

## SAMPLING AS THE COMPUTE

# Thermodynamic and Probabilistic Hardware for Generative and Embodied AI: A Survey

A generative model draws samples from a distribution. This report reviews hardware that draws them from physics instead of computing them first.

**David Jean Charlot, PhD**

Dean of Physical AI · The Charlot Lab, Institute for Physical AI @ BMI

Correspondence: [contact@physicalai-bmi.org](mailto:contact@physicalai-bmi.org) · [physicalai-bmi.org](https://physicalai-bmi.org) · The Charlot Lab: [labs.physicalai-bmi.org/charlot](https://labs.physicalai-bmi.org/charlot)

Interactive companion: [physicalai-bmi.org/research/charlot-lab#topic-thermo](https://physicalai-bmi.org/research/charlot-lab#topic-thermo)

**ABSTRACT.** A generative model, reduced to its core operation, draws a sample from a probability distribution. A conventional accelerator does this in two stages: it computes the distribution with matrix multiplication, then draws from it with a random-number generator. Most of the energy is spent in the first stage, on a quantity that a random draw immediately collapses. Thermodynamic and probabilistic hardware removes the first stage. It encodes the distribution in the physics of a device and lets thermal fluctuation produce the sample, so the noise a digital design spends energy to suppress becomes the computation. This report reviews the approach. Section 2 states the review method. Section 3 gives the energy-based-model formulation and the sampling algorithms that hardware realizes physically, with the equations. Section 4 develops a first-order energy comparison between computed and physical sampling. Sections 5 to 7 survey thermodynamic sampling units, probabilistic bits, and Ising machines, and Section 8 states the reversible-computing limit set by Landauer's principle. Section 9 gives the current status: the reported energy advantages are large and the silicon is early. The report reports no new experimental measurements.

## 1. The stochastic half of a policy

An embodied system does not only evaluate learned functions. It also samples. It draws an action from a distribution over actions, a hypothesis about a world it observes only in part, a set of candidate futures, and the successive denoising steps of a diffusion-based controller. Generative modeling in general is the drawing of samples from a learned distribution. Conventional hardware does this indirectly. It uses dense linear algebra to compute the parameters of the distribution, a vector of logits or a scalar energy, and then a separate step draws a random sample from those parameters. The costly stage is the matrix multiplication, and the quantity it produces is used only to seed a random draw. For a system that runs on a body's power budget this is a poor allocation of energy, and it raises a direct question: can the sample be produced by a physical process that is already random, without computing the distribution first?

## 2. Scope and method

This is a review of published theory, algorithms, and hardware for thermodynamic and probabilistic computing, with attention to the generative and embodied workloads relevant to Physical AI. It draws on peer-reviewed work, preprints, and primary hardware disclosures, and favors the earliest source for each method so the lineage is visible. The energy comparison in Section 4 uses the same first-order model as the companion report on multiply-free arithmetic: fixed per-operation coefficients from a 45 nm survey<sup>[21]</sup>, counting arithmetic and ignoring control and interconnect. Two limitations apply throughout. The reported efficiency figures for thermodynamic hardware come largely from prototypes and simulation, not volume parts; Section 9 marks this. And the model quality of physics-based samplers is quoted from the originating work, not re-measured here.

## 3. Energy-based models and physical sampling

An energy-based model defines a distribution through an energy function  $U$ . The probability of a configuration  $x$  at temperature  $T$  is the Boltzmann form

$$p(x) = (1/Z) \cdot \exp(-U(x) / T) , \quad Z = \sum_x \exp(-U(x)/T) \quad (1)$$

so low-energy configurations are the probable ones and the temperature sets how sharply the probability concentrates on them<sup>[22]</sup>. This is the object of statistical mechanics. The Ising model of interacting spins, from 1925<sup>[1]</sup>, is an energy-based model; the Hopfield network stores memories as energy minima<sup>[2]</sup>; and the Boltzmann machine learns a distribution by adjusting the couplings of a stochastic Ising-like network<sup>[3]</sup>. If a device is built so that its physical energy landscape is the model's  $U$ , then bringing it to thermal equilibrium and reading its state produces a sample from  $p(x)$ . The sample is measured, not computed.

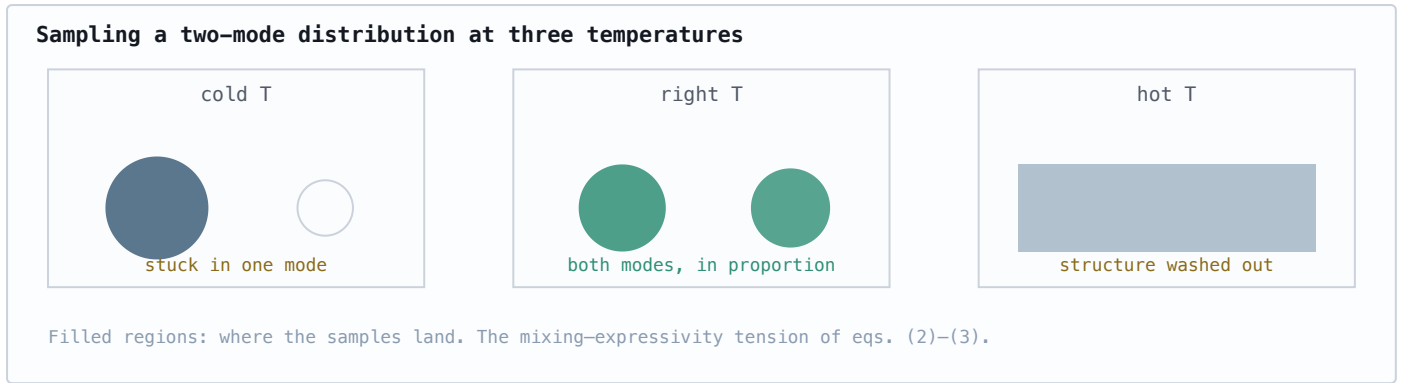
Reaching equilibrium is done by a Markov chain. Two forms, one continuous and one discrete, recur in hardware. Overdamped Langevin dynamics moves a continuous state downhill in energy under a thermal force,

$$dx = -\nabla U(x) \cdot dt + \sqrt{(2T)} \cdot dW \quad (2)$$

and the Metropolis rule accepts a proposed discrete move with probability<sup>[4]</sup>

$$a = \min(1, \exp(-\Delta U / T)) \quad (3)$$

Both have the target distribution (1) as their stationary distribution. Their difficulty is governed by a single tension, shown in Figure 1. At low temperature a rugged landscape traps the chain in one mode, so the sample set misrepresents the distribution: poor mixing. At high temperature the chain moves freely but the structure of the distribution washes out: poor expressivity. The useful regime lies between. Simulated annealing lowers the temperature on a schedule to find minima<sup>[7]</sup>; diffusion models transform noise into structure through a sequence of denoising steps and so avoid the mixing barrier<sup>[19,20]</sup>, an idea recent hardware adopts directly<sup>[15]</sup>.



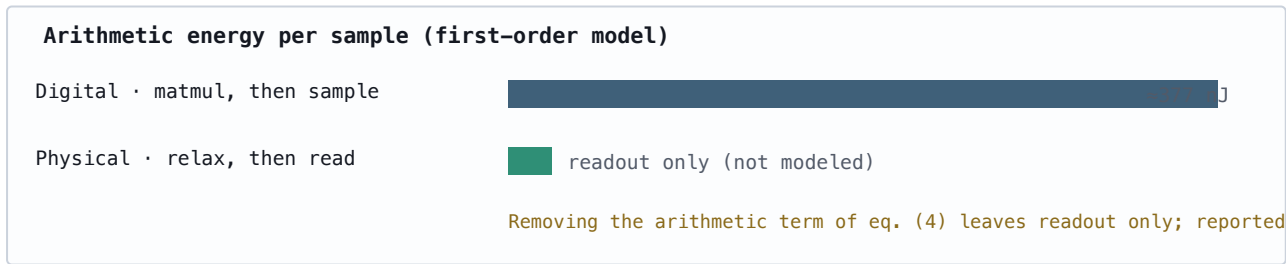
**Figure 1.** The mixing-expressivity tension in sampling a two-mode distribution. Cold temperatures trap the chain in one mode; hot temperatures erase the structure; an intermediate temperature reproduces both modes in their correct proportion.

#### 4. The energy of computing a sample versus drawing one

Consider drawing one sample from a model of  $N$  parameters by Markov-chain Monte Carlo, which evaluates the energy  $K$  times, once per chain step. On digital hardware each energy evaluation is a matrix–vector product, so the cost of a sample is

$$E_{\text{digital}} \approx K \cdot N \cdot (E_{\text{mul}} + E_{\text{add}}) \quad (4)$$

On a device whose physics is the energy landscape, the chain steps are performed by the hardware's own relaxation and cost essentially the readout, so the arithmetic term of (4) is removed. Figure 2 shows the two on a common axis using the arithmetic coefficients of the companion report and a modest model ( $N = 4096$ ,  $K = 20$ ). The developers of thermodynamic sampling hardware report advantages of order  $10^4$  against a GPU on sampling workloads<sup>[11]</sup>; the first-order model here reproduces a gap of that scale for the computed-versus-physical comparison, though the two figures are not measured the same way and should not be equated.



**Figure 2.** Arithmetic energy to draw one sample under the first-order model of eq. (4), for a model of 4096 parameters and 20 chain steps. The digital bar ( $\approx 377$  nJ) is a model estimate of the matrix-multiply term. The physical sampler removes that term; its residual cost is readout-dominated and is not modeled here. The  $\sim 10^4\times$  device advantage is the developers' separately reported result<sup>[11]</sup>, measured differently.

#### 5. Thermodynamic sampling units

A thermodynamic sampling unit builds an energy-based model into a physical substrate and reads its thermal fluctuations as samples. The design premise is that generative inference reduces to sampling, and that a circuit engineered to sample from a programmable distribution by controlled physical noise can perform that inference at a small fraction of the energy of a digital accelerator that computes the distribution first. Two efforts define the current commercial frontier. One has published prototype units, an open modeling library, and the order- $10^4$  energy figure noted above<sup>[11]</sup>. Another has reported the tape-out

of an early thermodynamic chip for datacenter use. Both are recent, and the figures are from prototypes and analysis rather than production silicon.

## 6. Probabilistic bits

A complementary approach builds the elementary stochastic primitive directly. A probabilistic bit, or p-bit, fluctuates between 0 and 1 with a controllable probability. The most developed realization uses a stochastic magnetic tunnel junction, a low-barrier nanomagnet whose magnetization flips under thermal agitation at gigahertz rates, so a single compact device supplies randomness that would otherwise cost thousands of transistors<sup>[12]</sup>. Networks of coupled p-bits implement Ising models and perform Boltzmann sampling natively, and have run optimization and inference tasks, including integer factorization on a hardware p-bit array<sup>[13]</sup>. The approach has been scaled to a programmable machine of a million p-bits performing on the order of  $10^{12}$  state updates per second<sup>[14]</sup>, and extended toward diffusion-style generation<sup>[16]</sup>. Because p-bits are built from CMOS-compatible magnetic devices, the approach can use the existing fabrication base.

## 7. Ising machines

When the task is combinatorial optimization rather than generation, the same physical principle appears as the Ising machine: encode the objective as the energy of a spin system and relax the system toward its minimum. Coherent Ising machines realize this with networks of optical parametric oscillators<sup>[17]</sup>; quantum annealers use adiabatic evolution toward a ground state<sup>[18]</sup>; digital annealers implement simulated annealing in dedicated hardware. Table 1 lists representative thermodynamic, probabilistic, and Ising hardware and its status.

**Table 1.** Representative physics-based sampling and optimization hardware.

| SYSTEM                          | MECHANISM                            | TASK                    | STATUS                                    | REF .   |
|---------------------------------|--------------------------------------|-------------------------|-------------------------------------------|---------|
| Thermodynamic sampling unit     | programmable noise in silicon        | generative sampling     | prototype; $\approx 10^4 \times$ reported | [11]    |
| Thermodynamic chip (datacenter) | physics-based linear algebra         | sampling / HPC          | early tape-out                            | [11]    |
| p-bit array (sMTJ)              | stochastic magnetic tunnel junctions | optimization, inference | silicon; scaled to $10^6$ p-bits          | [12,14] |
| Coherent Ising machine          | optical parametric oscillators       | optimization            | laboratory demonstrations                 | [17]    |
| Quantum / digital annealer      | adiabatic / simulated annealing      | optimization            | commercial (annealing)                    | [18]    |

## 8. The thermodynamic limit

The field's lower bound is set by thermodynamics. Landauer's principle states that erasing one bit of information dissipates at least

$$E_{\min} = kT \cdot \ln 2 \approx 2.9 \times 10^{-21} \text{ J} \quad (T = 300 \text{ K})$$

(5)

while logically reversible operations, which erase nothing, carry no such bound<sup>[8,9,10]</sup>. Reversible and adiabatic computing pursue this headroom by recovering the energy of a computation. Thermodynamic and probabilistic machines are of interest partly because they treat the thermal environment as a resource rather than a loss, and so operate in the regime where equation (5) is the relevant limit rather than the switching energy of a transistor.

## 9. Status

The claims here are among the largest in computing, and a review should separate what is established from what is projected. The energy-based-model theory is settled and decades old. The sampling algorithms are mature. p-bit devices and small Ising machines are demonstrated in silicon and have run real tasks. What remains early is scaled, general-purpose thermodynamic hardware. The headline efficiency figures derive substantially from prototypes, analysis, and simulation, and the first datacenter-scale tape-outs are recent. The honest reading matches the companion report on ternary arithmetic: the physics and the algorithms are real, the small-scale demonstrations are real, and the volume silicon is a work in progress. Stating this identifies the open problem, which is scaling and integration.

## 10. Discussion and conclusion

A policy has a deterministic path, the forward evaluation of a learned function, and a stochastic path, the sampling treated here. A companion report argues that multiply-free arithmetic serves the deterministic path near its energy floor; this report argues that physics-based sampling serves the stochastic path near its floor. The two are complementary. The deterministic path performs no sampling and the stochastic path performs no matrix multiplication, so a system that runs each on the substrate suited to it need not pay for the operation it does not use. On that reading an energy-proportional edge is plausibly a heterogeneous device rather than a single one. The position rests on open prior art: the Ising model from 1925<sup>[1]</sup>, Boltzmann machines from 1985<sup>[3]</sup>, and Landauer's bound from 1961<sup>[8]</sup>. Sampling is the operation that physics performs most naturally, by letting a stochastic system fall into a distribution rather than computing that distribution first. The reported efficiencies are large and the silicon is young. For Physical AI at the edge, physics-based sampling is the natural stochastic complement to multiply-free deterministic arithmetic.

## References

1. E. Ising. Beitrag zur Theorie des Ferromagnetismus. Zeitschrift für Physik, 1925.
2. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. PNAS 79, 1982.
3. D. H. Ackley, G. E. Hinton, T. J. Sejnowski. A learning algorithm for Boltzmann machines. Cognitive Science 9, 1985.
4. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller. Equation of state calculations by fast computing machines. J. Chem. Phys., 1953.
5. R. J. Glauber. Time-dependent statistics of the Ising model. J. Math. Phys., 1963.
6. S. Geman, D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE PAMI, 1984.
7. S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi. Optimization by simulated annealing. Science 220, 1983.
8. R. Landauer. Irreversibility and heat generation in the computing process. IBM J. Res. Dev., 1961.
9. C. H. Bennett. Logical reversibility of computation. IBM J. Res. Dev., 1973.
10. E. Fredkin, T. Toffoli. Conservative logic. Int. J. Theoretical Physics, 1982.
11. Extropic. Thermodynamic Computing: From Zero to One; TSU 101; and the *thrml* library, 2023–26; G. Verdon, T. Coles, et al.

12. K. Y. Camsari, S. Datta, et al. A full-stack view of probabilistic computing with p-bits: devices, architectures and algorithms. arXiv:2302.06457, 2023.
13. W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno, S. Datta. Integer factorization using stochastic magnetic tunnel junctions. Nature 573, 2019.
14. Programmable Probabilistic Computer with 1,000,000 p-bits. arXiv:2606.25313, 2026.
15. An efficient probabilistic hardware architecture for diffusion-like models. arXiv:2510.23972, 2025.
16. From Independent to Correlated Diffusion: Generalized Generative Modeling with Probabilistic Computers. arXiv:2603.27996, 2026.
17. T. Inagaki et al. A coherent Ising machine for 2000-node optimization problems. Science 354, 2016.
18. T. Kadowaki, H. Nishimori. Quantum annealing in the transverse Ising model. Phys. Rev. E, 1998.
19. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. ICML, 2015.
20. J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. NeurIPS, 2020.
21. M. Horowitz. Computing's energy problem (and what we can do about it). IEEE ISSCC, 2014.
22. Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang. A tutorial on energy-based learning. Predicting Structured Data, MIT Press, 2006.

**AI-use disclosure.** Preparation of this report used a large language model (Claude, Anthropic) for drafting and editing text, organizing the reviewed literature, and preparing the figures and the interactive companion. Cited references were checked to resolve to their sources. The author reviewed the content and is solely responsible for it. Consistent with ICMJE, COPE, and IEEE guidance, the model is a tool and is not credited as an author.