

THE OPEN FIELD OF COMPUTING

Post-von Neumann and Energy-Efficient Computing Paradigms for Physical AI at the Edge: A Survey

From the orbital gigawatt datacenter to the sub-microwatt microcontroller, with attention to how old each method is.

David Jean Charlot, PhD

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

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org · The Charlot Lab: labs.physicalai-bmi.org/charlot

Interactive companion: physicalai-bmi.org/research/charlot-lab#topic-compute

ABSTRACT. The dominant cost of contemporary artificial intelligence is energy, and most of it is spent moving data across the von Neumann boundary and performing floating-point multiplication. This report surveys the computing paradigms that reduce or remove those costs. It organizes about forty methods into nine families along two axes, paradigm and deployment scale, spanning energy-harvesting microcontrollers to orbital datacenters. One observation runs through the survey: the roots of nearly every method are decades old, and are matters of published prior art rather than proprietary advantage. Balanced ternary hardware dates to 1958, the memristor to 1971, the Ising model to 1925, reversible computing to 1961. Section 2 states the review method. Section 3 gives the taxonomy and a timeline of these root dates. Sections 4 through 12 survey the families. Section 13 argues that the near-term advance in edge computing is not a single new paradigm but the convergence of existing open ones, and in particular the pairing of a multiply-free deterministic substrate with a physics-based stochastic one. The report reports no new experimental measurements.

1. Introduction

The von Neumann architecture fetches instructions and data from a memory that is separate from the processor, computes, and writes back. Its defining feature is also its defining inefficiency: memory and computation are physically and energetically separate. For workloads dominated by dense linear algebra, as machine learning is, two costs dominate. The first is the multiply-accumulate operations, in which multiplication is more expensive than addition^[A]. The second is the movement of operands between memory and the arithmetic units, which at scale exceeds the arithmetic energy by an order of magnitude or more.

For Physical AI, meaning embodied systems that sense, model, and act within a body's power budget, these costs are a hard constraint rather than an abstraction. A model that runs comfortably in a datacenter may

be inadmissible on a battery. The design metric becomes energy per decision, which motivates a broad reconsideration of how computation is physically realized. This report collects the principal alternatives. Its aim is breadth and orientation rather than depth in any single method; for each it gives the operating principle, the source of the energy or mathematical advantage, the maturity, and the historical root.

2. Scope and method

This is a review. It surveys published methods and hardware for computing beyond the standard von Neumann model, drawing on peer-reviewed papers, preprints, and primary hardware disclosures across a period from 1925 to 2026. For each claim it favors a primary source, and for each method it cites the earliest source that established it, so the record of prior art is visible. Coverage is deliberately broad and therefore shallow per method; readers are directed to the cited work for depth. The report states a research position in Section 13 and reports no original experiments. Its principal limitation is that the field moves quickly and several of the most recent hardware results are simulations or prototypes rather than production parts; the survey marks maturity where it is relevant.

3. A two-axis taxonomy, and the age of the field

The survey organizes methods along two axes. The first is the paradigm family, meaning the physical or mathematical mechanism that performs the computation. The second is deployment scale, a spectrum of power running from sub-microwatt energy-harvesting microcontrollers, through edge accelerators and datacenter systems, to orbital datacenters. Table 1 lists the nine families. Figure 1 places the root year of a representative method from each on a single timeline, which makes the central observation of this report concrete: the ideas are old.

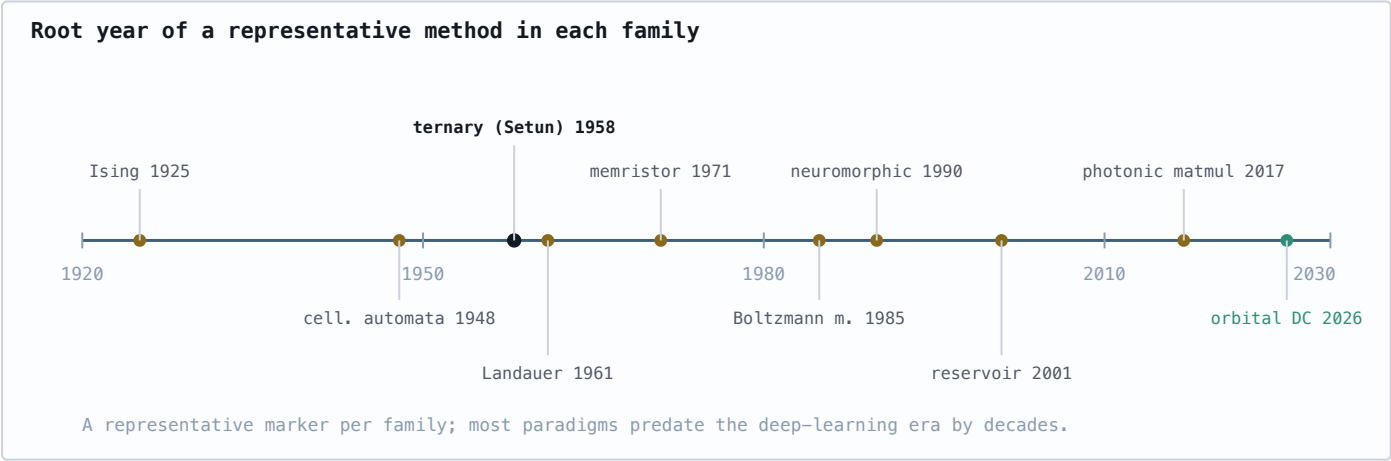


Figure 1. Root years of representative methods across the nine families. The span from the Ising model (1925) to orbital datacenters (2026) shows that the paradigms surveyed here are, with few exceptions, long-published prior art. An interactive version that maps every method by family and deployment scale is at physicalai-bmi.org/research/charlot-lab.

Table 1. The nine paradigm families and representative methods.

FAMILY	REPRESENTATIVE METHODS	EARLIEST ROOT
Number & arithmetic	ternary, logarithmic, posit, microscaling FP, residue, stochastic, approximate	1925 / 1958
In-memory & dataflow	SRAM compute-in-memory, memristor crossbar, PCM, MRAM, FeFET, systolic arrays, PIM	1969

Neuromorphic	spiking networks, analog neuromorphic, memristive synapses, event cameras, many-core fabrics	1952 / 1989
Thermodynamic & probabilistic	thermodynamic sampling units, p-bits, Ising machines, reversible/adiabatic, annealers	1925 / 1961
Optical & photonic	photonic matrix multiply, photonic reservoir, optical Ising, photonic hyperdimensional	1960s
Quantum	gate-model, annealing, measurement-based, analog simulation	1982
Exotic substrate	superconducting (SFQ), spintronic/magnonic, molecular/DNA, biological/organoid, reaction-diffusion	1962 / 1994
Representation & dynamics	hyperdimensional/VSA, reservoir computing, cellular automata, physical reservoir, continuous-time	1931 / 1948
Scale & deployment	orbital datacenter, energy-harvesting MCU, sub-threshold, edge accelerators, wafer-scale	1968 / 1972

4. Number systems and arithmetic

The cheapest way to make multiplication less expensive is to change the numbers. Ternary weights, restricted to $\{-1, 0, +1\}$, turn a product into an add, a subtract, or a skip and remove the multiplier. Realized in the Setun computer in 1958^[1], the method returned with BitNet b1.58^[2], which trained language models at 1.58 bits per weight to full-precision quality, and with BitVLA for embodied policies^[3]. Logarithmic number systems make a multiply an addition of exponents^[5], an idea recent work has applied to approximate floating-point multiplication by integer addition^[6]. Posits allocate precision where it is used^[7], and microscaling block formats share one exponent across a block of low-bit values^[8]. Older still, the residue number system gives carry-free parallel arithmetic^[9], and stochastic computing encodes numbers as bitstream probabilities so a multiplier becomes a single gate^[10].

5. In-memory and dataflow architectures

If the bottleneck is moving data to the arithmetic, the response is to compute where the data sits. Compute-in-memory performs the multiply-accumulate inside the memory array, an idea introduced as logic-in-memory in 1969^[11]. Analog crossbars of memristors^[12,13], phase-change memory^[14], and ferroelectric devices compute a matrix-vector product in a single physical step, with the weight stored as a device conductance. Where computation cannot be pushed fully into memory, systolic and dataflow arrays stream operands through a grid of processing elements to minimize fetches^[15], the organizing principle of modern tensor accelerators.

6. Neuromorphic systems

Neuromorphic computing takes the brain's organization as a template: co-located memory and computation, parallelism, and communication by sparse events. Analog neuromorphic circuits run transistors in the subthreshold regime as silicon neurons^[16]. Spiking networks compute only when a neuron fires, so energy tracks activity rather than model size^[17]; large digital realizations include Loihi^[18] and TrueNorth^[19]. The same principle at the sensor yields the event camera, whose pixels report only change at microwatts.

7. Thermodynamic and probabilistic computing

A generative step is a draw from a distribution. Conventional hardware computes the distribution with matrix multiplication and then samples; thermodynamic hardware encodes the distribution in a physical system and lets thermal fluctuation produce the sample. The lineage runs from the Ising model^[20] and Hopfield networks^[21] through Boltzmann machines to thermodynamic sampling units^[26] and probabilistic p-bits built from stochastic magnetic tunnel junctions^[27], now scaled to a million bits^[28]. Ising machines and annealers^[22] solve optimization by relaxing a physical system toward its minimum. At the limit sit reversible and adiabatic computing, which approach the Landauer bound of $kT \ln 2$ per erased bit^[23,24,25]. This family is the subject of a companion report.

8. Optical and photonic computing

Light performs linear operations at low energy: a matrix–vector product can be realized by propagation through a mesh of interferometers, with most energy spent at the electro-optic interfaces^[30]. Photonic reservoirs exploit optical dynamics for temporal processing at high throughput^[31], and optical Ising machines settle networks of parametric oscillators into optimization solutions. The field's roots in Fourier optics are decades deep; its present constraint is integration and the cost of converting between optical and electronic domains.

9. Quantum computing

For completeness the survey includes quantum computing, which uses superposition and entanglement to obtain, for specific problems, advantages unavailable classically^[39]. Gate-model, measurement-based, and analog-simulation approaches target general and special-purpose computation, and quantum annealing addresses optimization by adiabatic evolution^[40]. Quantum systems are datacenter-scale and cryogenic; for the embodied edge they are complementary rather than substitutive, and appear here as one boundary of the field.

10. Exotic substrates

Computation does not require CMOS, and several communities pursue other substrates. Superconducting single-flux-quantum logic switches at attojoule energies and tens of gigahertz at the cost of cryogenic operation^[38]. Spintronic and magnonic devices compute with spin rather than charge. Molecular and DNA computing use chemical parallelism^[36], reaction–diffusion systems compute with propagating waves, and biological and organoid computing use living neural tissue^[37]. These remain frontier research, valuable less as near-term targets than as evidence that the space of physical computation is far larger than the transistor.

11. Alternative representations and dynamics

A parallel line changes the representation rather than the substrate. Hyperdimensional computing operates on very high-dimensional vectors that distribute information holographically, which gives robustness to noise and cheap binding operations^[32]. Reservoir computing couples a fixed random dynamical system to a single trained readout, moving nearly all cost out of training^[33]; when the reservoir is a physical medium, the nonlinearity is free. Cellular automata obtain global, Turing-complete computation from local rules^[34],

and continuous-time analog machines solve differential equations by letting a physical system evolve, a tradition older than the digital computer^[35].

12. Scale and deployment

Independent of paradigm is scale, and the field's two extremes both move quickly. At the largest scale, orbital datacenters, proposed in principle with space solar power in 1968^[42], reached operation when a model was trained aboard a spacecraft in 2025^[43], trading launch cost against abundant solar power and radiative cooling. At the smallest, energy-harvesting microcontrollers compute intermittently with no battery, drawing on decades of sub-threshold design^[41]. Between them sit edge accelerators and wafer-scale integration, the latter removing off-chip communication by making the chip the size of the wafer. Physical AI runs across this whole spectrum, and the appropriate paradigm depends on where in it a system must operate.

13. Discussion: convergence

Two observations shape the conclusion. First, no single paradigm dominates across workloads and scales; each buys efficiency by specializing. Second, two of them are complementary in a specific way. An embodied policy has a deterministic path, the forward evaluation of a learned function, and a stochastic path, the sampling of actions, futures, and beliefs. Multiply-free arithmetic, of which ternary is the sharpest case, serves the deterministic path near its energy floor. Physics-based sampling, of which thermodynamic hardware is the sharpest case, serves the stochastic path near its floor. Neither pays for the other's dominant cost.

The report therefore advances a modest thesis. The near-term advance in edge computation is unlikely to come from a newly discovered paradigm, since the record shows the paradigms are already known, and more likely to come from the co-integration of open, complementary ones on a single heterogeneous device. This is a research programme, not a product claim, and it rests on the open prior art the survey has mapped. The future of computing, on this reading, is a commons, and the work is to converge it.

14. Conclusion

This report surveyed the principal paradigms of post-von Neumann and energy-efficient computing across nine families and the full deployment spectrum, and emphasized the depth and openness of the underlying prior art. The energy cost of intelligence is the organizing constraint of embodied computation, and the field's response is neither singular nor new but plural and old. For Physical AI at the edge, the most promising direction is the disciplined convergence of these open methods, beginning with the pairing of deterministic ternary arithmetic and stochastic thermodynamic sampling.

References

1. N. P. Brusentsov et al. Setun: development and operation of a ternary computer. Moscow State University, 1958–65.
2. S. Ma, H. Wang, et al. The Era of 1-bit LLMs (BitNet b1.58). arXiv:2402.17764, 2024; BitNet b1.58 2B4T Technical Report, arXiv:2504.12285, 2025.
3. H. Wang, S. Xiong, et al. BitVLA: 1-bit Vision-Language-Action Models for Robotics Manipulation. arXiv:2506.07530, 2025.
4. Sparse-BitNet. arXiv:2603.05168, 2026.
5. H. Wang, S. Ma, et al. BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv:2310.11453, 2023.
6. J. N. Mitchell. Computer multiplication and division using binary logarithms. IRE Trans. Electronic Computers, 1962.

7. H. Luo et al. Addition is All You Need for Energy-efficient Language Models (L-Mul). arXiv:2410.00907, 2024.
8. J. L. Gustafson, I. Yonemoto. Beating Floating Point at its Own Game: Posit Arithmetic. Supercomputing Frontiers and Innovations, 2017.
9. B. D. Rouhani et al. Microscaling Data Formats for Deep Learning. arXiv:2310.10537, 2023 (Open Compute Project MX specification).
10. H. L. Garner. The residue number system. IRE Trans. Electronic Computers, 1959.
11. B. R. Gaines. Stochastic computing systems. Advances in Information Systems Science, 1969.
12. W. H. Kautz. Cellular logic-in-memory arrays. IEEE Trans. Computers, 1969.
13. L. O. Chua. Memristor—the missing circuit element. IEEE Trans. Circuit Theory, 1971.
14. D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams. The missing memristor found. Nature 453, 2008.
15. S. R. Ovshinsky. Reversible electrical switching phenomena in disordered structures. Phys. Rev. Lett., 1968.
16. H. T. Kung, C. E. Leiserson. Systolic arrays (for VLSI). Sparse Matrix Proc., 1978.
17. C. Mead. Neuromorphic electronic systems. Proc. IEEE 78(10), 1990.
18. W. Maass. Networks of spiking neurons: the third generation of neural network models. Neural Networks, 1997.
19. M. Davies et al. Loihi: a neuromorphic manycore processor with on-chip learning. IEEE Micro, 2018.
20. P. A. Merolla et al. A million spiking-neuron integrated circuit (TrueNorth). Science 345, 2014.
21. E. Ising. Beitrag zur Theorie des Ferromagnetismus. Zeitschrift für Physik, 1925.
22. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. PNAS 79, 1982.
23. T. Inagaki et al. A coherent Ising machine for 2000-node optimization problems. Science 354, 2016.
24. R. Landauer. Irreversibility and heat generation in the computing process. IBM J. Res. Dev., 1961.
25. C. H. Bennett. Logical reversibility of computation. IBM J. Res. Dev., 1973.
26. E. Fredkin, T. Toffoli. Conservative logic. Int. J. Theoretical Physics, 1982.
27. Extropic. Thermodynamic Computing: From Zero to One and related reports, 2023–26; G. Verdon, T. Coles, et al.
28. K. Y. Camsari, S. Datta, et al. A full-stack view of probabilistic computing with p-bits. arXiv:2302.06457, 2023.
29. Programmable Probabilistic Computer with 1,000,000 p-bits. arXiv:2606.25313, 2026.
30. Y. Shen et al. Deep learning with coherent nanophotonic circuits. Nature Photonics 11, 2017.
31. Scalable photonic reservoir computing for parallel machine learning. Nature Communications, 2025.
32. P. Kanerva. Sparse Distributed Memory, MIT Press, 1988; Hyperdimensional computing, Cognitive Computation, 2009.
33. H. Jaeger. The echo-state approach, 2001; W. Maass, T. Natschläger, H. Markram. Real-time computing without stable states (LSM). Neural Computation, 2002.
34. J. von Neumann, A. W. Burks (ed.). Theory of Self-Reproducing Automata, 1966; M. Gardner. Conway's Game of Life, Sci. Am., 1970.
35. V. Bush. The differential analyzer. J. Franklin Institute, 1931.
36. L. M. Adleman. Molecular computation of solutions to combinatorial problems. Science 266, 1994.
37. B. J. Kagan et al. In vitro neurons learn and exhibit sentience when embodied in a simulated game-world (DishBrain). Neuron 110, 2022.
38. K. K. Likharev, V. K. Semenov. RSFQ logic/memory family. IEEE Trans. Applied Superconductivity, 1991.
39. R. P. Feynman. Simulating physics with computers. Int. J. Theoretical Physics, 1982; D. Deutsch, Proc. R. Soc. Lond. A, 1985.
40. T. Kadowaki, H. Nishimori. Quantum annealing in the transverse Ising model. Phys. Rev. E, 1998.
41. R. M. Swanson, J. D. Meindl. Ion-implanted complementary MOS transistors in low-voltage circuits. IEEE J. Solid-State Circuits, 1972.
42. P. E. Glaser. Power from the sun: its future. Science 162, 1968.
43. Starcloud / NVIDIA. First AI model trained in orbit, Dec. 2025; Google. Project Suncatcher, 2025.
44. M. Horowitz. Computing's energy problem (and what we can do about it). IEEE ISSCC, 2014.

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.