + sMTJ architectures.

RESULTS

Probabilistic inference with heterogeneous p-computers

In the p-bit formulation, we define probabilistic inference as generating samples from a specified distribution which is the Gibbs-Boltzmann distribution for a given network (see Supplementary Section I for details). This is a computationally hard problem [41], and is at the heart of many important applications involving Bayesian inference [42], training probabilistic models in machine learning [43], statistical physics [44] and many others [45]. Due to the broad applicability of probabilistic inference, improving key figures of merit such as probabilistic flips per second (sampling throughput) and energy-delay product for this task are extremely important.

To demonstrate this idea, we evaluate probabilistic inference on a probabilistic version of the full adder (FA) [39] as shown in FIG. 2a. The truth table of the FA is given in FIG. 2b. The FA performs 1-bit binary addition and it has three inputs (A, B, Carry in= C_{in}) and two outputs (Sum=S, and Carry out= C_{out}). The probabilistic FA can be described in a 5 p-bit, fully-connected network (FIG. 2a). When the network samples from its equilibrium, it samples

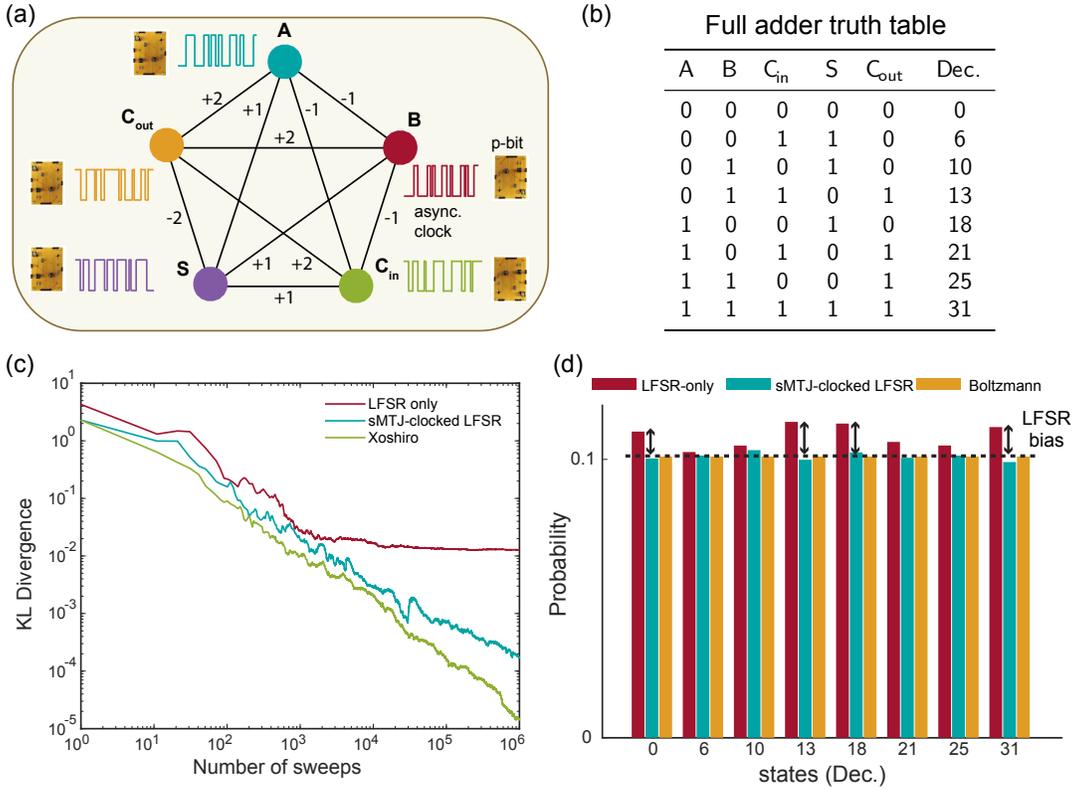


Fig. 2: Inference on a probabilistic full adder. (a) Fully-connected full adder network [39], where p-bits are clocked by the sMTJs. (b) Truth table of the full adder where Dec. represents the decimal representation of the state of [A B C_{in} S C_{out}] from left to right. (c) Kullback-Leibler (KL) divergence between the ideal and measured distributions vs. the number of sweeps. Results are shown for LFSR-based p-bit (red line), sMTJ-clocked LFSR-based p-bit (blue line), and Xoshiro-based p-bit (green line). (d) Histogram for the measured and ideal distributions at the 10⁶ sweep. The red, blue, and yellow bars show LFSR, sMTJ-clocked LFSR, and Boltzmann distribution, respectively. The histogram shows all 8 high probability states denoted in (b) and with a clear bias for the LFSR distribution (see Supplementary Section VII for full histograms for all PRNGs including Xoshiro).

states corresponding to the truth table, according to the Boltzmann distribution.

We demonstrate probabilistic sampling on the probabilistic FA using the digital p-bits with standalone Linear Feedback Shift Registers LFSRs (only using the FPGA), sMTJ-clocked LFSRs (using sMTJ-based p-bits and the FPGA), and standalone Xoshiro RNGs (only using the FPGA). Our main goal is to compare the quality of randomness obtained by inexpensive but low-quality PRNGs such as LFSRs [46] with sMTJ-enhanced LFSRs and high-quality but expensive PRNGs such as Xoshiro [47] (see Supplementary Section VI).

FIG. 2c shows the comparison of these three different solvers where we measure the Kullback-Leibler (KL) divergence [48] between the cumulative distribution based on the number of sweeps and the ideal Boltzmann distribution of the FA:

$$\text{KL}[P_{\text{exp}} || P_{\text{ideal}}] = \sum_x P_{\text{exp}}(x) \log \frac{P_{\text{exp}}(x)}{P_{\text{ideal}}(x)}, \quad (1)$$

where P_{exp} is the probability obtained from the experiment (cumulatively measured) and P_{ideal} is the probability obtained

from the Boltzmann distribution. For LFSR (red line), the KL divergence saturates when the number of sweeps exceeds $N = 10^4$, while for sMTJ-clocked LFSR (blue line) and Xoshiro (green line), the KL divergence decreases with increasing the number of sweeps. The persistent bias of the LFSR is also visible in the partial histogram of probabilities measured at $N = 10^6$ sweeps as shown in FIG. 2d (see Supplementary Section VII for the full histograms). It is important to note here in our present context where sMTJs are limited to a handful of devices, we use sMTJ-based p-bits used to drive low-quality LFSRs, observing how they perform similar to high-quality PRNGs. In integrated implementations however, sMTJ-based p-bits can be directly used as p-bits themselves, without any supporting PRNG (see Supplementary Section XVI for details on projections of integrated implementations).

The mechanism of how the sMTJ-clocked LFSRs produce random numbers is interesting: even though the next bit in an LFSR is always perfectly determined, the randomness in the arrival times of clocks from the sMTJs makes their output unpredictable. Over the course of the full network's

evolution, each LFSR produces an unpredictable bitstream, functioning as truly random bits.

The observed bias of the LFSR can be due to several reasons: first, the LFSRs generally provide low-quality random numbers and do not pass all the tests in the NIST statistical test suite [49] (see Supplementary Section XII). Second, we take whole words of random bits from the LFSR to generate large random integers. This is a known danger when using LFSRs [50, 51], which can be mitigated by the use of phase shifters that scramble the parallelly obtained bits to reduce their correlation [52]. But such measures increase the complexity of PRNG designs further limiting the scalable implementation of digital p-computers (see Supplementary Section XI for detailed experimental analysis of LFSR bias).

The quality of randomness in Monte Carlo sampling is a rich and well-studied subject (see, for example, [53–55]). The main point we stress in this work is that even compact and inexpensive simple PRNGs can perform as well as sophisticated, high-quality RNGs when augmented by truly random nanodevices such as sMTJs.

Boltzmann Learning with heterogeneous p-computers

We now show how to train deep Boltzmann machines (DBM) with our heterogeneous sMTJ + FPGA-based computer. Unlike probabilistic inference, in this setting, the weights of the network are unknown and the purpose of the training process is to obtain desired weights for a given truth table, such as the full adder (see Supplementary Section IX for an example of arbitrary distribution generation using the same learning algorithm). We consider this demonstration as a proof-of-concept for eventual larger-scale implementations (FIG. 3a,b). Similar to probabilistic inference, we compare the performance of three solvers: LFSR-based, Xoshiro-based and sMTJ+LFSR-based RNGs. We choose a 32-node Chimera lattice [56] to train a probabilistic full adder with 5 visible nodes and 27 hidden nodes in a 3-layer DBM (see FIG. 3b top panel). Note that this deep network is significantly harder to train than training fully-visible networks whose data correlations are known a priori [29], necessitating positive and negative phase computations (see Supplementary Section VII and Algorithm 1 for details on the learning algorithm and implementation).

FIG. 3c,d show the KL divergence and the probability distribution of the full adder Boltzmann machines based on the fully digital LFSR/Xoshiro and the heterogeneous sMTJ-clocked LFSR RNGs. The KL divergence in the learning experiment is performed like this: after each epoch during training, we save the weights in the classical computer and perform probabilistic inference to measure the KL distance between the learned and ideal distributions. The sMTJ-clocked LFSR and the Xoshiro based Boltzmann machines produce probability distributions that eventually closely approximate the Boltzmann distribution of the full adder. On the other hand, the fully digital LFSR based Boltzmann machine produces the incorrect states $[A B C_{in} S C_{out}] = 2$ and 29 with a significantly higher probability than the correct peaks, and grossly underestimates the probabilities of states

0, 6, 18, and 25 (see FIG. Supplementary Figure 4 for full histograms that are avoided here for clarity). As in the inference experiment (FIG. 2a), the KL divergence of the LFSR saturates and never improves beyond a point. The increase in the KL divergence for Xoshiro and sMTJ-clocked LFSR towards the end is related to hyperparameter selection and unrelated to RNG quality [57]. For this reason, we select the weights at epoch=400 for testing to produce the histogram in FIG. 3d.

In line with our previous results, the learning experiments confirm the inferior quality of LFSR-based PRNGs, particularly for learning tasks (see Supplementary Section X for an MNIST training comparisons between p-bits based on Xoshiro and LFSR). While LFSRs can produce correct peaks with some bias in optimization problems, they fail to learn appropriate weights for sampling and learning, rendering them unsuitable for these applications. In addition to these results, statistical tests on the NIST test suite corroborate our findings that sMTJ-clocked LFSRs and high-quality PRNGs such as Xoshiro outperform the pure LFSR-based p-bits (see Supplementary Section XII).

Our learning result demonstrates how asynchronously interacting p-bits can creatively combine with existing CMOS technology. Scaled and integrated implementations of this concept could lead to a resurgence in training powerful deep Boltzmann machines [58].

Energy and transistor count comparisons

Given our prior results stressing how the quality of randomness can play a critical role in probabilistic inference and learning, it is beneficial to perform precise, quantitative comparisons with the various digital PRNGs we built in hardware FPGAs with sMTJ-based p-bits [15]. Note that for this comparison, we do not consider augmented CMOS p-bits, but directly compare sMTJ-based mixed signal p-bits with their digital counterparts (see Supplementary Section XVI for details on projections of integrated implementations using sMTJ-based mixed signal p-bits). Moreover, instead of benchmarking the voltage comparator based p-bit circuit shown in FIG. 1 or other types of spin-orbit torque based p-bits [3, 59], we benchmark the 3T/1MTJ based p-bit first reported in [15]. The reason for this choice is that this design allows the use of fast in-plane sMTJs whose fluctuations can be as fast as micro to nanoseconds. We also note that the table-top components we use in this work are not optimized but used for convenience.

For the purpose of benchmarking and characterization, we synthesize circuits for LFSR and Xoshiro PRNGs and these PRNG-based p-bits using the ASAP 7nm Predictive process development kit (PDK) that uses SPICE-compatible FinFET device models [60]. Our synthesis flow, explained in detail in Supplementary Section XIII, starts from hardware description level (HDL) coding of these PRNGs and leads to transistor-level circuits using the experimentally benchmarked ASAP 7nm PDK. As such, the analysis we perform here offers a high degree of precision in terms of transistor counts and quantitative energy consumption.

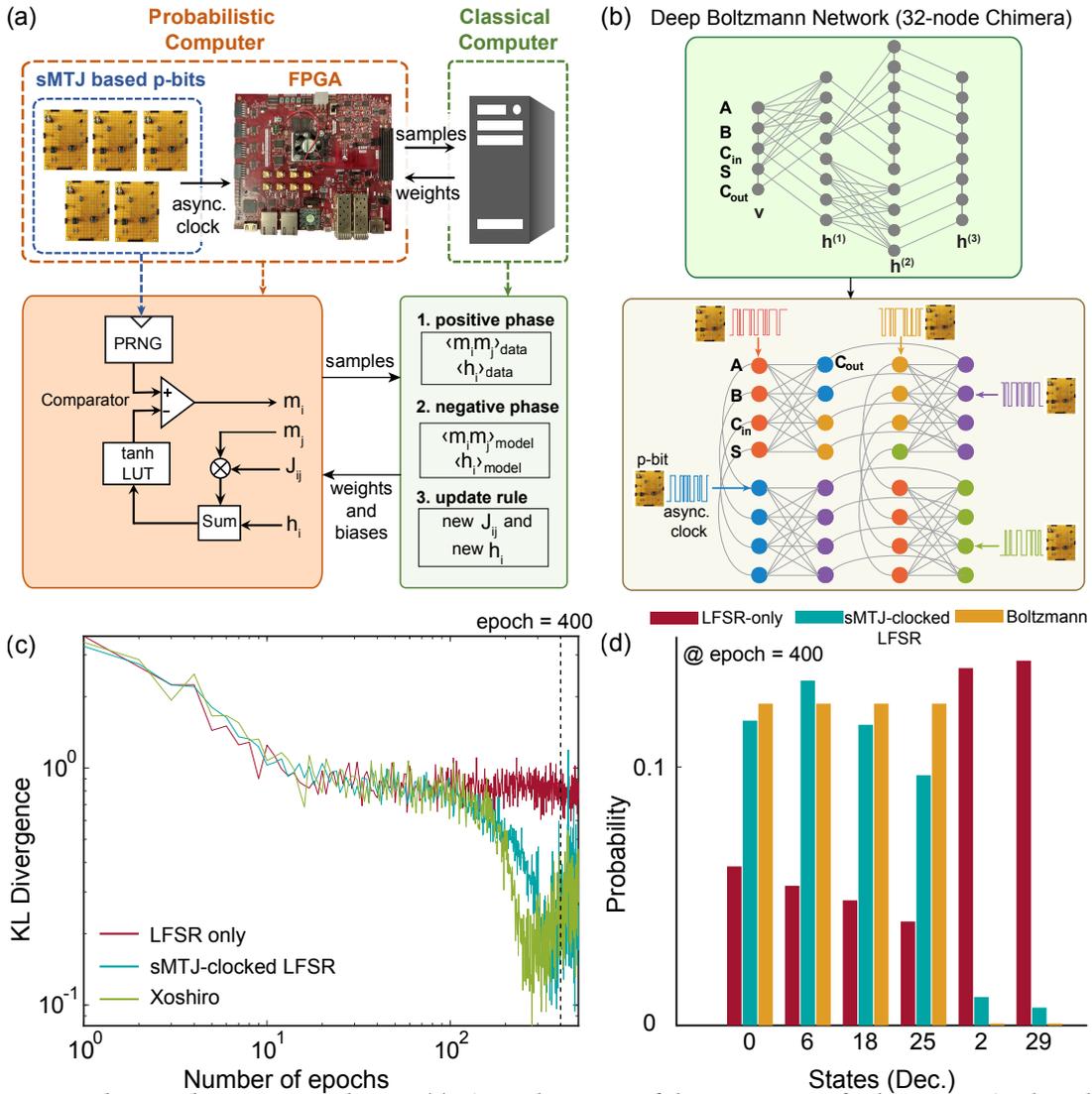


Fig. 3: Learning deep Boltzmann machines. (a) The architecture of the p-computer for learning. The digital p-bits in FPGA are fed by sMTJ-based p-bits output similar to probabilistic inference. The weights J_{ij} and biases h_i are updated in the CPU for a specified number of epochs. (b) (Top) The 32-node Chimera graph is used as a deep BM. (Bottom) An asynchronous clocking scheme is shown with node coloring. (c) KL divergence as a function of the number of epochs for LFSR (red line), LFSR clocked by sMTJ-based p-bit (blue line), and Xoshiro (green line) (d) The distribution of full adder with learned weights and biases at epoch = 400 where the number of sweeps per epoch = 400 for LFSR-only and the number of sweeps per epoch = 16000 for sMTJ-clocked LFSR. The Boltzmann distribution was obtained with $\beta = 3$. The red, blue, and yellow bars show LFSR and LFSR clocked by sMTJ-based p-bit, and Boltzmann, respectively. The histogram shows 4 correct (0, 6, 18, 25) and 2 incorrect (2, 29) states, out of the 32 possible states. sMTJ-based p-bit closely approximates the ideal Boltzmann distribution whereas the LFSR underestimates correct states and completely fails with states 2 and 29 (see Supplementary Section VII for full histograms for all PRNGs including Xoshiro).

FIG. 4a shows the transistor count for p-bits using 32-bit PRNGs. Three pieces make up a digital p-bit: PRNG, LUT (for the activation function) and a digital comparator (typically small). To understand how each piece contributes to the transistor count, we separate the PRNG from the LUT contributions in FIG. 4a.

First, we reproduce earlier results reported in Ref. [28], corresponding to the benchmarking of the design reported in [15] and find that a 32-bit LFSR requires 1122 transistors which is very close to the custom-designed 32-bit LFSR with

1194 transistors in Ref. [28]. However, we find that the addition of a LUT, ignored in [28], adds significantly more transistors. Even though the inputs to the p-bit are 10-bits (s[6][3]), the saturating behavior of the tanh activation allows reductions in LUT size. In our design, the LUT stores 2^8 words of 32-bit length that are compared to the 32-bit PRNG. Under this precision, the LUT increases the transistor count to 5150, and more would be needed for finer representations. Note that the compact sMTJ-based p-bit design proposed in [15] uses 3 transistors plus an sMTJ which we estimate as

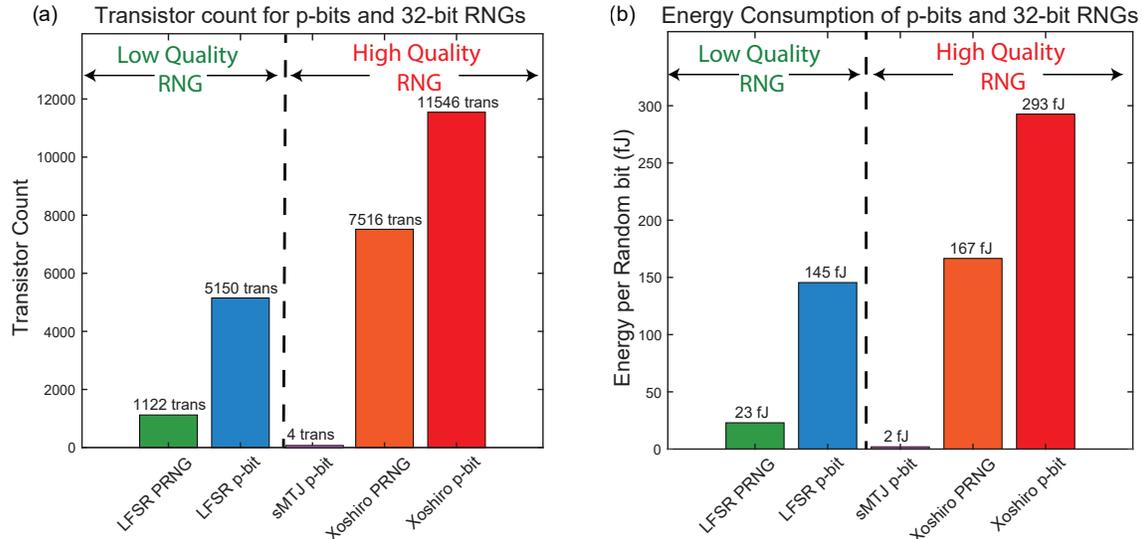


Fig. 4: Transistor counts and energy consumption for p-bit and RNG implementations The digital p-bits and PRNGs are synthesized by the ASAP7 PDK and simulated in HSPICE in transistor-level simulations. The PRNGs are 32-bits long and LUTs store 2^8 words that are 32-bits long to be compared with 32-bit RNGs. The sMTJ-based p-bit result is repeated from [28]. To activate the LUT, a periodic input signal with low inputs to the p-bit has been used. See the text and Supplementary Section XV for details on the energy calculation.

having an area of 4 transistors, following Ref. [28]. In this case, there is no explicit need for a LUT or a PRNG.

Additionally, the results presented in FIG. 2 and 3 indicate that to match the performance of the sMTJ-based p-bits, more sophisticated PRNGs like Xoshiro must be used. In this case, merely the PRNG cost of a 32-bit Xoshiro is 7516 transistors. The LUT costs are the same as LFSR-based p-bits which is about ≈ 4029 transistors.

Collectively, these results indicate that to truly replicate the performance of an sMTJ-based p-bit, the actual transistor cost of a digital design is about 11,000 transistors which is an order of magnitude worse than the conservative estimation performed in Ref. [28].

In FIG. 4b we show the energy costs of these differences. We focus on the energy required to produce one random bit. Once again, our synthesis flow, followed by ASAP7 based HSPICE simulations reproduces the results presented in Ref. [28]. We estimate a 23 fJ energy per random bit from the LFSR based PRNG where this number was reported to be 20 fJ in Ref. [28].

Similar to the transistor count analysis, we consider the effect of the LUT on the energy consumption which was absent in [28]. We first observe that if the LUT is not active, i.e., if the input I_i to the p-bit is not changing, the LUT does not change the energy per random bit very much. In a real p-circuit computation, however, I_i would be continuously changing activating the LUT repeatedly. To simulate these working conditions, we create a variable I_i pulse that wanders around the stochastic window of the p-bit by changing the least significant bits of the input (see Supplementary Section XV). We choose a 1 GHz frequency for this pulse mimicking an sMTJ with a lifetime of 1 ns. We observe that in this case the total energy to create a

random bit on average increases by a factor of $6\times$ for the LFSR, reaching 145 fJ per bit.

For the practically more relevant Xoshiro, the average consumption per random bit reaches around 293 fJ. Once again, we conclude that the 20 fJ per random bit, reported in Ref. [28] underestimates the costs of RNG generation by about an order of magnitude when the RNG quality and other peripheries such as LUTs are carefully taken into account. In this paper, we do not reproduce the energy estimation of the sMTJ-based p-bit but report the estimate in Ref. [28] which assumes an sMTJ-based p-bit with \approx nanosecond fluctuations.

Our benchmarking results highlight the true expense of high-quality, digital p-bits in silicon implementations. Given that functionally interesting and sophisticated p-circuits require above 10000 to 50000 p-bits [5], using a 32-bit Xoshiro-based p-bit in a digital design would consume up to *0.1 to 0.5 Billion transistors*, just for the p-bits. In addition, the limitation of not being able to parallelize or fit more random numbers in hardware would limit the throughput [61] and the probabilistic flips per second, a key metric measuring the effective sampling speed of a probabilistic computer (see for example, [62–64]). As discussed in detail in Supplementary Section XVI, near-term projections with $N = 10^4$ p-bits using sMTJs with in-plane magnetic anisotropy (IMA) ($\tau \approx 1$ ns [19]) can reach $\approx 10^4$ flips/ns in sampling throughput. These results clearly indicate that a digital solution beyond 10000 to 50000 p-bits, as required by large-scale optimization, probabilistic machine learning, and optimization tasks, will remain prohibitive. To solve these traditionally expensive but practically useful problems, the heterogeneous integration of sMTJs holds great promise both in terms of scalability and energy-efficiency.

DISCUSSIONS

This work demonstrates the first hardware demonstration of a heterogeneous computer combining versatile FPGAs with stochastic MTJs for probabilistic inference and deep Boltzmann learning. We introduce a new variation tolerant p-bit circuit that is used to create an asynchronous clock domain, driving digital p-bits in the FPGA. In the process, the CMOS + sMTJ computer shows how commonly used and inexpensive PRNGs can be augmented by magnetic nanodevices to perform as well as high-quality PRNGs (without the resource overhead), both in probabilistic inference and learning experiments. Our CMOS + sMTJ computer also shows the first demonstration of training a deep Boltzmann network in a 32-node Chimera topology, leveraging the asynchronous dynamics of sMTJs. Careful comparisons with existing digital circuits show the true potential of integrated sMTJs which can be scaled up to million p-bit densities far beyond the capabilities of present day CMOS technology (see Supplementary Section XVI for detailed benchmarking and a p-computing roadmap).

METHODS

sMTJ fabrication and circuit parameters

We employ a conventional fixed and free layer sMTJ, both having perpendicular magnetic anisotropy. The reference layer thickness is 1 nm (CoFeB) while the free layer is 1.8 nm (CoFeB), deliberately made thicker to reduce its energy barrier [28, 35]. The stack structure of the sMTJs we use is, starting from the substrate side, Ta(5)/Pt(5)/[Co(0.4)/Pt(0.4)]₆/Co(0.4)/Ru(0.4)/[Co(0.4)/Pt(0.4)]₂/Co(0.4)/Ta(0.2)/CoFeB(1)/MgO(1.1)/CoFeB(1.8)/Ta(5)/Ru(5), where the numbers are in nanometers (FIG. 1a). Films are deposited at room temperature by dc/rf magnetron sputtering on a thermally oxidized Si substrate. The devices are fabricated into a circular shape with a 40-80 nm diameter using electron beam lithography and Ar ion milling and annealed at 300 °C for 1 hour by applying a 0.4 T magnetic field in the perpendicular direction. The average tunnel magnetoresistance ratio (TMR) and resistance area product (RA) are 65% and 4.7 Ωμm², respectively. The discrete sMTJs used in this work are first cut out from the wafer and the electrode pads of the sMTJs are bonded with wires to IC sockets. The following parameters are measured by sweeping DC current to the sMTJ and measuring the voltage. The resistance of the P state R_P is 4.4-5.7 kΩ, the resistance of the AP state R_{AP} is 5.9-7.4 kΩ, and the current at which P/AP fluctuations are 50% is defined as $I_{50/50}$, in between 14-20 μA. At the output of the new p-bit design, we use an extra branch with a bipolar junction transistor (BJT) that acts as a buffer to the peripheral module (PMOD) pins of the Kintex UltraScale KU040 FPGA board. Given the electrostatic sensitivity of the sMTJs, this branch also protects the circuit from any transients that might originate from the FPGA.

Digital synthesis flow

HDL codes are converted to gate-level models using the Synopsys Design Compiler. Conversion from these models to Spice netlists is done using Calibre Verilog-to-LVS. Netlist post-processing is done by a custom Mathematica script to make it HSPICE compatible. Details of the synthesis flow (shown in FIG. 4), followed by HSPICE simulation results for functional verification and power analysis are provided in Supplementary Sections XIII, XIV and XV.

DATA AVAILABILITY

All processed data generated in this study are provided in the main text and Supplementary Information. The data that support the plots within this paper and other findings of this study are available from the corresponding author upon request.

CODE AVAILABILITY

The computer code used in this study is available from the corresponding author upon request.

References

- [1] Thomas N Theis and H-S Philip Wong. The end of moore's law: A new beginning for information technology. *Computing in Science & Engineering*, 19(2):41–50, 2017.
- [2] Jack Dongarra and Francis Sullivan. Guest editors introduction to the top 10 algorithms. *Computing in Science & Engineering*, 2(01):22–23, 2000.
- [3] K. Y. Camsari et al. Stochastic p-bits for invertible logic. *Physical Review X*, 7(3):031014, 2017.
- [4] Brian Sutton et al. Autonomous probabilistic coprocessing with petaflips per second. *IEEE Access*, 8:157238–157252, 2020.
- [5] Navid Anjum Aadit, Andrea Grimaldi, Mario Carpentieri, Luke Theogarajan, John M Martinis, Giovanni Finocchio, and Kerem Y Camsari. Massively parallel probabilistic computing with sparse ising machines. *Nature Electronics*, 5(7):460–468, 2022.
- [6] Jan Kaiser and Supriyo Datta. Probabilistic computing with p-bits. *Applied Physics Letters*, 119(15):150503, 2021.
- [7] Kerem Y Camsari, Brian M Sutton, and Supriyo Datta. P-bits for probabilistic spin logic. *Applied Physics Reviews*, 6(1):011305, 2019.
- [8] Shuvro Chowdhury, Andrea Grimaldi, Navid Anjum Aadit, Shaila Niazi, Masoud Mohseni, Shun Kanai, Hideo Ohno, Shunsuke Fukami, Luke Theogarajan, Giovanni Finocchio, et al. A full-stack view of probabilistic computing with p-bits: devices, architectures and algorithms. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, 9(1):1–11, 2023.
- [9] Kyung Seok Woo, Jaehyun Kim, Janguk Han, Woohyun Kim, Yoon Ho Jang, and Cheol Seong Hwang. Probabilistic computing using cu0. 1te0. 9/hfo2/pt diffusive memristors. *Nature Communications*, 13(1):5762, 2022.
- [10] Yixuan Liu, Qiao Hu, Qiqiao Wu, Xuanzhi Liu, Yulin Zhao, Donglin Zhang, Zhongze Han, Jinhui Cheng, Qingting Ding, Yongkang Han, et al. Probabilistic circuit implementation based on p-bits using the intrinsic random property of rram and p-bit multiplexing strategy. *Micromachines*, 13(6):924, 2022.

- [11] Tae Joon Park, Kemal Selcuk, Hai-Tian Zhang, Sukriti Manna, Rohit Batra, Qi Wang, Haoming Yu, Navid Anjum Aadit, Subramanian KRS Sankaranarayanan, Hua Zhou, et al. Efficient probabilistic computing with stochastic perovskite nickelates. *Nano Letters*, 22(21):8654–8661, 2022.
- [12] Sheng Luo, Yihan He, Baofang Cai, Xiao Gong, and Gengchiao Liang. Probabilistic-bits based on ferroelectric field-effect transistors for probabilistic computing. *IEEE Electron Device Letters*, 44(8):1356–1359, 2023.
- [13] William Whitehead, Zachary Nelson, Kerem Y Camsari, and Luke Theogarajan. Cmos-compatible ising and potts annealing using single photon avalanche diodes. *Nature Electronics*, 6(12):1009–1019, 2023.
- [14] Charles Roques-Carmes, Yannick Salamin, Jamison Sloan, Seou Choi, Gustavo Velez, Ethan Koskas, Nicholas Rivera, Steven E Kooi, John D Joannopoulos, and Marin Soljacic. Biasing the quantum vacuum to control macroscopic probability distributions. *Science*, 381(6654):205–209, 2023.
- [15] Kerem Yunus Camsari, Sayeef Salahuddin, and Supriyo Datta. Implementing p-bits with embedded mtj. *IEEE Electron Device Letters*, 38(12):1767–1770, 2017.
- [16] Suresh Cheemalavagu, Pinar Korkmaz, Krishna V Palem, Bilge ES Akgul, and Lakshmi N Chakrapani. A probabilistic cmos switch and its realization by exploiting noise. In *IFIP International Conference on VLSI*, pages 535–541, 2005.
- [17] Akio Fukushima, Takayuki Seki, Kay Yakushiji, Hitoshi Kubota, Hiroshi Imamura, Shinji Yuasa, and Koji Ando. Spin dice: A scalable truly random number generator based on spintronics. *Applied Physics Express*, 7(8):083001, 2014.
- [18] Laura Rehm, Corrado Carlo Maria Capriata, Misra Shashank, J Darby Smith, Mustafa Pinarbasi, B Gunnar Malm, and Andrew D Kent. Stochastic magnetic actuated random transducer devices based on perpendicular magnetic tunnel junctions. *Phys. Rev. Appl.*, 19:024035, 2023.
- [19] Christopher Safranski, Jan Kaiser, Philip Trouilloud, Pouya Hashemi, Guohan Hu, and Jonathan Z Sun. Demonstration of nanosecond operation in stochastic magnetic tunnel junctions. *Nano Letters*, 21(5):2040–2045, 2021.
- [20] Keisuke Hayakawa, Shun Kanai, Takuya Funatsu, Junta Igarashi, Butsurin Jinnai, WA Borders, H Ohno, and S Fukami. Nanosecond random telegraph noise in in-plane magnetic tunnel junctions. *Physical Review Letters*, 126(11):117202, 2021.
- [21] Leo Schnitzspan, Mathias Kläui, and Gerhard Jakob. Nanosecond true-random-number generation with superparamagnetic tunnel junctions: Identification of joule heating and spin-transfer-torque effects. *Phys. Rev. Appl.*, 20:024002, Aug 2023.
- [22] Jan Kaiser, Avinash Rustagi, Kerem Y Camsari, Jonathan Z Sun, Supriyo Datta, and Pramey Upadhyaya. Subnanosecond fluctuations in low-barrier nanomagnets. *Physical Review Applied*, 12(5):054056, 2019.
- [23] 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*, 10:1–5, 2019.
- [24] Shun Kanai, Keisuke Hayakawa, Hideo Ohno, and Shunsuke Fukami. Theory of relaxation time of stochastic nanomagnets. *Physical Review B*, 103(9):094423, 2021.
- [25] Takuya Funatsu, Shun Kanai, Jun'ichi Ieda, Shunsuke Fukami, and Hideo Ohno. Local bifurcation with spin-transfer torque in superparamagnetic tunnel junctions. *Nature communications*, 13(1):4079, 2022.
- [26] Kerem Y Camsari, Mustafa Mert Torunbalci, William A Borders, Hideo Ohno, and Shunsuke Fukami. Double-free-layer magnetic tunnel junctions for probabilistic bits. *Phys. Rev. Appl.*, 15:044049, 2021.
- [27] Keito Kobayashi, Keisuke Hayakawa, Junta Igarashi, William A Borders, Shun Kanai, Hideo Ohno, and Shunsuke Fukami. External-field-robust stochastic magnetic tunnel junctions using a free layer with synthetic antiferromagnetic coupling. *Physical Review Applied*, 18(5):054085, 2022.
- [28] William A Borders et al. Integer factorization using stochastic magnetic tunnel junctions. *Nature*, 573:390–393, 2019.
- [29] Jan Kaiser, William A Borders, Kerem Y Camsari, Shunsuke Fukami, Hideo Ohno, and Supriyo Datta. Hardware-aware in situ learning based on stochastic magnetic tunnel junctions. *Physical Review Applied*, 17(1):014016, 2022.
- [30] Jia Si, Shuhan Yang, Yunuo Cen, Jiaer Chen, Zhaoyang Yao, Dong-Jun Kim, Kaiping Cai, Jerald Yoo, Xuanyao Fong, and Hyunsoo Yang. Energy-efficient superparamagnetic ising machine and its application to traveling salesman problems. *arXiv preprint arXiv:2306.11572*, 2023.
- [31] Sidra Gibeault, Temitayo N Adeyeye, Liam A Pocher, Daniel P Lathrop, Matthew W Daniels, Mark D Stiles, Jabez J McClelland, William A Borders, Jason T Ryan, Philippe Talatchian, et al. Programmable electrical coupling between stochastic magnetic tunnel junctions. *arXiv preprint arXiv:2312.13171*, 2023.
- [32] John Daniel, Zheng Sun, Xuejian Zhang, Yuanqiu Tan, Neil Dilley, Zhihong Chen, and Joerg Appenzeller. Experimental demonstration of an integrated on-chip p-bit core utilizing stochastic magnetic tunnel junctions and 2d-mos $\text{-}\{2\}$ fets. *arXiv preprint arXiv:2308.10989*, 2023.
- [33] Naeimeh Mohseni, Peter L McMahon, and Tim Byrnes. Ising machines as hardware solvers of combinatorial optimization problems. *Nature Reviews Physics*, 4(6):363–379, 2022.
- [34] Damir Vodenicarevic, Nicolas Locatelli, Alice Mizrahi, Joseph S Friedman, Adrien F Vincent, Miguel Romera, Akio Fukushima, Kay Yakushiji, Hitoshi Kubota, Shinji Yuasa, et al. Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing. *Physical Review Applied*, 8(5):054045, 2017.
- [35] Bradley Parks, Mukund Bapna, Julianne Igbokwe, Hamid Almasi, Weigang Wang, and Sara A Majetich. Superparamagnetic perpendicular magnetic tunnel junctions for true random number generators. *AIP Advances*, 8(5):055903, 2018.
- [36] Vaibhav Ostwal and Joerg Appenzeller. Spin-orbit torque-controlled magnetic tunnel junction with low thermal stability for tunable random number generation. *IEEE Magnetics Letters*, 10(4503305):1–5, 2019.
- [37] Yang Lv, Brandon R Zink, and Jian-Ping Wang. Bipolar random spike and bipolar random number generation by two magnetic tunnel junctions. *IEEE Transactions on Electron Devices*, 69(3):1582–1587, 2022.
- [38] Zhenxiao Fu, Yi Tang, Xi Zhao, Kai Lu, Yemin Dong, Amit Shukla, Zhifeng Zhu, and Yumeng Yang. An overview of spintronic true random number generator. *Frontiers in Physics*, 9:638207, 2021.
- [39] S Smithson et al. Efficient cmos invertible logic using stochastic computing. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 66(6):2263–2274, 2019.
- [40] William T Coffey and Yuri P Kalmykov. Thermal fluctuations of magnetic nanoparticles: Fifty years after brown. *Journal of Applied Physics*, 112(12):121301, 2012.
- [41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

- [42] Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50:95–125, 2003.
- [43] Philip M Long and Rocco A Servedio. Restricted boltzmann machines are hard to approximately evaluate or simulate. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, page 703–710. Omnipress, 2010.
- [44] Werner Krauth. *Statistical mechanics: algorithms and computations*, volume 13. OUP Oxford, 2006.
- [45] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50:5–43, 2003.
- [46] Christof Paar and Jan Pelzl. *Understanding cryptography: a textbook for students and practitioners*. Springer Science & Business Media, 2009.
- [47] David Blackman and Sebastiano Vigna. Scrambled linear pseudorandom number generators. *ACM Trans. Math. Softw.*, 47(4), 2021.
- [48] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.
- [49] Andrew Rukhin, Juan Soto, James Nechvatal, Miles Smid, Elaine Barker, Stefan Leigh, Mark Levenson, Mark Vangel, David Banks, Alan Heckert, James Dray, and San Vo. A statistical test suite for random and pseudorandom number generators for cryptographic applications. Technical report, 2010.
- [50] William H Press, William T Vetterling, Saul A Teukolsky, and Brian P Flannery. *Numerical recipes*. Citeseer, 1988.
- [51] Donald Knuth. *The art of computer programming, 2 (Seminumerical Algorithms)*. Addison-Wesley, 1981.
- [52] Janusz Rajski and Jerzy Tyszer. Design of phase shifters for bist applications. In *Proceedings. 16th IEEE VLSI Test Symposium (Cat. No. 98TB100231)*, pages 218–224. IEEE, 1998.
- [53] Giorgio Parisi and Federico Rapuano. Effects of the random number generator on computer simulations. *Physics Letters B*, 157(4):301–302, 1985.
- [54] Thomas Filk, Mihail Marcu, and Klaus Fredenhagen. Long range correlations in random number generators and their influence on monte carlo simulations. *Physics Letters B*, 165(1):125–130, 1985.
- [55] I. Vattulainen, T. Ala-Nissila, and K. Kankaala. Physical tests for random numbers in simulations. *Phys. Rev. Lett.*, 73:2513–2516, 1994.
- [56] Kelly Boothby, Paul Bunyk, Jack Raymond, and Aidan Roy. Next-generation topology of d-wave quantum processors. *arXiv preprint arXiv:2003.00133*, 2020.
- [57] Lennart Dabelow and Masahito Ueda. Three learning stages and accuracy–efficiency tradeoff of restricted boltzmann machines. *Nature communications*, 13(1):5474, 2022.
- [58] Shaila Niazi, Navid Anjum Aadit, Masoud Mohseni, Shuvro Chowdhury, Yao Qin, and Kerem Y Camsari. Training deep boltzmann networks with sparse ising machines. *arXiv preprint arXiv:2303.10728*, 2023.
- [59] Jialiang Yin, Yu Liu, Bolin Zhang, Ao Du, Tianqi Gao, Xiangyue Ma, Yi Dong, Yue Bai, Shiyang Lu, Yudong Zhuo, et al. Scalable ising computer based on ultra-fast field-free spin orbit torque stochastic device with extreme 1-bit quantization. In *2022 International Electron Devices Meeting (IEDM)*, pages 36–1. IEEE, 2022.
- [60] Lawrence T. Clark, Vinay Vashishtha, Lucian Shifren, Aditya Gujja, Saurabh Sinha, Brian Cline, Chandarasekaran Ramamurthy, and Greg Yeric. Asap7: A 7-nm finfet predictive process design kit. *Microelectronics Journal*, 53:105–115, 2016.
- [61] Shashank Misra, Leslie C. Bland, Suma G. Cardwell, Jean Anne C. Incorvia, Conrad D. James, Andrew D. Kent, Catherine D. Schuman, J. Darby Smith, and James B. Aimone. Probabilistic neural computing with stochastic devices. *Advanced Materials*, page 2204569, 2022.
- [62] Tobias Preis, Peter Virnau, Wolfgang Paul, and Johannes J Schneider. Gpu accelerated monte carlo simulation of the 2d and 3d ising model. *Journal of Computational Physics*, 228(12):4468–4477, 2009.
- [63] Kun Yang, Yi-Fan Chen, Georgios Roumpos, Chris Colby, and John Anderson. High performance monte carlo simulation of ising model on tpu clusters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2019.
- [64] Joshua Romero et al. High performance implementations of the 2d ising model on gpus. *Computer Physics Communications*, 256:107473, 2020.
- [65] Rafatul Faria, Jan Kaiser, Kerem Y Camsari, and Supriyo Datta. Hardware design for autonomous bayesian networks. *Frontiers in Computational Neuroscience*, 15:584797, 2021.
- [66] Emile Aarts and Jan Korst. *Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing*. John Wiley & Sons, Inc., 1989.
- [67] Ahmed Zeeshan Pervaiz, Supriyo Datta, and Kerem Y Camsari. Probabilistic computing with binary stochastic neurons. In *2019 IEEE BiCMOS and Compound semiconductor Integrated Circuits and Technology Symposium (BCICTS)*, pages 1–6. IEEE, 2019.
- [68] Orchi Hassan, Supriyo Datta, and Kerem Y. Camsari. Quantitative evaluation of hardware binary stochastic neurons. *Phys. Rev. Applied*, 15:064046, Jun 2021.
- [69] airhdl.com. airhdl VHDL/SystemVerilog Register Generator. <https://airhdl.com>.
- [70] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.
- [71] Asja Fischer and Christian Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47:25–39, 2014.
- [72] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [73] Nike Dattani, Szilard Szalay, and Nick Chancellor. Pegasus: The second connectivity graph for large-scale quantum annealing hardware. *arXiv preprint arXiv:1901.07636*, 2019.
- [74] Nihal Sanjay Singh, Shaila Niazi, Shuvro Chowdhury, Kemal Selcuk, Haruna Kaneko, Keito Kobayashi, Shun Kanai, Hideo Ohno, Shunsuke Fukami, and Kerem Yunus Camsari. Hardware demonstration of feedforward stochastic neural networks with fast mtj-based p-bits. In *2023 International Electron Devices Meeting (IEDM) Proceedings*. IEEE, 2023.
- [75] Benjamin Block, Peter Virnau, and Tobias Preis. Multi-gpu accelerated multi-spin monte carlo simulations of the 2d ising model. *Computer Physics Communications*, 181(9):1549–1556, 2010.
- [76] 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(10):2467–2478, 2014.
- [77] Andrea Grimaldi, Kemal Selcuk, Navid Anjum Aadit, Keito Kobayashi, Qixuan Cao, Shuvro Chowdhury, Giovanni Finocchio, Shun Kanai, Hideo Ohno, Shunsuke Fukami, and Kerem Y. Camsari. Experimental evaluation of simulated

quantum annealing with mtj-augmented p-bits. In *2022 International Electron Devices Meeting (IEDM)*, pages 22.4.1–22.4.4, 2022.

- [78] K Lee, JH Bak, YJ Kim, CK Kim, A Antonyan, DH Chang, SH Hwang, GW Lee, NY Ji, WJ Kim, et al. 1gbit high density embedded stt-mram in 28nm fdsoi technology. In *2019 IEEE International Electron Devices Meeting (IEDM)*, pages 2–2. IEEE, 2019.

ACKNOWLEDGEMENTS

We are grateful to Subhasish Mitra and Carlo Gilardi for discussions regarding Linear Feedback Shift Registers and high-level synthesis. We gratefully acknowledge Kevin Cao and Mishel Jyothis Paul for their help with the configuration of ASAP7 PDK. We are grateful to Shuvro Chowdhury for his comments on an earlier version of this manuscript. The U.S. National Science Foundation (NSF) grant CCF 2106260, the Office of Naval Research Young Investigator Program (YIP) grant, SAMSUNG Global Research Outreach (GRO) grant, and an NSF CAREER grant are acknowledged by N.S.S., Q.C, K.S., T.H., S.N., N.A.A., and K.Y.C for supporting this research. Murata Science Foundation and Marubun Research Promotion Foundation are acknowledged by K.K. JST-CREST Grant No. JPMJCR19K3, JST-AdCORP Grant No. JPMJKB2305, and MEXT X-NICS Grant No. JPJ011438 are acknowledged by S.F. JST-PRESTO Grant No. JPMJPR21B2 is acknowledged by S.K.

AUTHOR CONTRIBUTIONS

KYC and SF conceived and supervised the study. NSS developed the ASAP7 synthesis flow, ran SPICE simulations, and performed circuit-level experiments with sMTJs along with KK, QC and KS. KK, SK, SF and HO fabricated sMTJs. KK, QC, and SK ran the device-level sMTJ experiments. NAA, SN, TH have implemented the FPGA design for the learning and inference experiments. All authors have discussed the results and participated in writing and improving the manuscript.

COMPETING INTERESTS

The Authors declare no Competing Financial or Non-Financial Interests

SUPPLEMENTARY INFORMATION

I. BASIC PRINCIPLES OF PROBABILISTIC COMPUTING WITH HARDWARE P-BITS

Probabilistic algorithms (e.g., sampling, inference, optimization) are performed with a network of p-bits interacting with each other [7]. The basic equation for the p-bit is given by:

$$m_i(t + \tau_N) = \text{sgn}\{\text{rand}(-1, 1) + \tanh[\beta I_i(t)]\}, \quad (1)$$

where $\text{rand}(-1, 1)$ represents a random number drawn from the uniform distribution in $[-1, 1]$, β is the inverse algorithmic temperature, and I_i is the local field of p-bit “ i ” received from its neighbors. τ_N in this equation is defined as the neuron evaluation time [65]. For the typical choice of 2-local (Ising-like) energy functions, I_i is given by:

$$I_i(t + \tau_S) = \sum J_{ij} m_j(t) + h_i, \quad (2)$$

where J_{ij} are the weights and h_i is the bias term for each individual p-bit. τ_S represents the synapse evaluation time. If the network of p-bits is symmetric ($J_{ij} = J_{ji}$), it is possible to define the following energy function:

$$E = - \left(\sum J_{ij} m_i m_j + \sum h_i m_i \right). \quad (3)$$

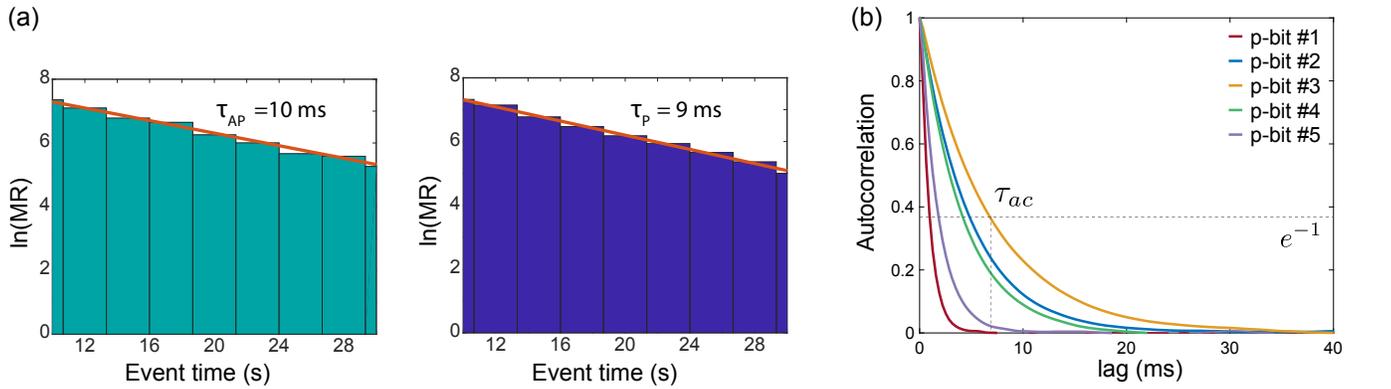
In such a network of N p-bits, there are 2^N states, $S = \{1, 2, \dots, 2^N\}$, and each state $j \in S$ is visited according to the Boltzmann-Gibbs distribution:

$$p_j = \frac{1}{Z} \exp(-\beta E_j). \quad (4)$$

The coupled evolution of Supplementary Eq. (1) and Supplementary Eq. (2) represents a dynamical system and a wide variety of problems can be mapped onto this system, including combinatorial optimization, machine learning and quantum simulation problems, all phrased in terms of powerful physics-inspired Monte Carlo algorithms [6, 8]. In this paper, we focus on the two settings of probabilistic inference and Boltzmann learning. In either case, we will be interested in a fixed β value (typically 1) and sample from the corresponding Boltzmann-Gibbs distribution of the network. We will show how the asynchronous and truly random bits generated by the sMTJs represent a low-level realization of physics-inspired probabilistic computation.

II. CHARACTERIZING SMTJS THROUGH P-BITS

We first characterize the statistics of the sMTJs (in Supplementary Figure 1) by making measurements on the p-bit circuit described later in Supplementary Figure 2b. The fluctuations of the p-bit circuit are controlled entirely by the sMTJs. We observe the outputs of 5 sMTJ-based p-bits and characterize their rate of fluctuations from event times and autocorrelations.



Supplementary Figure 1. sMTJ-based p-bit characterization. (a) Histogram of $\ln(\text{MR})$ (number of magnetic relaxation (MR) events) as a function of the event times t_{event} for p-bit #2, following the exponential distribution $1/\tau_{\text{P,AP}} \exp(-t_{\text{event}}/\tau_{\text{P,AP}})$, expected from a Poisson process (red lines). (b) Autocorrelation of sMTJ-based p-bits #1-5. We define τ_{ac} as the time at which the normalized autocorrelation decays to $1/e$.

In the main paper FIG. 1g shows the fluctuations of p-bits #1-5 at $I_{50/50}$, which is the current at which the sMTJs show 50/50 fluctuations for the high- and low-resistance states. The rate of fluctuations provides an estimate of the neuron evaluation time τ_N of Supplementary Eq. (1) after which the p-bit produces a new and independent random bit.

p-bit	Mean MR time (τ)	Autocorr. time (τ_{ac})	TMR	$I_{50/50}$
1	2.4 ms	0.97 ms	64%	16 μA
2	9.6 ms	4.8 ms	68%	19 μA
3	14.4 ms	6.8 ms	65%	20 μA
4	8.7 ms	4.2 ms	59%	17 μA
5	4.2 ms	1.8 ms	65%	14 μA

Supplementary Table 1. Table of results for all 5 p-bits reporting mean relaxation time $\tau = \sqrt{\tau_P \tau_{AP}}$, τ_{ac} , TMR and $I_{50/50}$ of sMTJs. $I_{50/50}$ is defined as the absolute value of the current at which the perpendicular sMTJs show 50/50 fluctuations by canceling the uncompensated dipolar field resulting from the fixed layer.

Even though the sMTJs were fabricated under the same conditions, slight differences in their volumes and shapes critically affect their relaxation times. The relaxation times $\tau_{P,AP}$ obeys a Néel-Arrhenius law [40], described by $\tau = \tau_0 \exp(\Delta/k_B T)$, where τ_0 is the attempt time and Δ is the energy barrier of the nanomagnet. In this paper, our magnets are not truly zero-barrier and their fluctuation rates exponentially depend on the energy barrier, Δ , however, detailed theoretical analysis shows that this dependence is much less pronounced in low barrier nanomagnets [22, 23]. Moreover, as we experimentally show in Sections 3-4, p-bits in symmetric networks are agnostic to update orders [3, 65, 66], therefore variations in these time scales should not play a prohibitive role in scaled implementations of p-computers (see Supplementary Section VIII for another experiment showing this update order invariance).

In Supplementary Figure 1b, we calculate the normalized autocorrelation of the p-bits using a 300 s time window with a sampling rate of 3.16 kHz, collecting $N \approx 949000$ samples. The discrete autocorrelation function is defined as:

$$C[m] = \frac{\sum_{n=0}^{N-1} V[n]V[n+m]}{\sum_{n=0}^{N-1} V[n]^2}, \quad (5)$$

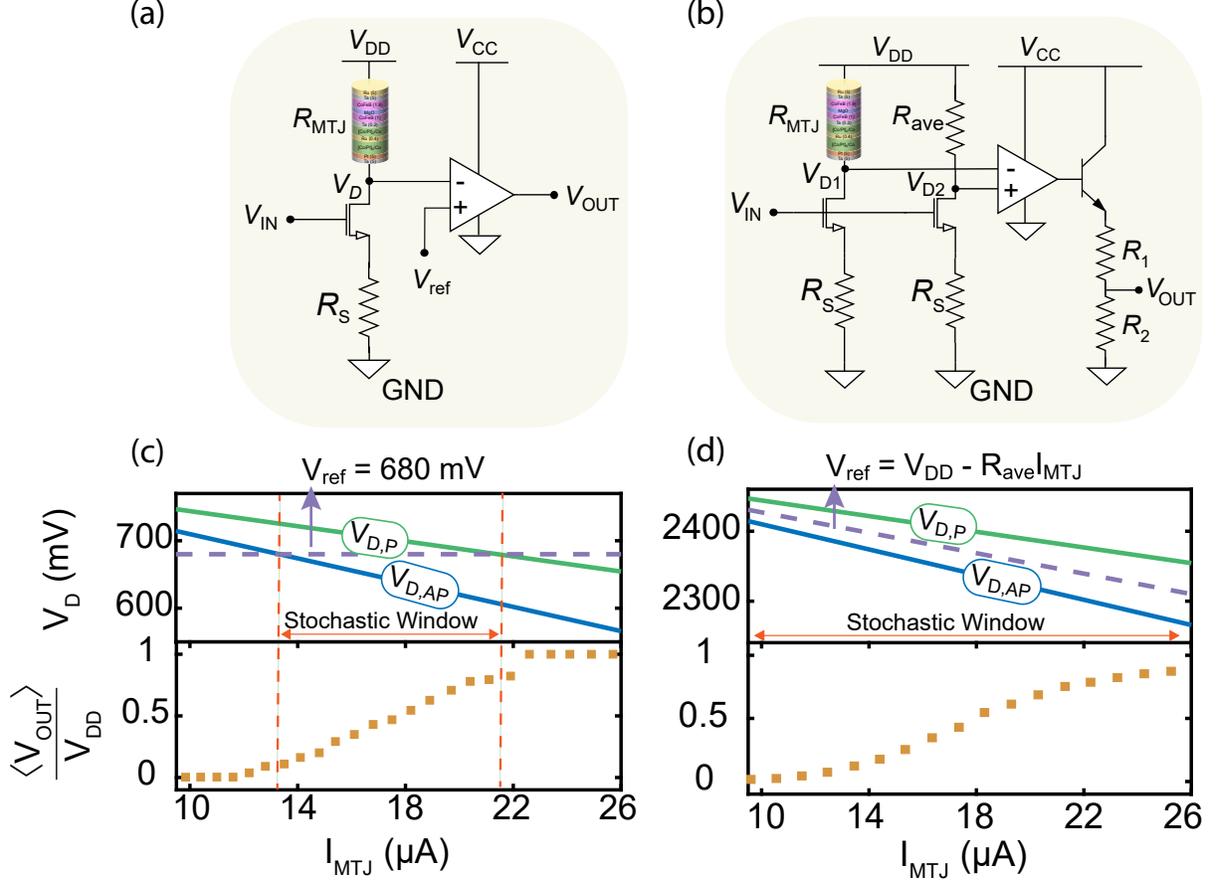
where $V[n]$ is the discrete sampled voltage read from the oscilloscope, and m represents the discretized lag time. Since autocorrelations typically show exponential decay [22], we extract an autocorrelation time τ_{ac} , by measuring and fitting the autocorrelation to a continuous function $C(t_{lag}) = \exp(-t_{lag}/\tau_{ac})$. As an additional measure to τ_{ac} , we also determine the frequency of magnetic relaxations that are characterized by an event time. The event time, t_{event} , is defined as the time between one event to the next. Measuring the distribution of event times from parallel and antiparallel configurations, we observe that the p-bit outputs are distributed according to the exponential distribution, indicating a Poisson process [25]. Overall, p-bits show a good degree of variation in the relaxation and autocorrelation times as well as in their TMR values and their 50/50 currents. These values are summarized in Supplementary Table 1.

III. A VOLTAGE-COMPARATOR BASED P-BIT

In this section, we describe a new p-bit circuit we designed in this work, comparing its characteristics to an earlier design implemented in [28]. Supplementary Figure 2a,b show both these designs. The earlier design was implemented in several small-scale experimental demonstrations using perpendicular sMTJs [28, 29, 67]. In the original theoretical proposal [15], however, circular or elliptical in-plane nanomagnets were used. In-plane low barrier magnets are very hard to pin, requiring spin polarized currents of ≈ 100 -500 μA or more for typical parameters [23]. If the magnetization provides continuous randomness providing all resistance values between R_P to R_{AP} , this allows a faithful realization of Supplementary Eq. (1) as carefully discussed in [68]. In such a case, spin-transfer-torque pinning is an unnecessary distraction, other than causing a read disturbance. Indeed, the fastest experimental p-bits are based on in-plane magnetic tunnel junctions [26, 27] and variations of the design proposed in Ref. [15] may still be useful in future implementations.

The experimental demonstrations including the present work have so far been primarily focused on perpendicular sMTJs to keep fluctuation speeds slow for practical reasons. Perpendicular sMTJs are easily pinned with spin currents around ≈ 10 to 20 μA for typical parameters [68]. Unlike in-plane sMTJs, however, perpendicular sMTJs switch telegraphically and they do not provide a uniform resistance between the two extremes. By a fortunate coincidence, the presence of the uncompensated dipolar field from the reference layer and easy pinning of perpendicular sMTJs appear to allow the realization of Supplementary Eq. (1) in hardware [28], since the spin-torque pinning changes the 50/50 fluctuations of the sMTJ [67].

In the design shown in Supplementary Figure 2a, the comparator has a fixed reference voltage. This means that as a function of V_{IN} , the drain voltage swings between two extremes, $V_{D,AP} = V_{DD} - R_{AP}I_{MTJ}$ and $V_{D,P} = V_{DD} - R_P I_{MTJ}$. As shown



Supplementary Figure 2. sMTJ-based p-bit circuits (a) Single branch p-bit circuit diagram based on [28] with the following typical parameters: $V_{DD} = 0.8$ V, $V_{CC} = 2$ V, $R_S = 10$ k Ω , $V_{ref} = 680$ mV, and NMOS (2N7000) [28] (b) Double branch p-bit circuit with a variable V_{ref} and fixed R_S . $V_{DD} = 2.5$ V, $V_{CC} = 5$ V, $R_S = 47$ k Ω , $R_{ave} = 1/2 (R_P + R_{AP})$, $R_1 = 100$ Ω , $R_2 = 220$ Ω , NMOS (ALD1101), and op-amp (ALD1702). (c) Single branch circuit of (a): the drain voltage as a function of current through the sMTJ, where ($V_{D,AP} = V_{DD} - R_{AP} I_{MTJ}$) and ($V_{D,P} = V_{DD} - R_P I_{MTJ}$). Vertical alignments show how the fixed reference voltage limits the stochastic window of the p-bit. (d) Double branch circuit of (b): the drain voltage as a function of the sMTJ current, showing the variable reference voltage to the op-amp.

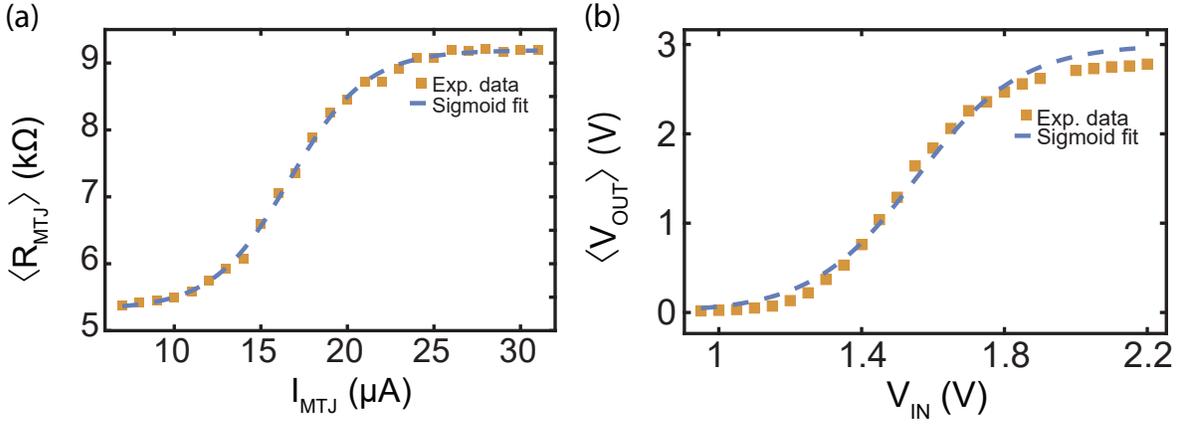
in Supplementary Figure 2c, this limits the stochastic window of the p-bit. For the circuit shown in Supplementary Figure 2b, we have two parallel branches, one with the R_{MTJ} and the other with R_{ave} , defined as $1/2 (R_P + R_{AP})$ of the sMTJ. The output nodes of these branches, taken from the drain of the transistor are compared by an operational amplifier. The key difference as shown in Supplementary Figure 2d is that the voltage reference is variable in the circuit of Supplementary Figure 2b and a larger stochastic window can be obtained. Additionally, the differential nature of the voltage comparison in the new design allows the use of a single source resistance across all sMTJs, unlike the circuit of Supplementary Figure 2a where the source resistance is adjusted individually [28]. Finally, the bipolar junction transistor at the output may not be necessary for integrated implementations, in this work, it simply functions as a buffer and lowers the output voltage to 3V, to safely interface with the FPGA.

IV. EXPERIMENTALLY DESIGNING THE VOLTAGE-COMPARATOR BASED P-BIT

Although the circuit shown in Supplementary Figure 2b is operated at only 50/50 fluctuations to provide asynchronous clocks, this circuit can also be used as a full p-bit whose output probability is tunable by the input voltage V_{IN} .

Each of the sMTJs used in this work has different characteristics, as such R_{ave} and V_{IN} should be customized for the specific sMTJ used in a p-bit circuit. We first characterize the specific sMTJ by experimentally measuring the $\langle R_{MTJ} \rangle - I_{MTJ}$ (Supplementary Figure 3a), the value of R_P , and the value of R_{AP} . The value of the R_{ave} resistor is given by $R_{ave} = 1/2 (R_P + R_{AP})$.

As shown in Supplementary Figure 2b, the NMOS transistor and the R_S in either branch of the circuit form a constant



Supplementary Figure 3. sMTJ-based p-bit circuit characterization (a) $\langle R_{\text{MTJ}} \rangle$, the time average (over a period of 3 minutes) of the sMTJ resistance, as a function of I_{MTJ} , the current flowing through the sMTJ. The yellow squares are experimental data, and the blue dashed line is a fit of the form $\langle R_{\text{MTJ}} \rangle = a + b / \{1 + \exp[-c(I_{\text{MTJ}} - d)]\}$, with $a = 5.32 \text{ k}\Omega$, $b = 3.87 \text{ k}\Omega$, $c = 0.45 (\mu\text{A})^{-1}$, $d = 16.6 \mu\text{A}$. (b) Experimentally measured $\langle V_{\text{OUT}} \rangle$, the time average (over a period of 3 minutes) of the output of the circuit shown in Supplementary Figure 2b, as a function of DC input voltage V_{IN} . The yellow squares are experimental data, and the blue dashed line is a fit of the form $\langle V_{\text{OUT}} \rangle = 1/2 V'_{\text{CC}} [\tanh[\beta(V_{\text{IN}} - V_0)] + 1]$, where $V_0 = 1.55 \text{ V}$, $\beta = 3.43 \text{ V}^{-1}$, $V'_{\text{CC}} = 3 \text{ V}$ is a reduced voltage from $V_{\text{CC}} = 5 \text{ V}$.

current source. The two current sources in the two branches form a current mirror, so they have the same I-V characteristics. Note that the I-V characteristics of the circuit (with ALD1101 NMOS transistors and $R_S = 47 \text{ k}\Omega$) do not depend on the R_P and R_{AP} values of the specific sMTJ used.

The ALD1101 current source can be characterized using the following setup: we replace the sMTJ and the R_{ave} in the left and right branches of the circuit with two $10 \text{ k}\Omega$ resistors, we connect V_{DD} to a 2.5 V power supply, and we leave V_{CC} unconnected. Supplementary Table 2 shows the experimentally measured I-V relationship.

Supplementary Table 2. ALD1101 constant current source characterization: experimentally measured $I_{\text{NMOS}} - V_{\text{IN}}$ relationship. The ALD1101 constant current source consists of the ALD1101 matched pair NMOS (the ALD1101 is a single IC consisting of two NMOS's with their threshold voltages matched to within 10 mV of difference) and the two R_S resistors, as shown in Supplementary Figure 2b.

V_{IN} (mV)	600	700	800	900	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000	2100	2200	2300	2400
I_{NMOS} (μA)	0.974	2.24	3.83	5.55	7.37	9.21	11.1	13.0	15.0	16.9	18.9	20.9	22.8	24.8	26.8	28.8	30.8	32.8	34.8

We discovered that the variation in the V_{TH} of each ALD1101 IC is important. While the shape of the $I_{\text{NMOS}} - V_{\text{IN}}$ relationship in Supplementary Table 2 is the same for each ALD1101 IC, the V_{TH} variations lead to offsets, shifting the $I_{\text{NMOS}} - V_{\text{IN}}$ curves up or down. Remeasuring of all the data points in Supplementary Table 2 is not required, only one measurement is needed: using the same characterization setup described above, we find and record the V_{IN} that produces $I_{\text{NMOS}} = 16.9 \mu\text{A}$. The difference of this V_{IN} with the value recorded in Supplementary Table 2 (1500 mV) gives the offset to the entire $I_{\text{NMOS}} - V_{\text{IN}}$ curve of the specific ALD1101 used. We simply add this offset to all data points in the reference $I_{\text{NMOS}} - V_{\text{IN}}$ curve. We use the experimentally measured $\langle R_{\text{MTJ}} \rangle - I_{\text{MTJ}}$ curve of the specific sMTJ and the $I_{\text{NMOS}} - V_{\text{IN}}$ curve of the specific ALD1101 IC to determine the stochastic range of V_{IN} .

V. FPGA ARCHITECTURE

A. p-bit and MAC Unit

To evaluate the performance of the random number generators, we first implemented fully digital probabilistic bits (p-bits) on a Kintex UltraScale KU040 FPGA Development Board. A single p-bit consists of a tanh lookup table (LUT), a random number generator (RNG) and a comparator to implement Supplementary Eq. (1). In this work, we used 32-bit LUTs and RNGs. There is also a multiplier-accumulator (MAC) unit to compute Supplementary Eq. (2) from the neighbor p-bits and provide the input signal for the LUT. The p-bits are interconnected in a fixed hardware topology where the weights and biases, J_{ij}, h_i are stored in 10-bit registers and a digital multiply-accumulate operation with $m_j \in \{0, 1\}$ selects a particular weight or not. In the FPGA, we switch from a bipolar p-bit $m_i \in \{-1, 1\}$ to a binary formulation $m_i \in \{0, 1\}$ [5]. In our hybrid CMOS + sMTJ based p-bits, the architecture remains the same except the sMTJs serve as the clocks for the LFSRs and the p-bits.

B. RNG Unit

As shown in Algorithm 1 in the Supplementary information, we used three different types of RNGs: linear feedback shift register (LFSR), Xoshiro [47] and sMTJ-driven LFSR to compare the quality of randomness. We used 32-bit RNGs and compared the RNG outputs with the 32-bit LUTs to implement Supplementary Eq. (1). LFSR involves a linear shift operation on all the bits and an XNOR operation on some bits based on the selected taps. Unique seeds were used for each RNG and unique taps were used for RNGs in each p-bit block while ensuring maximal-length outputs. In contrast, a Xoshiro RNG involves linear shift and rotation operations on 32-bit words as well as XORing between subsequent words. In the all-digital versions, LFSR and Xoshiro RNGs are driven by digital clocks. However, in the hybrid design where sMTJs serve as the clocks for the RNGs, sMTJ + LFSR is considered as a new RNG unit. Using sMTJ-based p-bit removes the need for the 32-bit RNG and the 32-bit LUT.

C. Clocking and Sampling Unit

For the fully digital CMOS p-bits, each p-bit and its RNG is driven by system clocks generated on the low-voltage differential signaling (LVDS) clock-capable pins of the FPGA. These clocks are generated using Xilinx LogiCORE™ IP clocking wizard and mixed-mode clock manager (MMCM) module. Each of the five clocks operates at 15 MHz with shifted phases and is highly accurate with low jitter noise. In the Boltzmann learning example and for the NIST tests, to make these clocks comparable to sMTJs, we used frequency divider circuits to slow down the clocks to ≈ 2 kHz. For the sMTJ-driven p-bits, sMTJs replace the FPGA clocks and drive the p-bits and the RNGs externally using the peripheral module (PMOD) interface. FPGA samples the sMTJ outputs with a fast system clock of 75 MHz.

D. Data Programming and Acquisition Unit

We used MATLAB 2022a as the host program to read-write data to and from the FPGA through the USB-JTAG interface. MATLAB communicates with the FPGA board via AXI4 (Advanced eXtensible Interface 4) protocol where MATLAB works as the AXI master to drive a slave memory-mapped registered bank and Block RAMs (BRAM) inside the FPGA. We used airHDL [69], a memory management tool to assign the memory addresses for the register bank and the BRAMs. The weights J, h of the full adder circuit are programmed through MATLAB. In the inference and the Boltzmann learning example, the p-bit outputs were read from MATLAB as batches of sweeps that were initially sampled at 2kHz and stored in a BRAM. For the NIST tests, however, we sampled and stored all the sweeps in the BRAM at a much higher frequency (≈ 10 kHz) compared to the clock frequency of ≈ 2 kHz and then downsampled the data to a designated frequency. This procedure ensured that we did not lose any samples. MATLAB reads the BRAM data in burst mode and performs the downsampling.

VI. INFERENCE ON THE PROBABILISTIC FULL ADDER

Inside the FPGA, we construct digital p-bits that behave according to Supplementary Eq. (1) and interconnections between p-bits that behave according to Supplementary Eq. (2). Each p-bit has a PRNG, a LUT for the hyperbolic tangent function and 10-bit weights in fixed-precision, 1 sign, 6 integer, and 3 fractional bits (s[6][3]). Supplementary Eq. (2) is implemented by a multiply-accumulate unit inside the FPGA, whose multiplication reduces to simple multiplexing since a given weight J_{ij} is either taken or not if m_j is 0 or 1. An important consideration in ensuring the p-bit network reaches the equilibrium is the necessity of fast synapse times (τ_s) compared to neuron times τ_n [65]. In our context this requirement ($\tau_s < \tau_n$) is naturally satisfied because the combinational logic inside the FPGA, which computes Supplementary Eq. (2) with about 10 ns delays is orders of magnitude faster than both our deliberately slowed digital clocks and our sMTJs. In scaled and integrated implementations with fast p-bits with GHz fluctuations, this necessity requires careful design. In the case sMTJ clocks, there is also the theoretical possibility of parallel updates by simultaneously switching sMTJs. Practically this is not a concern due to the extremely low probability of such an event which would be washed over thousands of samples anyway. Our results with sMTJs in FIG. 2c,d and in FIG. 3c,d indicate that the sMTJs reproduce the ideal distributions well.

A full block diagram of the FPGA unit is shown in FIG. 3a. To rule out any spurious correlations, the starting states (seeds) used for the LFSR and Xoshiro are randomized for each p-bit. LFSRs also use unique sets of random taps while ensuring maximum-length outputs.

In this setting, for each RNG, we cumulatively sampled 10^6 states from the p-bits starting from a random initial state. A system state out of 5-p-bits can be defined from 0 to $2^N - 1$ such that state 0 is $p_1p_2p_3p_4p_5 = 00000$ and state 31 is $p_1p_2p_3p_4p_5 = 11111$. We define the single update of each p-bit according to Supplementary Eq. (1)-(2) as a sweep. Due to their digital nature, defining exact times to perform a sweep for LFSR and Xoshiro is straightforward. With a driving clock frequency of 2 kHz, we perform one sweep and then record it as a new state. However, sampling states from sMTJ-clocked LFSR p-bits is not straightforward due to the analog nature of fluctuations of the sMTJs. The relaxation time of the slowest sMTJ is ≈ 20 ms (50Hz) that is $40\times$ slower than the sampling frequency of 2 kHz. For this reason, we collect $40\times$ more data points from sMTJ-based p-bits and downsampled them to obtain independent samples. The oversampling and subsequent downsampling of sMTJ data is due to our $40\times$ faster readout process, not an inherent sMTJ limitation, and is implemented

to use the common sampler for each RNG setup. The sampling frequency can easily be adjusted, removing the need for oversampling and downsampling.

Algorithm 1: Learning the full adder on Deep BMs

```

Input : number of samples  $N$ , number of truth table lines  $T$ , epochs  $N_L$ , learning rate  $\varepsilon$ , regularization  $\lambda$ 
RNG: LFSR, Xoshiro, sMTJ-clocked LFSR
Output : learned weights  $J_{\text{new}}$  and biases  $h_{\text{new}}$ 
 $J_{\text{new}} \leftarrow 0.01 * \text{randi}([-1, 1]);$ 
 $h_{\text{new}} \leftarrow \text{randi}([-1, 1]);$ 
for  $i \leftarrow 1$  to  $N_L$  do
     $J_{\text{FPGA}} \leftarrow J_{\text{new}};$ 
    /* positive phase * /
    for  $j \leftarrow 1$  to  $T$  do
         $h_T \leftarrow$  clamping to truth table values;
         $h_{\text{FPGA}} \leftarrow h_T + h_{\text{new}};$ 
         $\{r\} \leftarrow \text{RNG}();$ 
         $\{m\} \leftarrow \text{FPGA}(N, \{r\});$ 
    end
     $\langle m_i m_j \rangle_{\text{data}} = \{m\} \{m\}^T;$ 
    /* negative phase * /
     $h_{\text{FPGA}} \leftarrow h_{\text{new}};$ 
     $\{r\} \leftarrow \text{RNG}();$ 
     $\{m\} \leftarrow \text{FPGA}(N \times T, \{r\});$ 
     $\langle m_i m_j \rangle_{\text{model}} = \{m\} \{m\}^T;$ 
    /* update weights and biases * /
     $J_{\text{new}} \leftarrow J_{\text{new}} + \Delta J_{ij};$ 
     $h_{\text{new}} \leftarrow h_{\text{new}} + \Delta h_i;$ 
end

```

VII. LEARNING THE FULL ADDER ON THE 32-NODE CHIMERA LATTICE

Boltzmann machines can be trained using the contrastive divergence algorithm. There are two phases during the training of Boltzmann machines as shown in FIG. 3a and Algorithm 1. The first one is the positive phase where the network operates in its clamped condition under the direct influence of the training samples. The next one is the negative phase when the network is allowed to run freely without having any environmental input. The update rules can be obtained by minimizing the KL divergence between the data and the model distributions [70]:

$$\Delta J_{ij} = \varepsilon \left(\langle m_i m_j \rangle_{\text{data}} - \langle m_i m_j \rangle_{\text{model}} - \lambda J_{ij} \right) \quad (6)$$

$$\Delta h_i = \varepsilon \left(\langle m_i \rangle_{\text{data}} - \langle m_i \rangle_{\text{model}} - \lambda h_i \right), \quad (7)$$

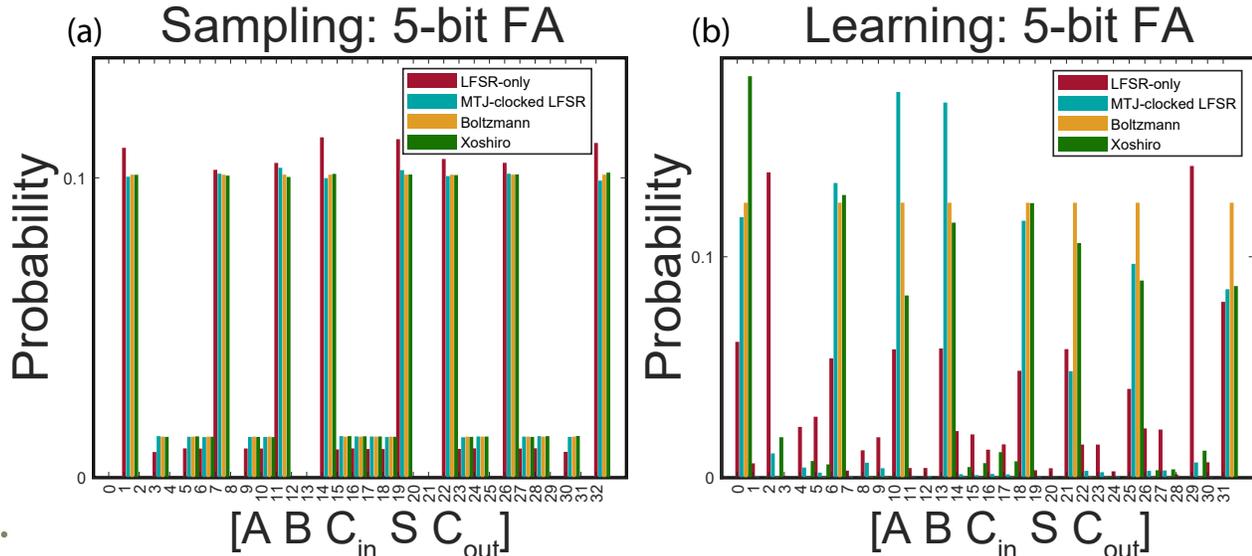
where ε is the learning rate, λ is the regularization parameter, $\langle m_i m_j \rangle_{\text{data}}$ is the average correlation between two neurons in the positive phase, and $\langle m_i m_j \rangle_{\text{model}}$ is the average correlation between two neurons in the negative phase.

As described in the previous section, five sMTJ-based p-bits are used here as clocks for digital p-bits with 32-bit LFSRs in FPGA and each LFSR starts with a unique seed. We also used 5 different sets of taps for the LFSRs, ensuring maximal-length output. The $\langle m_i m_j \rangle_{\text{data}}$ is obtained by clamping visible bits to the eight lines in the truth table of the full adder (FIG. 2b). Then the $\langle m_i m_j \rangle_{\text{model}}$ is calculated in the negative phase where the clamp is removed. We obtain the updated weights and biases using Supplementary Eqs. (6),(7) and repeat the weight learning for 500 epochs. We also naturally adopt the persistent contrastive divergence (PCD) algorithm that runs a continuous Markov Chain from the last state of the previous update to the next [71]. This is because the FPGA holds the previous state of the chain and continues sampling from that state as new weights are loaded. The hyperparameters we used while learning are as follows: inverse temperature $\beta = 1$, learning rate $\varepsilon = 0.003$, and regularization $\lambda = 0.005$.

The 32-node Chimera graph is bipartite. This means that p-bits in a Chimera topology can be updated in parallel in two blocks. In order to make our system resemble eventual, fully-asynchronous systems, we distributed our 5 available sMTJ clocks over these two blocks, ensuring that no sMTJ clock serviced two p-bits of the same block to avoid parallel updates (FIG. 3b bottom panel shows our clock distribution over the p-bits). For the fully digital setup, we distributed the digital clocks similarly. As in the inference experiments in the previous section, LFSR and Xoshiro RNGs were driven at 2 kHz. We used the same

5 sMTJ-based p-bits we characterized (Supplementary Figure 1) and sampled them at 2 kHz. To train the full adder on the LFSR and Xoshiro RNG-based p-bit network, we took 400 sweeps per epoch for a total of 500 epochs. For the sMTJ-clocked LFSR based p-bit network, we took 16000 sweeps per epoch and a total of 500 epochs. We took $40\times$ more sweeps for the sMTJ because they were sampled at 2 kHz ($40\times$ faster than their autocorrelation) in order to produce the same number of independent sweeps for all solvers.

In Supplementary Figure 4 we provide the full histograms (32-states) of the full adder for the sampling and learning experiments. Only parts of the histograms are shown in the main text for clarity. In Supplementary Figure 8 we show that this bias is not random and never goes away, rather it is a consistent systematic bias by starting the LFSRs at different uniform random initial conditions (seeds) and different maximum-length taps.



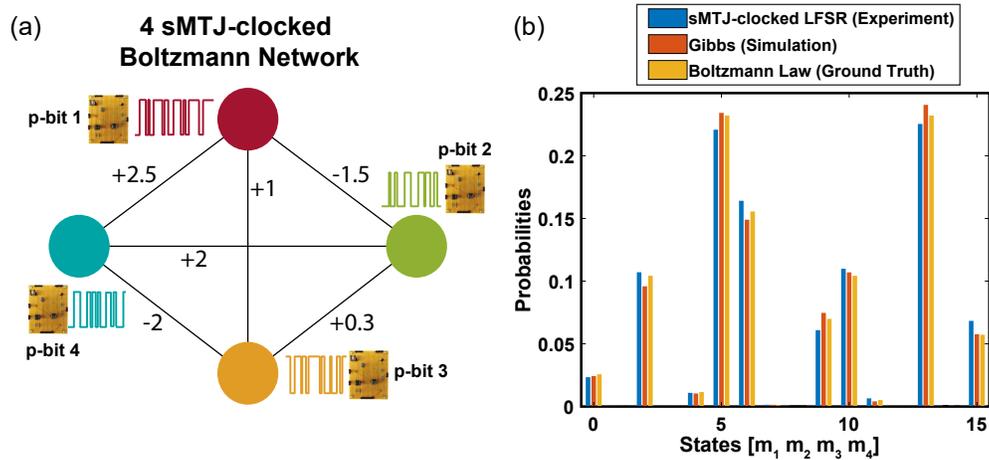
Supplementary Figure 4. Extended data for probabilistic sampling and learning experiments, showing all 32 states of the 5-bit full adder for LFSR and sMTJ-clocked LFSR. (a) Sampling data is taken at 10^6 sweeps, (b) Learning data is taken at epoch = 400 where the number of sweeps per epoch = 400 for LFSR-only and the number of sweeps per epoch = 16000 for the sMTJ-clocked LFSR. (c) Sampling data for Xoshiro (d) Learning data for Xoshiro

VIII. UPDATE ORDER INVARIANCE OF BOLTZMANN NETWORKS

In Supplementary Figure 5, we perform additional experiments to study how undirected Boltzmann networks are invariant to update orders in the system, significantly easing experimental difficulties in scaled-out implementations of p-bit networks. In this experiment, a fully connected undirected p-bit network had nodes clocked asynchronously with independent sMTJs. Even with only 10^4 samples, a very close match is seen between the distributions corresponding to sMTJ-clocked LFSRs and Gibbs sampling in simulation. It is important to note that Gibbs sampling was performed with *randperm* updates, where a new update order (out of $4!=24$ possibilities) was sampled (without replacement) at each iteration. Despite this highly random updating scheme, both Gibbs (ideal simulation) and the sMTJ-clocked system eventually reach the ground truth, represented by the Boltzmann distribution. This remarkable feature of update invariance in Boltzmann networks [66] significantly eases fabrication difficulties in eventual sMTJ-based large-scale p-bit networks. Going further, the investigation of “time-to-equilibrium” that determines the model averages and correlations in Algorithm 1 remains to be one of the future challenges of the field.

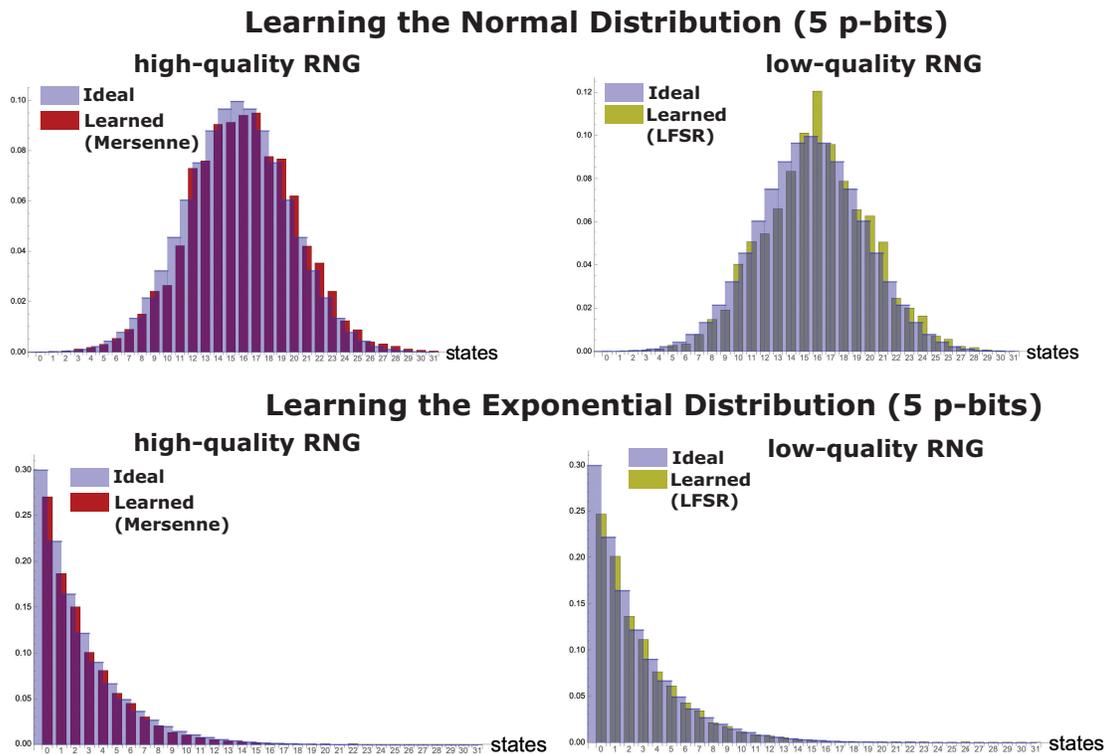
IX. SAMPLING FROM ARBITRARY DISTRIBUTIONS

In this section, we show how the CD algorithm can be used to create networks of p-bits that can sample from arbitrary distributions up to M-bit accuracy. We start by defining a PDF for a given distribution. As an example, we choose a normal distribution with $\mu = 16$ and $\sigma = 4$ to be approximated by a $N = 5$ p-bit network with $2^N = 32$ states. Once the PDF is specified, one approach is to create a truth table matrix of size $V = NT \times N$, where the rows of the truth table are repeated according to the specified PDF. Assuming a fully visible network, we then calculate the data correlations from the truth table, $V^\dagger V$. We then perform the standard contrastive divergence algorithm. Supplementary Figure 6 shows inference and learning of a normal and exponential distribution with Mersenne Twister-based high-quality RNGs and LFSR-based low-quality RNGs. Even though the Mersenne-based learning looks marginally better, we do not see a strong LFSR bias in such low-dimensional

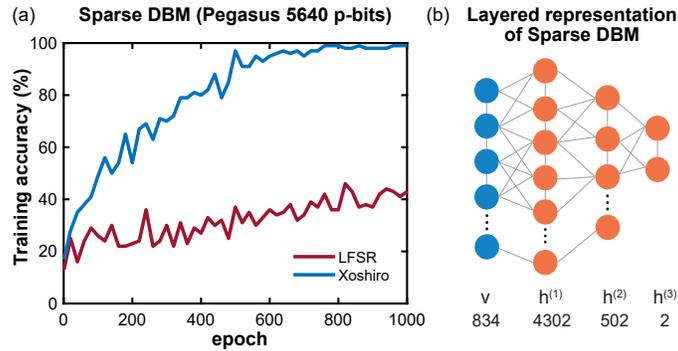


Supplementary Figure 5. (a) Fully connected undirected Boltzmann network with 4 nodes. Different sMTJs drive the LFSRs in the p-bits corresponding to each node. (b) Experimental results of 4 sMTJ-clocked LFSR-based p-bits ($1e4$ iterations) performing probabilistic inference. This asynchronous clocking of LFSR-based p-bits matched the random Gibbs Sampling simulation results ($1e4$ iterations), which also matched the Boltzmann Law ground truth for these undirected networks. Note that the Gibbs simulation randomized the update order of the network, choosing one of $4!=24$ possible permutations randomly at each iteration. The weight matrix is shown on the nodal interconnections, and the bias was 0 for p-bit 1, 3 and 4, but set to 0.3 for p-bit 2.

learning tasks. Next, we test the difference between low and high-quality RNGs in much larger networks including MNIST handwritten digits.



Supplementary Figure 6. Networks of p-bits can be connected to sample from arbitrary probability distributions: on the left of the top panel is a network of 5 p-bits approximating a normal distribution where training and inference are both done with high-quality Mersenne Twister-based PRNGs. The blue-filled lines are the exact probability density functions (PDF), discretely sampled in both figures. Bottom panel shows a similar result to learn the exponential distribution. Histograms show the inference on a 5 p-bit network that has been trained using the contrastive divergence algorithm. On the right is the same result (inference followed by learning) using 5 p-bits using LFSRs only (32-bit with random taps and initial conditions). All histograms are obtained with 50,000 samples. All models are trained on a 5-bit all-to-all fully visible network (weights and biases not shown).



Supplementary Figure 7. (a) Training accuracy of 100 handwritten digits from MNIST with a sparse Deep Boltzmann Machine (Pegasus 5640 p-bits) for 1000 epochs using both LFSR and Xoshiro. This shows a significant discrepancy in training accuracy depending on whether p-bits are implemented with LFSR or Xoshiro. Despite being identical excluding the RNG implementation, the former does not suffice due to the low accuracy at around 40% even after 1000 epochs, the latter reaches 100% accuracy with the same number of epochs and the same set of hyper-parameters. (b) Illustration of sparse DBM (Pegasus 5640 p-bits) with 1 layer of visible units and 3 layers of hidden units where both the inter-layer and intra-layer connections are allowed.

X. TRAINING MNIST WITH DEEP BOLTZMANN MACHINE: LFSR VS XOSHIRO

We now present a more complex task of training a subset of MNIST handwritten digits [72] using Algorithm 1 that also suffers from the low-quality randomness of LFSR. Pegasus 5640 p-bits [73] is chosen as the sparse DBM where we randomly distribute the visible and hidden units which yield multiple hidden layers as shown in Supplementary Figure 7b. The details of this method are described in [58]. Here, we choose a subset of MNIST data including 100 images with no down-sampling and train them in 10 mini batches using the CD algorithm for both LFSR and Xoshiro. All the hyperparameters remain the same for these two cases which are as follows: inverse temperature $\beta = 1$, learning rate ε varies linearly from 0.03 to 0.003 for 1000 epochs.

Supplementary Figure 7a shows the comparison of training accuracy of 100 MNIST digits between the implementation of LFSR and Xoshiro. While LFSR-based training barely reaches 40% accuracy in 1000 epochs, the Xoshiro-based training goes to 100% strongly indicating how the quality of randomness matters in terms of training such a large network.

XI. ANALYZING LFSR BIAS OVER DIFFERENT TAPS AND INITIAL CONDITIONS

Supplementary Figure 8 presents a detailed analysis of LFSR as a random number generator. LFSR RNG quality is measured by its ability to perform bias-free sampling on the probabilistic full adder. Across different seeds and taps, even for 10^6 iterations, a consistent bias is always seen as demonstrated by the variations in histogram peaks.

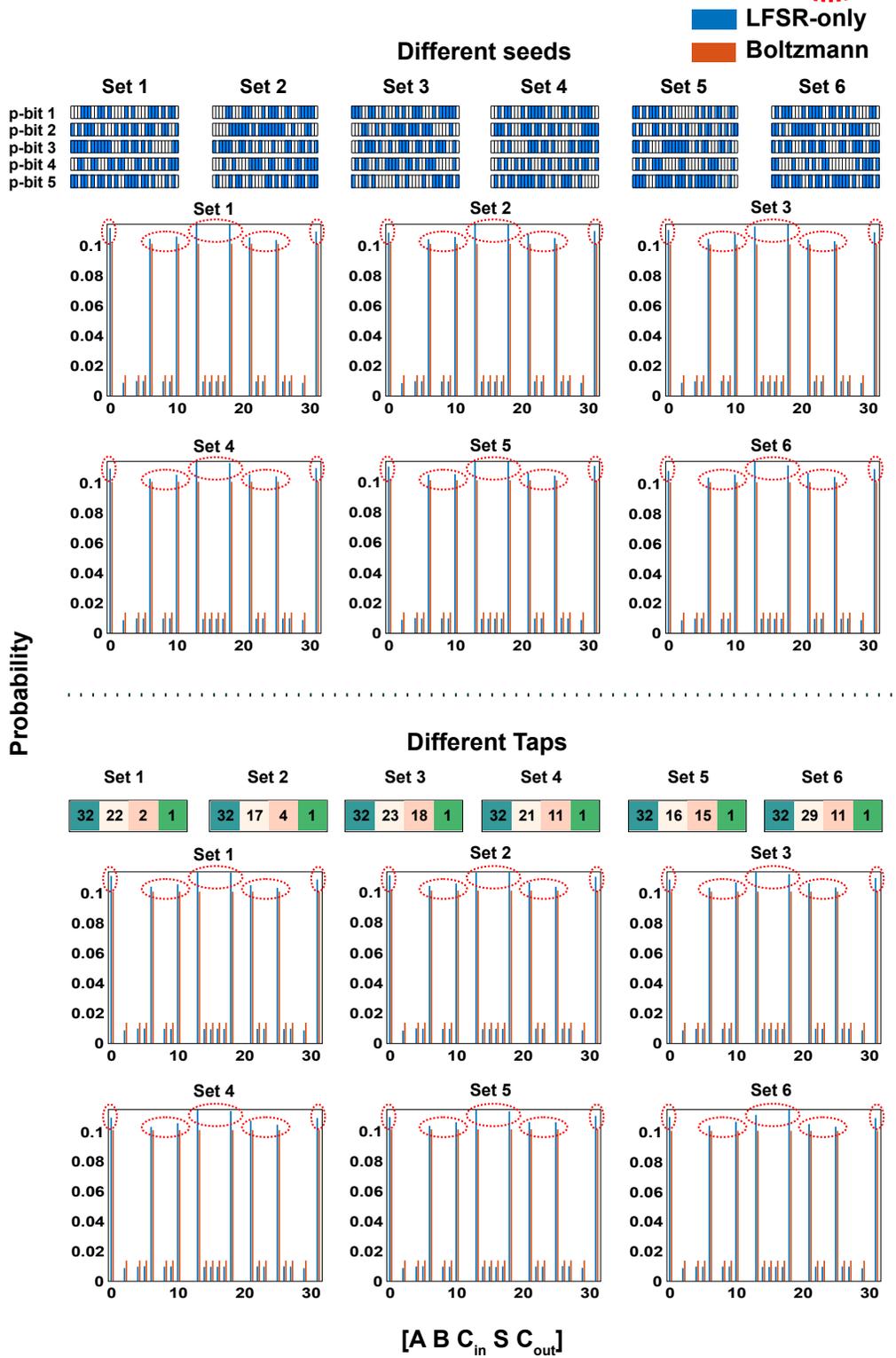
XII. NIST TESTS ON XOSHIRO, LFSR AND SMTJ CLOCKED LFSRS

NIST tests are widely used to evaluate the quality of randomness, so we conducted standard randomness tests on the bitstreams generated by the LFSR, Xoshiro, and LFSR+sMTJ using the NIST Statistical Test Suite [49]. We applied all 16 different NIST tests to the bitstreams. For LFSR and Xoshiro, we used MATLAB to generate bitstreams since given taps and initial conditions these bitstreams are fully reproducible.

We used a modified experimental setup to obtain the bitstreams for LFSR+sMTJ. First, the inverse temperature value β in Eq. (1) was set to 0 to get 50/50 fluctuations. Second, we sampled bitstreams of 650000 bits with a 10 kHz sampling rate in BRAM blocks of the FPGA. Then, we downsampled the bitstream by a factor k . This is equivalent to sampling the bitstream with frequency f_k that $f_k = 10000 \text{ Hz}/k$. The length of the bitstream then becomes $650000/k$. The reason for this downsampling is to obtain independent samples from the sMTJs whose fluctuations times are far above 100 microseconds.

We used p-bit #3 to generate the bitstreams for LFSR+sMTJ. The result ‘Random’ represents that the bitstream passes the tests whereas ‘Non-Random’ means the bitstream fails. Supplementary Table 3 summarizes the results of the NIST tests for the LFSR+sMTJ with $k = 601$ and $k = 2001$, Xoshiro and LFSR. The results without oversampling issue show the good quality of randomness generated by LFSR+sMTJ when $k=601$ or $f_k = 16.63 \text{ Hz}$, corresponding to a period of 60 milliseconds, of the same order of our sMTJ fluctuations reported in Supplementary Table 1. The result of $k = 2001$ fails Maurer’s Universal Test (Test #9) because the downsampled bitstream with length $650,000/2001 \approx 324$ is not long enough for applying that test. Corroborating our results in the main text, LFSR+sMTJ and Xoshiro pass all the tests, while LFSR fails one test.

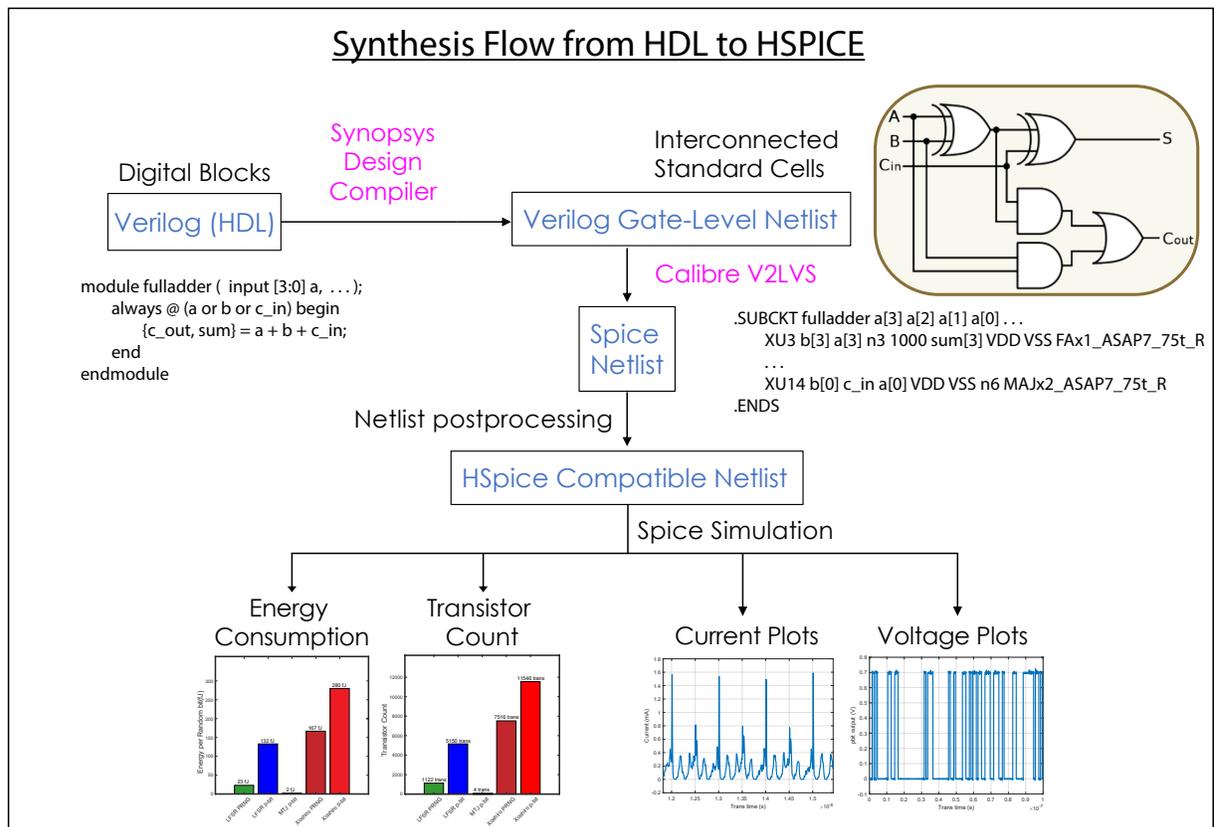
Systematic bias in sampling Full Adder with LFSR Bias



Supplementary Figure 8. Demonstration of consistent systematic bias in sampling the 5 p-bit full adder using multiple sets of seeds and taps for the LFSR. The seeds are chosen uniform randomly and the taps are chosen to provide the maximum length of the LFSR. Irrespective of the seeds and taps, the bias never goes away, indicating a systematic bias.

Supplementary Table 3. Results of tests on bit streams specified in standard NIST SP800-22a

Test #	Test Name	Sub-tests	LFSR + sMTJ	LFSR+sMTJ	Xoshiro	LFSR
			$k = 601$ $f_k = 16.63$ Hz	$k = 2001$ $f_k = 5.00$ Hz		
1	Frequency	1	Random	Random	Random	Random
2	Frequency within a Block	1	Random	Random	Random	Random
3	Runs	1	Random	Random	Random	Random
4	Longest run of ones	1	Random	Random	Random	Random
5	Rank	1	Random	Random	Random	Random
6	Discrete Fourier Transform	1	Random	Random	Random	Random
7	Non-overlapping T. M.	1	Random	Random	Random	Random
8	Overlapping T.M.	1	Random	Random	Random	Random
9	Maurer's Universal	1	Random	Non-Random	Random	Random
10	Linear complexity	1	Random	Random	Random	Non-Random
11	Serial	2	Random	Random	Random	Random
12	Approximate Entropy	1	Random	Random	Random	Random
13	Cumulative sums	2	Random	Random	Random	Random
14	Random Excursions	8	Random	Random	Random	Random
15	Random Excursions Variant	18	Random	Random	Random	Random

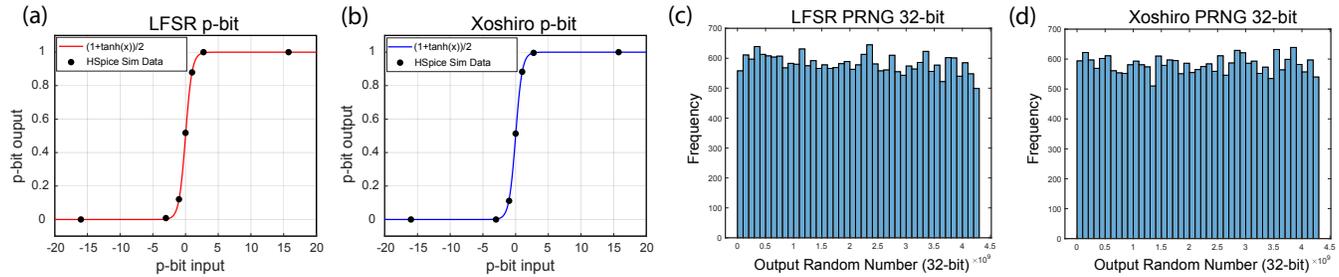
**Supplementary Figure 9.** Flowchart highlighting the key steps in the digital synthesis flow from hardware description languages such as Verilog all the way down to SPICE.

XIII. SYNTHESIS FLOW

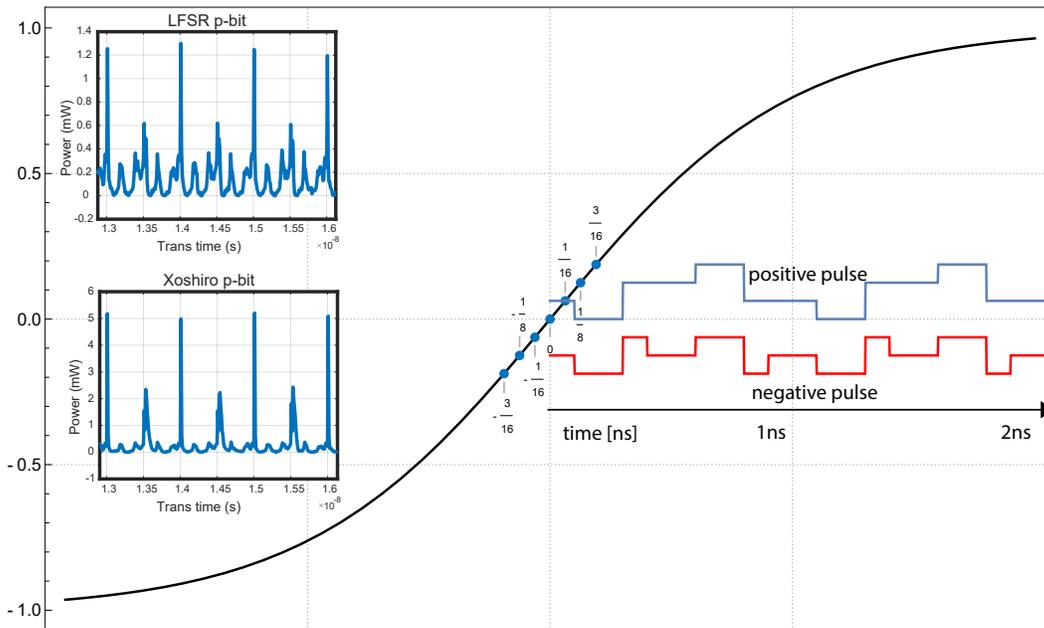
The synthesis process followed involves the conversion of HDL codes modeling the digital p-bit to a functionally equivalent SPICE netlist (Supplementary Figure 9). This HDL-to-SPICE conversion first involves using Synopsys Design Compiler (DC) to generate the HDL gate-level netlist from the open-source ASAP7 PDK library files. Subsequently, we used Calibre's Verilog-to-LVS (V2LVS) tool to translate the gate-level netlist into a SPICE compatible netlist. After post-processing using a custom Mathematica script to get the netlist HSPICE compatible, we run transient simulations to obtain energy consumption, transistor counts, and current and voltage plots. With Synopsys DC, we used the regular voltage threshold (RVT) database files, and with V2LVS we used the RVT CDL files. For the HSPICE simulations, a VDD value of 0.7 V was used, and clock frequencies from

100 MHz to 1 GHz were tested, with 1 GHz being used for all results reported in this work.

XIV. FUNCTIONAL VERIFICATION OF P-BITS AND PRNGS



Supplementary Figure 10. (a) Verifying the functionality of 32-bit LFSR and (b) Xoshiro-based p-bits synthesized using ASAP7 PDK. Input current is converted from the 2's complement s43 representation to decimal equivalent. The y-axis indicates the probability of the p-bit being in a 1 state calculated over 2500 clock cycles of a transient p-bit simulation generating random bits. (c) Verifying the functionality of 32-bit LFSR and (d) Xoshiro PRNGs synthesized using ASAP7 PDK. 32-bit binary outputs were mapped to their decimal equivalent, and a 25000 clock cycle transient simulation generated random 32-bit values where $f_{\text{clk}} = 1$ GHz.



Supplementary Figure 11. Input profile to activate the LUT: 1 GHz positive and negative pulse are applied to the input of the p-bit to activate LUT transistors. Insets show representative transient simulations for instantaneous power consumption for LFSR and Xoshiro-based p-bits.

As seen from Supplementary Figure 10a,b, the experimental data obtained from HSPICE by varying the p-bit inputs (8-bit LUT input) of the synthesized circuit falls on the theoretically expected $1 + \tanh(x)/2$ for both LFSR and Xoshiro-based p-bits. Varying the p-bit input allows us to tune the probability with which the p-bit fluctuates in accordance with the sigmoid seen in Supplementary Figure 10a,b. In Supplementary Figure 10c,d, we observe that across 25000 experimental samples, the distribution of outputs obtained from the digitally synthesized PRNG (LFSR or Xoshiro) are uniformly random. These two results verify the functionality of the digital p-bit synthesized using ASAP7 by a transistor-level simulation performed in HSPICE.

XV. POWER ANALYSIS OF P-COMPUTERS

A. Synapse power

It is important to note that our calculations do not include any power analysis for the synapse (Supplementary Eq. (2)) so far. Earlier estimations [4] indicated that the synapse power is at least 50% of the overall power consumption. Depending on the implementation, for example, using analog crossbars or in-memory computing techniques, the synapse power could show a large degree of variation and we do not explore these possibilities in this paper.

B. Digital p-bit power

Insets of Supplementary Figure 11 show high resolution, representative section of the power plots for LFSR and Xoshiro p-bits. Energy analysis was performed by integrating the power over the transient time followed by averaging over 100 clock cycles of a 1 GHz clock, using trapezoidal numerical integration (`trapz`) in MATLAB.

In order to estimate the energy contribution of the LUT to the energy of generating a random bit, we vary the least significant 3 bits of the p-bit input to keep the LUT actively switching to simulate normal operating conditions during probabilistic computations. We do this in two different ways by generating 2 switching sequences of the form 0000xxx and 1111xxx, where “x” switches between 0 and 1 (see Supplementary Figure 11 for the pulse shapes). The first switching pattern is shown by the positive pulse in blue, where the LUT is traversing the sigmoid just above the zero point. The second switching pattern shown by the negative pulse in red has the LUT switching between values just below the zero point. We choose these input variations to have a 1 GHz frequency and measure the average power and energy dissipation over 100 clock cycles. For the positive pulse, the results are shown in FIG. 4b. For the negative pulse we obtain an energy consumption for a 32-bit LFSR p-bit as 119 fJ, and that of a 32-bit Xoshiro p-bit as 267 fJ, similar to what we observed for the positive pulse, reported in the main text, FIG. 4b.

XVI. BENCHMARKING AND PROJECTION OF P-BITS: ROADMAP

In this section, we provide projections and benchmarks for probabilistic inference and sampling hardware at device, circuit and systems levels which are summarized in Supplementary Table 4, below.

Device-level: At the sMTJ level, a key performance metric is the speed of fluctuations which can be measured by autocorrelation and magnetic relaxation times. Here, we seek order-of-magnitude estimates and represent average fluctuations by a single number τ for simplicity. There are two main methods to design sMTJs, one by employing nanomagnets with perpendicular anisotropy (PMA) and another by employing nanomagnets with in-plane anisotropy (IMA). PMA magnets of the type we use in this paper typically have slow fluctuation rates compared to IMA [22, 24]. PMA magnets whose energy barriers are around 10-15 $k_B T$ fluctuate in the millisecond range [28, 29]. On the other hand, recent device-level experiments using IMA magnets have shown fluctuations with 1-10 nanosecond rates [19–21]. sMTJs made out of IMA magnets possess other favorable properties such as bias-independence [26] to build robust p-bit circuits.

Circuit-level: Despite advances at the individual device level, p-bit circuits using sMTJs have not yet caught up with the fastest sMTJs. Even though we use a different current-mirror topology for the p-bit circuit in this paper, similar to earlier work, our p-bits use PMA MTJs with milisecond fluctuations. A recent report showed the fastest p-bit circuits with microsecond fluctuations, using IMA magnets [74]. Reaching nanosecond fluctuations with p-bit circuits might require integrated solutions, of the type that are sought initially in Ref. [32]. Given that CMOS circuitry could operate at picosecond timescales, there should be no fundamental obstacles to building GHz p-bit circuits.

System-level: At the system level, two main performance metrics for p-computers have been identified. One is the number of nodes in the network, N . The other one is sampling throughput that measures the number of probabilistic decisions taken by a system. Highly optimized standalone GPUs/TPUs of similar size graphs provide a sampling throughput of 10 flips/ns, establishing an optimized conventional baseline [62, 63, 75, 76]. These chips consume around 100W of power. Given this GPU/TPU background, we provide established experimental data and projections for CMOS and CMOS+sMTJ-based probabilistic computers.

The FPGA columns represent our fully digital FPGA work with different quality RNGs (LFSR and Xoshiro). Depending on the bit precision and network connectivity, these digital solvers can sustain up to $N = 10^4$ p-bits in hardware [5, 58]. Running in parallel with around 10 MHz frequencies, they reach 100 flips/ns [5] in sampling throughput, about an order of magnitude higher than typical GPU and TPUs. On the other hand, the sMTJs with perpendicular magnetic anisotropy (PMA) used in our present heterogeneous p-computers currently have fluctuation times around $\tau = 1$ ms. However, near-term projections with $N = 10^4$ p-bits using sMTJs with in-plane magnetic anisotropy (IMA) ($\tau \approx 1$ ns [19–21]) can reach 10^4 flips/ns in sampling throughput. In the case of heterogeneous p-computers driven by sMTJs, the external p-bits can drive a large number of *digital* p-bits inside the FPGA. In such a case, around $N=10,000$ digital p-bits can be driven by sMTJs and the sampling throughput would be limited by the synapse time inside the FPGA (or the digital ASIC). This number shows that even in this modest scale, heterogeneous computers can already provide computational advantages over-optimized typical TPU/GPUs. Ultimate,

fully-integrated and sMTJ-based computers with $N = 10^6$ p-bits can reach sampling throughputs of 10^6 flips/ns, 5 orders of magnitude faster than typical TPU/GPUs. TPU/GPU references discussed have been plotted on a power consumption versus sampling throughput scale in [77] and [74].

For sampling throughput projections at large scale, we use N/τ , assuming each p-bit flips independently of each other. This basic formula assumes that each flip is communicated to neighboring p-bits *before* a new flip is attempted, as otherwise flips may not be useful. If the network density scales as $O(k \cdot N)$ with some small k corresponding to sparse networks, the fast communication assumption is warranted and ideal parallelism can be achieved.

For FPGA-based p-computers, the total power consumption is around 30W, including peripheral and unrelated circuits beyond the digital synthesis of our design. For heterogeneous computers of the type we consider in this work (5 MTJs + FPGA), the total power is also dominated by the FPGA power and is also around 30W. Unlike the present work where we used sMTJs to clock digital PRNGs such as LFSRs, in the future we envision the sMTJ-based analog p-bits as standalone blocks interacting through a CMOS underlayer without any seeding of CMOS PRNGs. In terms of power estimates for such fully sMTJ-based p-computers, detailed circuit simulations with experimentally established parameters indicate a power consumption of $10 \mu\text{W}$ per p-bit [68]. For fully sMTJ-based p-computers with $N = 10^6$, this would indicate a total p-bit power of 10 W, with an additional 10 W estimated synapse power [4, 74]. Considering how present day MRAM technology has been scaled up to 1 Gbit densities [78], integrating about 10^6 p-bits on top of CMOS should be reasonable.

Level of Comparison	GPUs/TPUs	← THIS WORK →				NEAR-TERM PROJECTION		LONG-TERM PROJECTION
		FPGA p-computer		Hetero FPGA+sMTJ p-computer		ASIC all-digital	Heterogeneous sMTJ+ASIC	
		LFSR-based	Xoshiro-based	PMA	IMA			IMA
DEVICE: sMTJ	N/A	N/A	N/A	$\tau \approx 1 \text{ ms}$ [28]	$\tau \approx 1 \mu\text{s}$ [27]	$\tau \approx 1 \text{ ns}$ [19]	N/A	$\tau \approx 1 \text{ ns}^\dagger$
CIRCUIT: p-bit	N/A	$(\tau)^{-1} \approx 10 \text{ MHz}$	$(\tau)^{-1} \approx 10 \text{ MHz}$	$\tau \approx 1 \text{ ms}$ [28]	$\tau \approx 1 \mu\text{s}$ [74]	$\tau \approx 1 \text{ ns}^\dagger$	$\tau \approx 1 \text{ ns}^\dagger$	$\tau \approx 1 \text{ ns}^\dagger$
SYSTEM: Number of p-bits (N)	N/A	$N \approx 10^4$ [5]	$N \approx 10^4$ [58]	$N = 32$ (*)	$N \approx 10^4$ (*) †	$N \approx 10^4$ (*) †	$N \approx 10^5$ †	$N \approx 10^6$ †
SYSTEM: Sampling throughput (flips/ns)	≈ 10	≈ 100	≈ 100	$\approx 32 \times 10^{-6}$	≈ 10 †	$\approx 10^4$ †	$\approx 10^5$ †	$\approx 10^6$ †
RNG Quality	PRNG/ High	PRNG/ Low ^o	PRNG/ High	TRNG/ High	TRNG/ High	TRNG/ High	PRNG/ High	PRNG/ High

Supplementary Table 4. Benchmarking probabilistic hardware from device, circuit and system perspectives. [Benchmarking probabilistic hardware from device, circuit and system perspectives. For device comparisons, we focus on experimentally demonstrated sMTJ fluctuations. p-bits are circuits that use sMTJs to produce binary stochastic neurons with tunable probability with fluctuations at τ^{-1} rates. At the system level, we focus on the number of p-bits in a network (N) and sampling throughput, which is given by N/τ , for asynchronous systems with fast synapses computing Supplementary Eq. 2 (see text). Also at the system level, we report published data for GPUs/TPUs handling similar probabilistic sampling tasks. Projections are shown using (†). (*) In heterogeneous computers of the type we consider in this paper, external sMTJ-based p-bits can drive a large number of digital p-bits in an FPGA or an ASIC. ^o RNG quality is deemed low / high for *sampling* problems rather than combinatorial optimization problems for which LFSR-based PRNGs seem sufficient [5].