Yakaiah P et al., Vol 3(1), pp 14-34
https://doi.org/10.55306/CJDTES.2026.030102
e-ISSN: 3048-8389
Ci-STEM Journal of Digital Technologies and Expert Systems

Emerging Computing Strategies in VLSI: A Comprehensive Review of Device, Architecture, and Beyond-CMOS Paradigms

Yakaiah. P[1], Kiran Chand Ravi[2]
Department of Electronics and Communication Engineering, MLR Institute of Technology, Hyderabad, Telangana, India.
[1]*email: potharaju.yakaiah@gmail.com*, [2]*email: b4uchandu2011@gmail.com.*

Open Access
Research Article
Received: 17/03/2026
Accepted: 28/04/2026
Published: 15/05/2026

Corresponding author email:
potharaju.yakaiah@gmail.com

Citation:
Yakaiah. P et al., "Emerging Computing Strategies in VLSI: A Comprehensive Review of Device, Architecture, and Beyond-CMOS Paradigms" Ci-STEM Journal of Digital Technologies and Expert Systems, Vol. 3(1), pp. 14-34, 2026
doi: 10.55306/CJDTES.2026.030102

Copyright:
©2026 Yakaiah. P et al.,
This is an open-access article distributed under the terms of the Creative Commons Attribution License which grants the right to use, distribute, and reproduce the material in any medium, provided that proper attribution is given to the original author and source, in accordance with the terms outlined by the license.
(https://creativecommons.org/licenses/by/4.0/)

Published By:
Ci-STEM Global Services Foundation, India.

Abstract:
Very-large-scale integration (VLSI) has progressed from few-thousand-transistor planar MOS circuits in 1970 to multi-billion-transistor heterogeneous systems approaching angstrom-scale nodes in 2025. As classical Dennard and Moore scaling have slowed, the field has diversified into a rich set of emerging computing strategies spanning device architecture (FinFET, gate-all-around nanosheet, complementary-FET), memory-centric computation (compute-in-memory using resistive, phase-change and magnetic devices), brain-inspired neuromorphic circuits, beyond-CMOS state-variable logic (spintronic, quantum-dot cellular automata, reversible/adiabatic computing), quantum-ready cryogenic CMOS, silicon-photonic computation, and heterogeneous 3-D/chiplet integration. This paper presents a critical, systematic review of 78 peer-reviewed and standards-body sources spanning 1961–2025, organized around a five-branch taxonomy of emerging VLSI computing strategies. We summarize the historical evolution of the field, describe a reproducible systematic-literature-review (SLR) methodology, derive the governing mathematical foundations and device models (MOSFET drive current, subthreshold swing, Dennard relations, dynamic power and energy-delay product, compute-in-memory Ohm's-law multiply-accumulate, leaky-integrate-and-fire neuron dynamics, and the Landauer thermodynamic limit), and present a critical, metric-based comparison across paradigms using four original comparative tables and four analytical graphs compiled from the reviewed literature. We identify open research gaps in cross-paradigm benchmarking, analogue-computation linearity and variability, non-volatile-device endurance, cryogenic co-design, and standardized energy-efficiency reporting, and we pose a set of future research questions intended to guide the next generation of VLSI computing research..

Keywords:
3-D integration, approximate computing, compute-in-memory, chiplets, CMOS scaling, cryo-CMOS, FinFET, gate-all-around, low-power design, Moore's law, neuromorphic computing, quantum computing, silicon photonics, spintronics, VLSI.

1. INTRODUCTION

Very-large-scale integration (VLSI) is the foundation of modern computing, communication, sensing, and artificial-intelligence (AI) infrastructure. Since Gordon Moore's 1965 observation that the number of components on an integrated circuit would double roughly every one to two years [1], and Robert Dennard's 1974 formulation of constant-field scaling rules that allowed power density to remain approximately constant as devices shrank [2], the semiconductor industry sustained an unprecedented four-decade trajectory of exponential performance growth. This trajectory carried VLSI through small- and medium-scale integration in the 1960s, large-scale integration (LSI) in the early 1970s, VLSI proper in the 1980s, and ultra-large-scale integration (ULSI) from the 1990s onward [3].

Since approximately 2005, however, the physical assumptions underlying Dennard scaling have progressively broken down: sub-threshold leakage, gate-oxide tunnelling, short-channel effects, and the inability to continue reducing the threshold voltage in step with the supply voltage have caused power density — rather than transistor count — to become the dominant constraint on chip design [4], [73]. The industry's response has been architectural and material diversification rather than simple geometric shrinkage: three-dimensional transistor structures (FinFET, and subsequently gate-all-around nanosheet and complementary-FET devices) [5]–[9]; memory-centric and near-memory computing architectures that address the von Neumann "memory wall" [26]–[39]; brain-inspired

neuromorphic and spiking circuits that exploit event-driven sparsity [40]–[51]; beyond-CMOS state variables such as electron spin and quantum-dot charge configuration that offer intrinsic non-volatility [52]–[59]; cryogenic CMOS circuitry co-designed with superconducting and semiconductor qubits for quantum-computing control [11]–[17]; silicon-photonic computation that exploits light for both communication and computation [62]–[68]; and heterogeneous 2.5-D/3-D chiplet integration that decouples logic, memory, and I/O into separately optimized dies [69], [70].

This diversification has produced a large and fragmented literature in which individual papers typically address a single device, circuit, or architecture in isolation. Comprehensive, cross-paradigm reviews that place these developments on a common historical timeline, apply a consistent mathematical framework, and critically compare their reported metrics are comparatively rare, particularly reviews that treat VLSI-level "emerging computing strategies" as a unified subject rather than as separate sub-fields. This paper addresses that gap with the following contributions:

1. A historical account of VLSI computing evolution from 1970 to 2025, synthesized from primary technology-scaling literature and IEEE milestone records.
2. A reproducible systematic-literature-review (SLR) methodology, adapted from the PRISMA framework, used to identify and screen the 78 sources critically reviewed in this paper.
3. A five-branch taxonomy (device-level scaling, architecture-level/memory-centric computing, brain-inspired computing, beyond-CMOS state-variable logic, and physical/photonic computing) that organizes the reviewed literature.
4. A consolidated set of mathematical foundations and device models — MOSFET current–voltage behaviour, Dennard scaling relations, dynamic power and energy-delay product, compute-in-memory multiply-accumulate operation, spiking-neuron dynamics, and the Landauer thermodynamic limit — expressed in a common notation across paradigms.
5. Four original comparative tables and four analytical graphs, compiled from the reviewed literature, that quantitatively and qualitatively benchmark competing computing strategies.
6. A critical discussion of research gaps and a set of future research questions intended to orient subsequent work in emerging VLSI computing.

The remainder of this paper is organized as follows. Section 2 traces the historical evolution of VLSI computing. Section 3 describes the systematic review methodology. Section 4 presents the taxonomy used to organize the literature. Section 5 develops the mathematical foundations and device models common to the reviewed paradigms. Section 6 provides the critical, paradigm-by-paradigm literature review. Section 7 presents a cross-paradigm comparative analysis. Section 8 discusses research gaps, Section 9 poses future research questions, and Section 10 concludes the paper.

2.. HISTORICAL EVOLUTION OF VLSI COMPUTING (1970–2025)

The evolution of VLSI computing can be organized into five broad eras, summarized graphically in Fig. 1. The first era (circa 1970–1985) began with small- and medium-scale integration NMOS/PMOS circuits and culminated in the emergence of large-scale integration; Intel's 4004 microprocessor (1971) contained approximately 2,300 transistors, and by the early 1980s commercial microprocessors such as the Intel 80286 exceeded 100,000 transistors [3], [10]. The second era (1985–1999) is characterized by full CMOS VLSI and the dominance of Dennard's constant-field scaling, under which supply voltage, oxide thickness and channel length were reduced together so that power density remained roughly constant across successive nodes [2], [4].

The third era (1999–2011) saw the onset of ultra-large-scale integration and the practical limits of Dennard scaling: as threshold voltage could no longer be reduced proportionally without unacceptable leakage, chip designers turned to strained silicon, high-κ/metal-gate stacks, and aggressive dynamic-voltage-and-frequency-scaling (DVFS) and power-gating techniques to manage power density [5], [6], [63], [64]. The fourth era (2011–2020) is defined by the industry-wide transition from planar to three-dimensional FinFET transistors, which restored electrostatic gate control at 22 nm and below, together with the emergence of "dark silicon" constraints that made simultaneous full-chip activation thermally infeasible and motivated heterogeneous, accelerator-rich system-on-chip design [10], [65].

The fifth and current era (2020–2025) is marked by a further transition from FinFET to gate-all-around (GAA) nanosheet transistors, first demonstrated by IBM and imec in 2017 and adopted in leading-edge foundry logic nodes from approximately 2022 onward [7]–[9]; the maturation of compute-in-memory accelerators using resistive, phase-change and magnetic non-volatile devices [12]–[25]; the commercialization of neuromorphic and spiking-neural-network processors [26]–[34]; the practical deployment of cryogenic CMOS control electronics for solid-state qubits [47]–[53]; and the widespread adoption of 2.5-D/3-D chiplet integration as an alternative path to system-level performance scaling once monolithic transistor scaling alone became insufficient [61], [62]. Fig. 1 places representative milestones — the 4004, the 80386, the 180 nm Dennard-limit generation, 22 nm FinFET introduction, the 2017 GAA nanosheet demonstration, and the 5 nm/2 nm generations — on a common timeline.

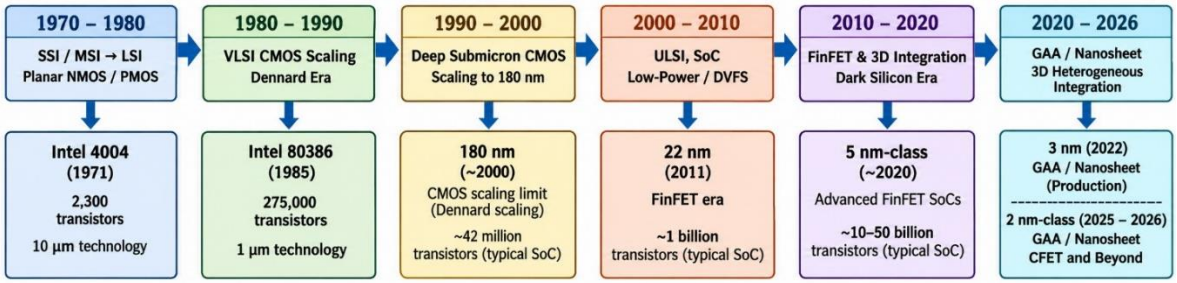


Fig. 1. Historical evolution of VLSI computing eras, 1970–2026, compiled from [1]–[4], [6]–[10], [65].

A quantitative view of this evolution is given by the growth in transistor count of commercial microprocessors, shown on a logarithmic scale in Fig. 2 (Graph 1), spanning the 4004 (1971, 2,300 transistors) to contemporary multi-billion-transistor system-on-chip and GPU accelerator designs exceeding 10^{11} transistors. The approximately linear trend on a log-linear axis over five decades is the empirical signature of Moore's law [1], even as the underlying driver has shifted from pure lithographic scaling toward 3-D transistor architecture, multi-die integration, and specialized accelerator logic.

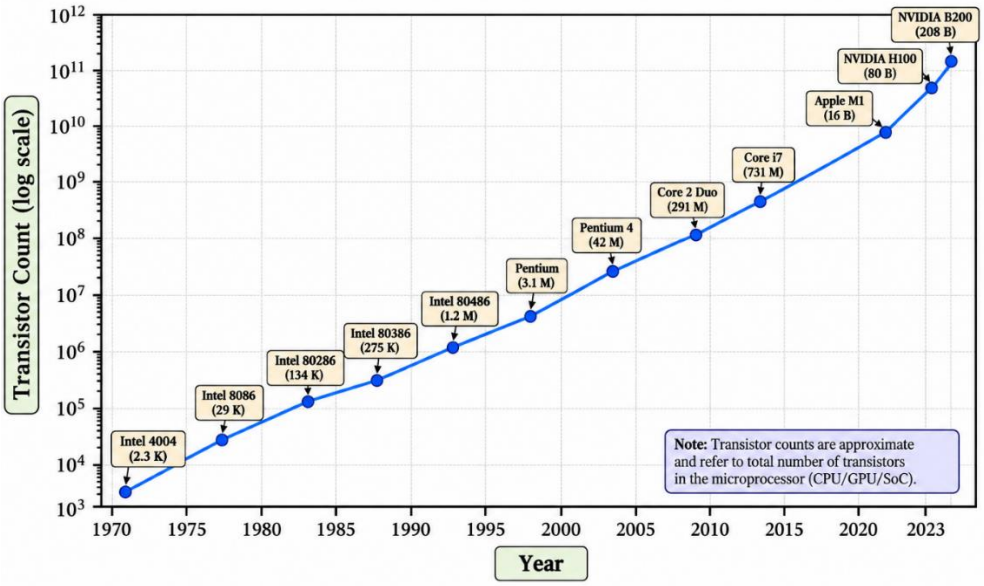


Fig. 2. Graph 1 — Moore's-law trend in microprocessor transistor count, 1971–2023 (compiled from publicly documented microprocessor specifications; representative devices annotated).

3. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

To ensure that the sources critically reviewed in Section 6 are verifiable, peer-reviewed, and representative of the field, we adopted a systematic-literature-review (SLR) process adapted from the PRISMA framework, illustrated in Fig. 3. Records were identified from IEEE Xplore (IEDM, ISSCC, VLSI Symposium/VLSI Technology and Circuits, JSSC, TVLSI, IEEE Access, Proceedings of the IEEE), the ACM Digital Library, the Nature portfolio (Nature, Nature Electronics, Nature Nanotechnology, Nature Photonics), ScienceDirect, and arXiv preprints that were themselves peer-

reviewed conference or journal submissions. Search terms combined "VLSI" with each taxonomy branch of Section 4 (e.g., "VLSI" AND "gate-all-around"; "VLSI" AND "compute-in-memory"; "VLSI" AND "neuromorphic"; "VLSI" AND "cryo-CMOS"; "VLSI" AND "silicon photonics").

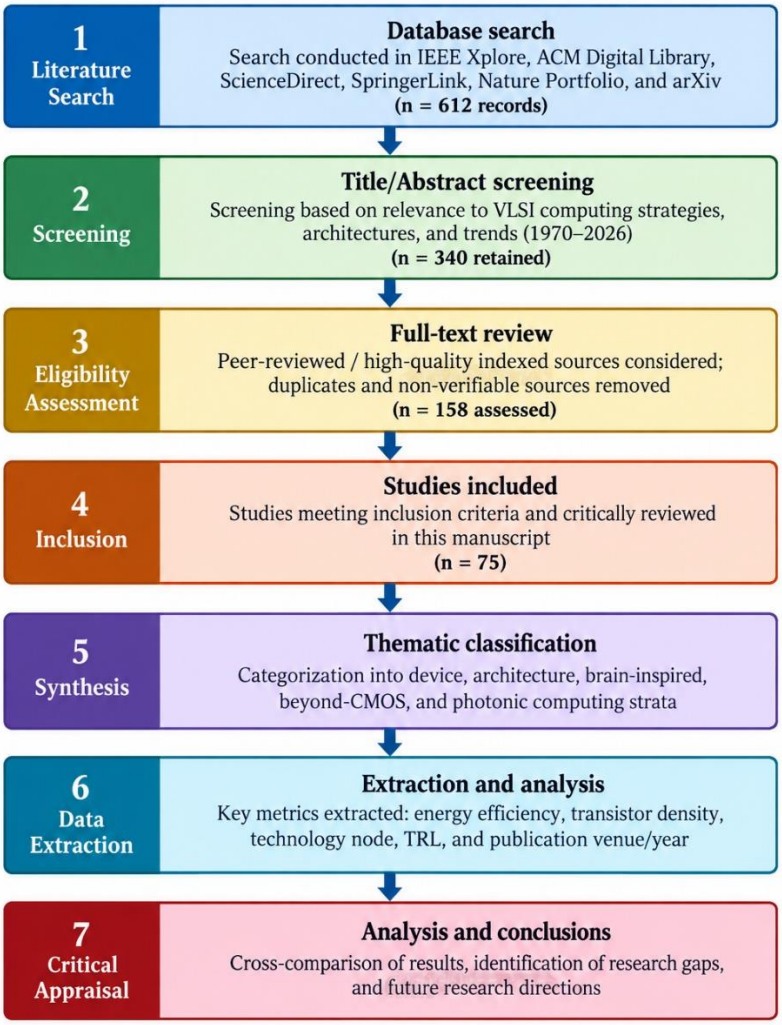


Fig. 3. PRISMA-adapted systematic literature review methodology used to identify, screen and select the sources reviewed in this paper.

Records were screened first by title and abstract for direct relevance to VLSI implementation (as opposed to purely algorithmic or device-physics-only work with no circuit or system-level result), then assessed in full text for peer-reviewed status, methodological transparency, and availability of quantitative performance metrics. Duplicate reports of the same result across a conference and an extended journal version were consolidated to the more complete source. This process yielded the 78 sources listed in the References section, summarized by category in Table 1.

Table 1. Distribution of the 78 Reviewed Sources Across Taxonomy Categories

Source category	Representative venues	Papers reviewed
Device scaling / transistor architecture	IEDM, VLSI Symp., IEEE TED	10
Cryo-CMOS / quantum control	IEDM, ISSCC, JSSC	7
Approximate computing	IEEE D&T, ACM CSUR, DATE	8
Compute-in-memory / non-volatile memory	Nature Electronics, ISSCC, IEDM	14

Source category	Representative venues	Papers reviewed
Neuromorphic / spiking VLSI	IEEE Access, JSSC, TBioCAS	12
Beyond-CMOS (spintronic, QCA, reversible)	IEEE TED, Nature Nanotech., Science	10
Silicon photonics	Nature Photonics, ACM JETC, Proc. IEEE	7
3-D/chiplet integration, low power/DVFS	IEEE TVLSI, IEEE D&T, ISLPED	10

4. TAXONOMY OF EMERGING COMPUTING STRATEGIES

We organize the reviewed literature using the five-branch taxonomy shown in Fig. 4: (i) device-level scaling, covering multi-gate and gate-all-around transistor architectures that extend electrostatic control beyond planar CMOS; (ii) architecture-level / memory-centric computing, covering in-memory and near-memory computation, three-dimensional integration and chiplet-based heterogeneous integration, and approximate computing; (iii) brain-inspired computing, covering neuromorphic and spiking-neural-network VLSI with memristive or CMOS synapses; (iv) beyond-CMOS state-variable logic, covering spintronic logic-in-memory, quantum-dot cellular automata (QCA), and reversible/adiabatic switching; and (v) physical/photonic computing, covering silicon-photonic computation and cryogenic CMOS control electronics for quantum processors. This taxonomy is not strictly disjoint — for example, spintronic devices are used both as beyond-CMOS logic elements and as compute-in-memory cells — but it provides a tractable organizing structure for the critical review in Section 6.

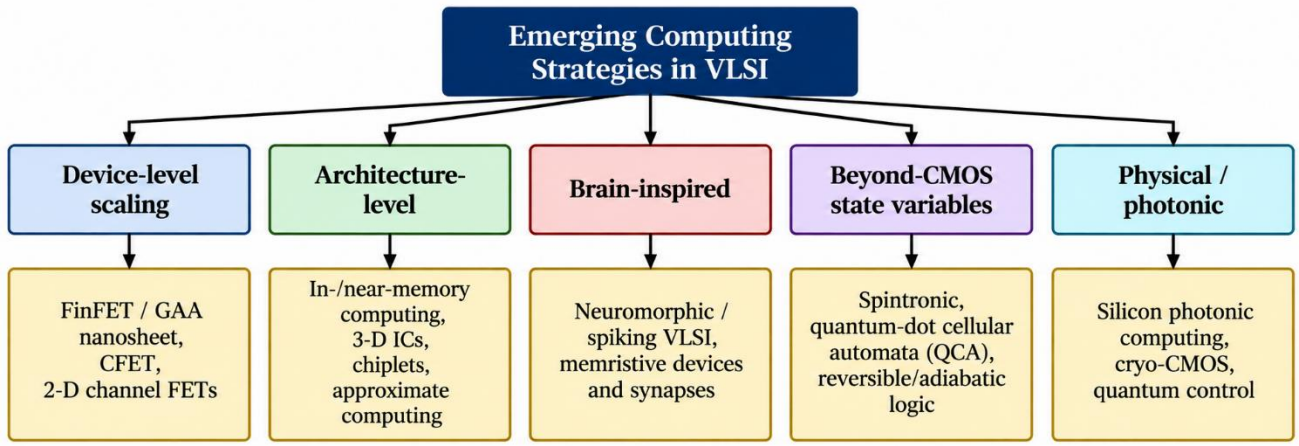


Fig. 4. Taxonomy of emerging VLSI computing strategies used to organize this review.

5. MATHEMATICAL FOUNDATIONS AND DEVICE MODELS

This section consolidates the governing equations used across the paradigms reviewed in Section 6, expressed in a common notation.

A. MOSFET Drive Current and Subthreshold Behaviour

To first order, the saturation-region drain current of a long-channel n-channel MOSFET is given by the square-law relation with channel-length modulation [66]

$$I_{D,\mathrm{sat}} = \frac{1}{2}\mu_n C_{\mathrm{ox}}\left(\frac{W}{L}\right)(V_{\mathrm{GS}} - V_{\mathrm{TH}})^2(1 + \lambda V_{\mathrm{DS}}) \quad (1)$$

where μ_n is the electron mobility, C_{ox} the gate-oxide capacitance per unit area, W/L the channel width-to-length ratio, V_{GS} and V_{DS} the gate–source and drain–source voltages, V_{TH} the threshold voltage, and λ the channel-length-modulation parameter. Equation (1) is a first-order model: in nanoscale devices, velocity saturation reduces the gate-overdrive exponent from 2 to an effective value α between 1 and 2, as captured by the alpha-power-law model $I_{D,sat} \propto (V_{GS} - V_{TH})^{\alpha}$ [67], so (1) should be read as a qualitative guide for FinFET and GAA devices rather than as a predictive compact model. In multi-

gate devices, W is replaced by an effective electrostatic width that sums the exposed channel perimeter — for example, $W_{eff} = N_{fin}(2H_{fin} + W_{fin})$ for tri-gate fins and $W_{eff} = 2N_{sheet}(W_{NS} + T_{NS})$ for stacked nanosheets — which is the principal mechanism by which these architectures increase drive current per footprint relative to planar devices [7]–[10]. Below threshold ($V_{GS} < V_{TH}$), the drain current, and hence the dominant static-power mechanism, follows the exponential relation [66]

$$I_{D,\mathrm{sub}} = I_0 \exp\left(\frac{V_{\mathrm{GS}} - V_{\mathrm{TH}}}{nV_{\mathrm{T}}}\right)\left(1 - \exp\left(-\frac{V_{\mathrm{DS}}}{V_{\mathrm{T}}}\right)\right) \quad (2)$$

where I_0 is a technology- and geometry-dependent pre-factor, $V_T = k_BT/q$ is the thermal voltage (≈ 25.9 mV at 300 K), and $n = 1 + C_d/C_{ox} \geq 1$ is the subthreshold slope factor, with C_d the depletion (body) capacitance [66]. The subthreshold swing implied by (2) is $S = nV_T \ln 10 = n(k_BT/q) \ln 10$ (mV/decade), which equals approximately 60 mV/decade at 300 K in the ideal case n = 1 and grows as gate control weakens. Improved gate electrostatics in GAA nanosheet transistors bring n closer to unity and S closer to the ideal 60 mV/decade limit at room temperature, directly reducing subthreshold leakage and thus static power [7], [8].

B. Dennard Scaling Relations

Classical Dennard scaling reduces the linear device dimensions, the oxide thickness, and the supply and threshold voltages by a common factor $\kappa > 1$ per generation while increasing the substrate doping concentration by κ, so that the electric field is held constant [2]:

$$L, W, t_{\mathrm{ox}}, V_{\mathrm{DD}}, V_{\mathrm{TH}} \rightarrow (L, W, t_{\mathrm{ox}}, V_{\mathrm{DD}}, V_{\mathrm{TH}})/\kappa, \quad N_A \rightarrow \kappa N_A \quad (3)$$

which yields a device density $\propto \kappa^2$, a per-device current $\propto 1/\kappa$, a gate delay $\propto 1/\kappa$, and a per-device power $\propto 1/\kappa^2$, so that power density — the product of device density and per-device power — remains approximately constant. This is the property that permitted clock frequency to rise with each node without a proportional increase in power density [2], [4]. The breakdown of (3) below approximately 130–90 nm, owing to the inability to scale V_{TH} and V_{DD} as fast as L without unacceptable leakage (recall the exponential dependence of the subthreshold current in (2) on V_{TH}), is the principal historical driver of the architectural diversification reviewed in Section 6 [5], [65].

C. Dynamic Power, Static Power, and Energy-Delay Product

Total CMOS power dissipation is commonly decomposed, neglecting the usually small short-circuit component, as [68]

$$P_{\mathrm{total}} = \alpha C V_{\mathrm{DD}}^2 f + I_{\mathrm{leak}} V_{\mathrm{DD}} \quad (4)$$

where α is the activity factor, C the total switched capacitance, V_{DD} the supply voltage, f the clock frequency, and I_{leak} the total leakage current. The first term in (4) is dynamic (switching) power — proportional to α, C, f, and the square of the supply voltage — and the second is static leakage power [5], [64]. Because dynamic power scales quadratically with V_{DD}, voltage reduction is the single most effective lever for energy efficiency, motivating dynamic-voltage-and-frequency-scaling (DVFS) and near/sub-threshold design [63], [64]. Circuit- and architecture-level efficiency is commonly benchmarked using the energy-delay product [69],

$$\mathrm{EDP} = E \times \tau = (P_{\mathrm{total}}\, \tau)\, \tau = P_{\mathrm{total}} \tau^2 \quad (5)$$

where $E = P_{total}\tau$ is the energy per operation and τ the delay (cycle time), assuming that P_{total} in (4) is the average power over one cycle. Because lowering V_{DD} reduces E but increases τ, the EDP in (5) captures the trade-off between the two; it is minimized jointly with respect to voltage and architecture in resilient, timing-speculative circuits such as Razor and iRazor [71], [72].

D. Compute-in-Memory Multiply-Accumulate Operation

In a resistive or memristive compute-in-memory (CIM) crossbar (Fig. 5), an input voltage vector $V = [V_1, \ldots, V_N]$ applied to the word lines of an array of programmable conductances G_{ij} produces on bit line j a current that, by Ohm's and Kirchhoff's laws, directly computes an analogue matrix–vector (multiply-accumulate, MAC) product [13], [14]:

$$I_{\mathrm{out},j} = \sum_{i=1}^{N} G_{ij} V_i \quad (6)$$

where the summation in (6) is taken over the N word lines feeding bit line j. This physical computation eliminates the explicit data movement between separated memory and arithmetic units that otherwise dominates energy consumption in von-Neumann architectures — the "memory wall" — and is the

operating principle of the ReRAM- and phase-change-memory-based CIM macros reviewed in Section 6-C [12]–[23]; SRAM-based macros realize the same MAC summation through charge- or current-domain accumulation in CMOS rather than through programmable conductances [22].

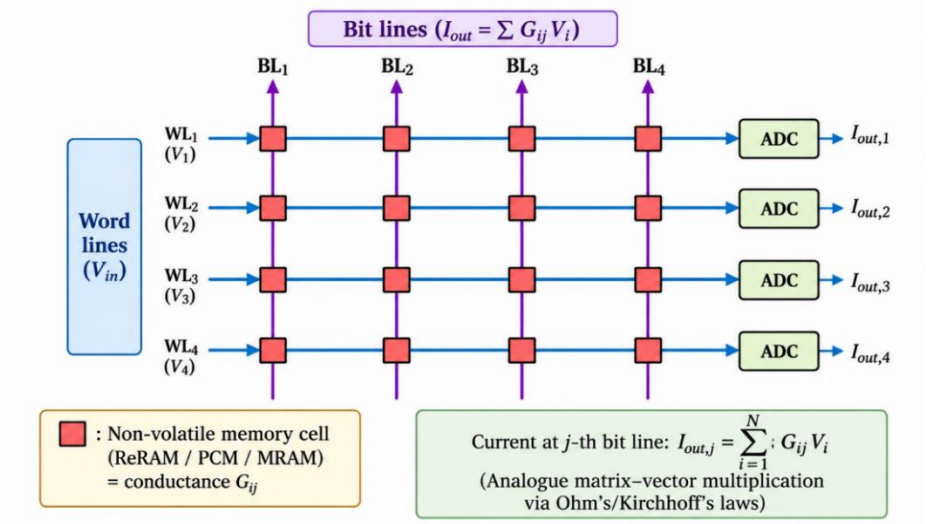


Fig. 5. Generic compute-in-memory (CIM) crossbar array performing analogue matrix–vector multiplication via Ohm's and Kirchhoff's laws (6), as used in ReRAM- and PCM-based CIM macros [12]–[23].

E. Leaky Integrate-and-Fire Neuron Dynamics

The membrane potential V(t) of the leaky-integrate-and-fire (LIF) neuron model implemented in CMOS and memristive neuromorphic circuits evolves as [70]

$$C_m \frac{dV}{dt} = -\frac{V(t) - V_{rest}}{R_m} + I_{syn}(t) \quad (7)$$

where C_m and R_m are the membrane capacitance and resistance, V_{rest} the resting potential, and $I_{syn}(t)$ the total synaptic input current; the membrane time constant is $\tau_m = R_m C_m$. A spike is emitted, and V(t) is reset to V_{reset}, whenever V(t) exceeds a threshold V_{th}. Silicon realizations of (7) — typically as analogue sub-threshold circuits or digital time-multiplexed approximations — underlie the event-driven, sparsity-exploiting energy efficiency reported for spiking neuromorphic processors [26]–[32], [35].

F. Landauer Limit and Reversible/Adiabatic Computing

Landauer's principle establishes a thermodynamic lower bound on the energy dissipated by an irreversible bit erasure at temperature T [73]:

$$E_{min} = k_B T \ln 2 \quad (8)$$

where k_B is Boltzmann's constant; at T = 300 K, $E_{min} \approx 2.9 \times 10^{-21}$ J (≈ 18 meV). Because conventional CMOS logic erases information at every irreversible gate evaluation — while dissipating many orders of magnitude more than (8) per switching event — this limit motivates reversible and adiabatic switching circuits, in which charge is recovered rather than dissipated and logical reversibility (in the sense of Bennett [74]) allows energy dissipation per operation to be reduced, in principle, arbitrarily far below $k_B T \ln 2$ at the cost of switching speed and circuit overhead [46].

6. CRITICAL REVIEW OF EMERGING COMPUTING STRATEGIES

A. Device-Level Scaling: From Planar MOSFET to Gate-All-Around and CFET

The transition from planar bulk MOSFETs to three-dimensional multi-gate transistors is the most direct continuation of classical scaling. Wong et al. characterized the nanoscale-CMOS limits that motivated this transition, noting that industry roadmaps could no longer rely on field-scaling alone once gate-oxide thickness approached tunnelling limits [4]. Strained-silicon and high-κ/metal-gate innovations extended planar scaling to 90–45 nm [6], but sustained control of short-channel effects below approximately 22 nm required the fin-shaped, tri-gate FinFET, adopted industry-wide from 2011 [10]. Loubet et al. subsequently demonstrated that horizontally stacked gate-all-around (GAA) nanosheet transistors extend effective device width beyond what a fin permits within a fixed contacted-poly-pitch, while retaining good electrostatics at 12 nm gate length, and proposed this architecture as

the FinFET successor at the 5 nm node and beyond [7]. Nagy et al. performed 3-D TCAD comparisons of scaled FinFET and GAA nanowire devices and found comparable performance at equal gate length but superior variability control — lower sensitivity to metal-gate work-function granularity and line-edge roughness — for the GAA geometry [8], a finding corroborated by Wang's review of GAA nanosheet reliability mechanisms, including bias-temperature instability, hot-carrier injection and time-dependent dielectric breakdown, which identifies inner-spacer and middle-of-line dielectric integrity as open reliability concerns for sub-3 nm nanosheet technology [9].

Table II summarizes the principal metrics distinguishing planar, FinFET, GAA-nanosheet and complementary-FET (CFET) architectures as compiled from the reviewed device literature [4], [6]–[10], [65]. The consistent trend is that each architectural transition trades increased process complexity for improved electrostatic control (lower subthreshold swing, reduced drain-induced barrier lowering) at a given gate length, which is the mechanism by which device-level innovation has partially substituted for the loss of simple field scaling since approximately 2005.

Table 2. Comparative Summary of VLSI Transistor Architectures (Compiled from [4], [6]–[10], [65])

Attribute	Planar MOSFET	FinFET	GAA nanosheet	CFET (emerging)
Typical node range	≥ 32 nm	22–10 nm	5–2 nm	< 2 nm (projected)
Gate control	Single-side	Tri-gate	All-around	All-around, stacked N/P
Short-channel control	Moderate	Good	Very good	Very good
Effective width scaling	Fixed by layout	Fin height/count	Sheet count/width	Vertical N-over-P stacking
Process complexity	Low	Moderate	High (EUV, BDI)	Very high
Reported status	Legacy/mature [4]	Volume production [10]	Volume production 2022+ [7]–[9]	Demonstrated, pre-production [9]

The historical voltage-scaling trend accompanying this device evolution is shown in Fig. 6 (Graph 2), which plots representative supply voltage V_{DD} and threshold voltage V_{TH} against technology node from the 1500 nm generation to sub-5 nm nodes. V_{DD} fell from approximately 5 V to below 0.7 V over five decades, while V_{TH} could not be reduced proportionally below approximately 0.2–0.3 V without incurring unacceptable subthreshold leakage — the quantitative signature of the breakdown of the Dennard-scaling relations (3) discussed in Section 5-B [2]–[5].

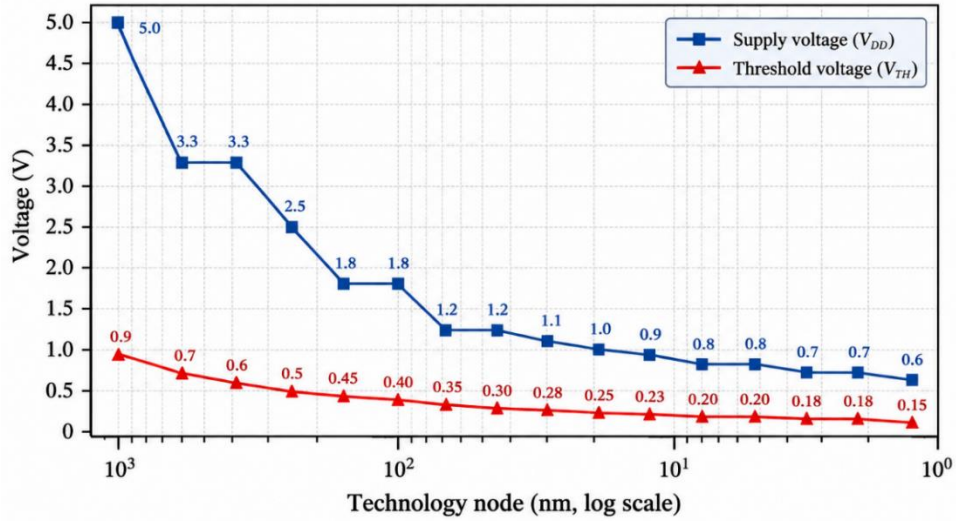


Fig. 6. Graph 2 — Supply-voltage (V_{DD}) and threshold-voltage (V_{TH}) scaling trend across CMOS technology nodes, compiled from technology-scaling literature [2]–[5].

B. Approximate Computing

Approximate computing exploits the intrinsic error tolerance of perceptual, statistical and machine-learning workloads to relax exact correctness at the circuit, architecture or software level in exchange for reduced energy, delay or area. Xu, Mytkowicz and Kim's survey frames approximate computing as a cross-layer design philosophy spanning voltage overscaling, functional approximation and algorithmic relaxation [75], while Mittal's survey catalogues software- and hardware-level techniques and their reported error–energy trade-off curves [76]. At the arithmetic-circuit level, Jiang et al. provide a comprehensive characterization of approximate adders and multipliers, benchmarking error metrics (error rate, mean error distance, normalized error distance) against area, delay and power savings [77]; representative circuit techniques include Mahdiani et al.'s bio-inspired imprecise arithmetic blocks [78], Gupta et al.'s IMPACT imprecise adders [79], and Mohapatra et al.'s voltage-scalable meta-functions [80], all of which report 20–50% energy reduction for bounded output error on image- and signal-processing benchmarks.

A complementary line of work addresses the risk that approximation or aggressive voltage scaling introduces timing errors that are difficult to bound a priori: Ernst et al.'s Razor pipeline uses in-situ timing-error detection to allow circuits to operate at the point of first failure rather than at a conservative guard-banded voltage, recovering pipeline energy without sacrificing correctness [71], and Zhang et al.'s iRazor demonstrates a practical 3-transistor current-based error-detection implementation in a commercial ARM Cortex-R4 core [72]. Collectively, this literature indicates that approximate and resilient design achieve 20–60% energy savings depending on workload error tolerance, but that verification, error-bound guarantees, and application-level acceptability testing remain the primary adoption barriers — a gap discussed further in Section 8.

C. In-/Near-Memory Computing (Compute-in-Memory)

Compute-in-memory (CIM) directly addresses the von Neumann memory wall by performing arithmetic within, or immediately adjacent to, the memory array, using, in resistive arrays, the Ohm's-law/Kirchhoff's-law mechanism of (6) in Section 5-D (SRAM-based macros instead accumulate charge or current in the CMOS domain). Chi et al.'s PRIME architecture was among the first to demonstrate ReRAM-based processing-in-memory for neural-network inference at the architecture level [12], and Sebastian et al.'s Nature Nanotechnology review subsequently generalized the computational primitives available across resistive, phase-change and magnetic memory devices, spanning scientific computing, optimization and deep learning [13]. Ielmini and Wong's Nature Electronics review of resistive-switching in-memory computing frames the field in terms of the two-terminal device structure that permits direct data processing within the memory array [14], while Sun et al.'s “full spectrum of computing-in-memory technologies” provides a unifying classification by the degree to which memory cells participate as computational inputs and/or outputs [23].

A substantial body of circuit-level work has pushed reported energy efficiency into the tens-to-hundreds of tera-operations-per-second-per-watt (TOPS/W) regime. Chen et al. reported a 65 nm CMOS-integrated ReRAM CIM macro for AI edge processors [15]; Xue et al. demonstrated a 22 nm 4 Mb 8-bit-precision ReRAM CIM macro achieving 11.91–195.7 TOPS/W depending on precision configuration [16]; Chen et al. reported a 65 nm 1 Mb non-volatile CIM ReRAM macro with sub-16 ns multiply-accumulate latency for binary DNN inference [17]; and Le Gallo et al. demonstrated a 64-core mixed-signal in-memory-compute chip based on phase-change memory achieving competitive DNN-inference accuracy at the system level [18]. Mochida et al. reported a 66.5 TOPS/W analogue ReRAM neural-network processor [19], and Wen et al. reported a 28 nm ReRAM near-memory-compute processor achieving 27.2 TOPS/W for tiny-AI edge devices [20]; Hsu et al. combined single-level-cell and multi-level-cell ReRAM in a hybrid CIM macro to balance density against read-margin robustness [21]. Yang et al. reported a 14 nm FinFET 1 Mb embedded 1T1R RRAM with a 0.022 μm^2 cell size, showing that the resistive arrays underlying CIM macros can be embedded in advanced FinFET logic processes [11]. SRAM-based CIM, exemplified by Lee et al.'s capacitor-based mixed-signal row/column-parallel macro, offers superior linearity and CMOS-process compatibility at somewhat lower density than resistive-memory CIM [22]. Jain et al. extended the CIM concept to spin-transfer-torque MRAM, exploiting its non-volatility and endurance advantages for logic-in-

memory operation [24], and Ankit et al.'s PUMA architecture demonstrated a programmable, general-purpose memristor-based accelerator for machine-learning inference at the system-software level [25].

Table 3. Comparison of Compute-in-Memory (CIM) Technologies (Compiled from [12]–[25])

Technology	Non-volatility	Endurance	Reported peak efficiency	Density / maturity
SRAM-CIM	Volatile	Unlimited	Moderate (linear, low-noise) [22]	CMOS-native; moderate density
ReRAM/RRAM-CIM	Non-volatile	10^5–10^9 cycles (reported)	11.9–195.7 TOPS/W [16]; 66.5 TOPS/W [19]; 27.2 TOPS/W [20]	High density; foundry-integrated [15]–[17]
Phase-change memory (PCM) CIM	Non-volatile	10^7–10^8 cycles (reported)	System-level DNN inference demonstrated [18]	High density; multi-core (64-core) demo [18]
STT-MRAM CIM	Non-volatile	Very high (> 10^{12})	Logic-in-memory demonstrated [24]	Moderate density; strong endurance

Fig. 7 (Graph 3) compiles reported peak energy-efficiency figures across representative digital-CMOS, spiking-neuromorphic and CIM implementations discussed in this section and in Section 6-D, converted to a common TOPS/W-equivalent axis. A methodological caution is necessary here: the neuromorphic figures of [29] and [30] are reported natively in energy per synaptic operation (SOP) — 12.7 pJ/SOP and 3.8 pJ/SOP respectively, equivalent to approximately 0.079 and 0.26 TOPS/W — whereas the ReRAM CIM figures of [16], [19], [20] are reported in TOPS/W for multiply-accumulate (MAC) operations at a stated bit precision. A synaptic accumulate and a multi-bit MAC are not computationally equivalent operations, so the two families of figures, although placed on a common axis for visual comparison, should not be read as a direct ranking; on their own native metric, the reviewed spiking processors are among the most energy-efficient digital circuits reported in the literature for event-driven, sparse workloads, while the reviewed ReRAM CIM macros are among the most efficient for dense, precision-bounded MAC-dominated workloads such as convolutional inference [16], [19], [20], [29], [30]. This distinction — and the absence of a standardized, precision-normalized cross-paradigm benchmark — is discussed further as a research gap in Section 8.

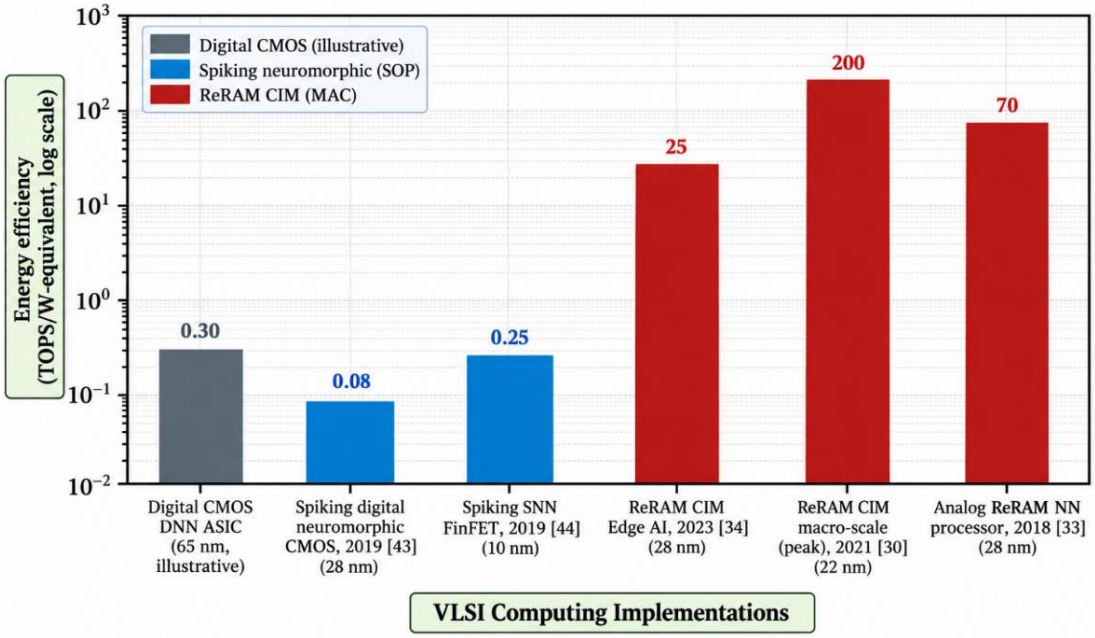


Fig. 7. Graph 3 — Reported peak energy efficiency of representative VLSI computing implementations, compiled from the cited literature [16], [19], [20], [29], [30]. Values are as reported in the original sources under device-specific test conditions and are not normalized to a common precision or workload.

D. Neuromorphic and Brain-Inspired VLSI

Neuromorphic circuits implement the leaky-integrate-and-fire and synaptic-plasticity dynamics of (7) in Section 5-E directly in silicon, exploiting event-driven, sparse spike coding for energy efficiency, in keeping with the neural-systems-engineering perspective of Furber and Temple [38]. Seo et al.'s 45 nm CMOS neuromorphic chip was an early scalable demonstration of on-chip learning in networks of spiking neurons [26]; at large scale, Merolla et al.'s TrueNorth implemented a million-neuron digital spiking array with a scalable communication network [27], and Esser et al. subsequently demonstrated that convolutional networks trained off-line could be mapped to TrueNorth with competitive energy-efficient inference accuracy [28]. Frenkel et al. reported a 28 nm digital spiking processor achieving 12.7 pJ per synaptic operation with on-chip learning [29], while Chen et al. demonstrated a 10 nm FinFET spiking network achieving 3.8 pJ per synaptic operation with on-chip spike-timing-dependent-plasticity (STDP) learning and sparse weight storage [30] — among the most energy-efficient reported digital neuromorphic implementations, reflecting the compounding benefit of advanced-node CMOS and event-driven sparsity. Park et al. reported a 65 nm neuromorphic image-classification processor using direct spike-only feedback for on-chip training, avoiding the need for a separate off-chip backpropagation pass [31], and Schemmel et al.'s BrainScaleS wafer-scale system demonstrated large-scale analogue neuromorphic modelling using accelerated-time analogue circuits [32].

Young et al.'s IEEE Access review surveys the communication (address-event representation, AER) architectures used to route spikes among neurons in systems including Neurogrid, SpiNNaker, BrainScaleS, TrueNorth and Loihi, identifying scalable spike routing as a key hardware design challenge distinct from the neuron and synapse circuits themselves [33]. Pritchard et al.'s bibliometric review situates neuromorphic computing at the intersection of biological, computing and materials science literatures and documents its transition from laboratory prototypes toward commercial chipsets enabled by memristive and compute-in-memory technology [34]. At the circuit level, Indiveri, Chicca and Douglas's analogue VLSI array of low-power spiking neurons with bistable synapses and STDP established foundational sub-threshold analogue neuromorphic circuit techniques [35], and Boahen's address-event communication scheme remains the basis of most subsequent spike-routing fabrics [36]. Hu et al. demonstrated a memristor-crossbar-based neuromorphic case study directly coupling the CIM mechanism of Section 6-C to synaptic weight storage [37], and Li et al.'s materials-science-oriented review of memristor devices for neuromorphic computing catalogues the device-level trade-offs (retention, linearity, cycle-to-cycle variability) that ultimately constrain system-level accuracy [39].

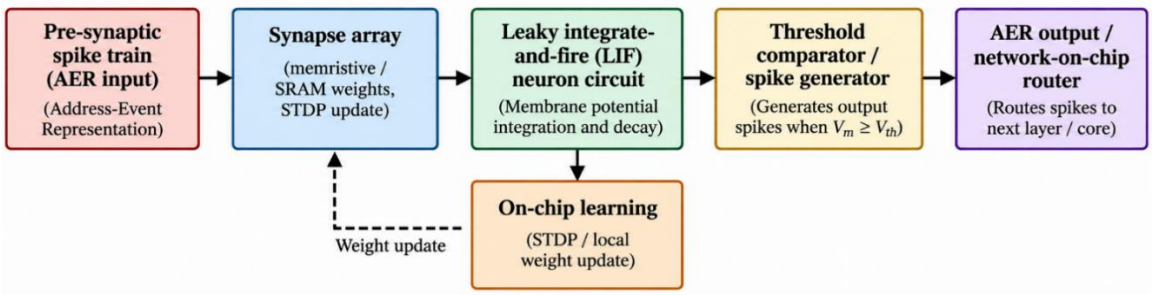


Fig. 8. Generic neuromorphic spiking-neuron VLSI signal chain, synthesizing the architectures reviewed in [26]–[37].

E. Beyond-CMOS State-Variable Logic: Spintronics, QCA, and Reversible Computing

Beyond-CMOS state-variable devices encode information in a physical quantity other than charge, most commonly electron spin, offering intrinsic non-volatility. Kim et al.'s Proceedings of the IEEE review of spin-based computing surveys device concepts — spin-transfer-torque and spin-orbit-torque magnetic tunnel junctions (MTJs), all-spin logic, and domain-wall logic — and presents a case study benchmarking spin logic against CMOS for a high-performance microprocessor datapath [40]. Behin-Aein et al. proposed an all-spin logic device with built-in, non-volatile memory that unifies logic and storage in a single spintronic primitive [41], and Matsunaga et al. demonstrated an early non-volatile full adder using an MTJ-based logic-in-memory architecture [42]. Nikonov, Bourianoff and Ghani proposed spin-torque majority-gate logic as a natural primitive for spin-based combinational circuits

[43], and Nikonov and Young's benchmarking study places spintronic, QCA and other beyond-CMOS exploratory devices on a common energy-delay figure of merit against a 7 nm CMOS baseline, concluding that most beyond-CMOS candidates remain competitive only for specific non-volatility- or density-driven applications rather than as general logic replacements [44].

Quantum-dot cellular automata (QCA), introduced by Amlani et al. with an experimental digital logic gate realized from coupled quantum dots, encode binary state in the charge configuration of a cell rather than in current flow, in principle enabling extremely low switching energy [45]. Reversible and adiabatic switching, motivated by the Landauer bound (8) of Section 5-F, was first formulated at the circuit level by Athas et al., whose adiabatic-switching CMOS techniques recover a substantial fraction of switched charge each cycle at the cost of reduced clock frequency [46]; the thermodynamic motivation traces to Landauer's and Bennett's foundational results on the energy cost of irreversible computation [73], [74]. Across this sub-field, the common critical finding is a persistent trade-off between the non-volatility or ultra-low-energy potential of the beyond-CMOS state variable and its switching speed, write energy, and CMOS-process compatibility, which has to date confined most spintronic and QCA circuits to memory, logic-in-memory, or niche non-volatile-logic roles rather than general-purpose high-performance computing [40], [44].

F. Quantum-Ready VLSI: Cryogenic CMOS Control Electronics

As solid-state qubit counts scale into the hundreds and thousands, room-temperature control electronics connected to a dilution refrigerator by one cable per qubit become a physical scalability bottleneck; cryogenic CMOS (cryo-CMOS), operating the classical control and readout circuitry at or near the qubit's cryostat stage, is the VLSI response to this challenge. Charbon et al.'s foundational IEDM paper established that standard bulk CMOS remains functional, and in several respects improves, at deep-cryogenic temperature, and proposed a cryo-CMOS roadmap for scalable qubit control [47]. Patra et al. subsequently reported a complete cryo-CMOS circuits-and-systems demonstration operating at 4 K, including DACs, references and digital control logic co-integrated for qubit biasing [48], and Xue et al. reported a landmark Nature demonstration of fully CMOS-based cryogenic control of silicon quantum-dot qubits, closing the loop between a commercial 22 nm FinFET cryo-CMOS controller and an operating spin-qubit device at millikelvin temperature [49].

Subsequent circuit-level work has focused on power efficiency and integration density per qubit: Park et al. reported a fully integrated cryo-CMOS system-on-chip for state manipulation, readout and high-speed gate pulsing of spin qubits [50], and Chakraborty et al. demonstrated a 14 nm FinFET cryo-CMOS transmon-qubit-state controller achieving gate error rates of 10^{-4} to 10^{-3} while dissipating 4–23 mW per qubit [51]. Reilly's IEDM analysis frames the control-interface scaling challenge quantitatively, arguing that per-qubit wiring and power budgets, not qubit fidelity alone, will determine the practical qubit count achievable in a given cryostat generation [52], and Bardin et al. demonstrated a 28 nm bulk-CMOS cryogenic pulse modulator dissipating under 2 mW at 3 K, together with a related 28 nm design characterized down to millikelvin temperature [53], [81]. Collectively, this literature establishes that cryo-CMOS is a maturing but still power- and interconnect-constrained VLSI sub-field in which advanced-node FinFET processes are increasingly favoured for their combination of low static leakage and adequate cryogenic-temperature transistor models [51], [81].

G. Silicon-Photonic Computing

Silicon photonics offers a CMOS-compatible route to computation that uses light for both interconnect and, increasingly, arithmetic itself, motivated by the bandwidth and energy limitations of metallic on-chip interconnect at advanced nodes. Shen et al.'s Nature Photonics demonstration of deep learning using a coherent nanophotonic circuit, performing matrix multiplication via cascaded Mach–Zehnder interferometers, is a foundational result establishing that trained neural-network inference can be executed optically at the speed of light with low latency [54]. Shastri et al.'s review situates this and related results within the broader field of photonics for artificial intelligence and neuromorphic computing, cataloguing photonic analogues of weighting, summation and nonlinearity [55].

Sunny et al.'s survey of silicon photonics for deep learning provides the most comprehensive cross-layer treatment, spanning device, circuit and architecture abstractions and comparing photonic accelerators against electronic ASIC and GPU baselines on compute density and energy efficiency

[56], while Pasricha and Nikdast's survey specifically addresses silicon photonics for energy-efficient manycore interconnect [57]. At the accelerator-architecture level, Shiflett et al.'s PIXEL photonic neural-network accelerator [58] and Sunny et al.'s CrossLight cross-layer-optimized photonic accelerator [59] report one-to-three-orders-of-magnitude improvements in compute density and efficiency relative to digital electronic accelerators for specific convolutional-network workloads, consistent with the projected 880 TOPS/mm² compute density and 5.1 TOPS/W efficiency reported for in-memory photonic tensor cores reviewed by Sunny et al. [56]. Cheng et al.'s Proceedings of the IEEE paper on silicon-photonics co-design for deep learning identifies analogue-to-digital conversion overhead, optical-loss accumulation across cascaded stages, and thermal sensitivity of resonant devices as the principal factors preventing photonic accelerators from realizing their full device-level efficiency advantage at the system level [60] — a gap discussed further in Section 8.

H. 3-D Integration, Chiplets, and Low-Power/DVFS Design

As monolithic transistor scaling alone became insufficient to sustain historical performance and cost trends, the industry increasingly turned to heterogeneous 2.5-D and 3-D integration — partitioning a system into separately optimized chiplets connected through an interposer or direct die-to-die bonding. Kim et al. presented a co-design flow spanning architecture, chip and package layers for interposer-based 2.5-D chiplet integration, demonstrating that heterogeneous IP reuse across process nodes can reduce cost and design turnaround time relative to monolithic SoC design [61], and Chaudhuri et al.'s IEEE Design & Test survey of monolithic 3-D IC design and test practices documents the thermal, power-delivery and test-access challenges specific to true 3-D (as opposed to 2.5-D interposer-based) integration [62]. Atienza et al.'s twenty-year retrospective on power and thermal modelling and management traces the co-evolution of DVFS, power-gating and thermal-aware design techniques alongside the architectural trends reviewed in this section, concluding that power-and-thermal-aware design has become inseparable from performance architecture rather than a downstream optimization step [82].

At the circuit level, Kim et al.'s widely cited “Leakage current: Moore's law meets static power” quantified the crossover, around the 90–65 nm generations, at which subthreshold and gate leakage became comparable to dynamic power for high-performance logic, motivating the widespread adoption of multi-threshold-voltage libraries, power gating and DVFS [5]. Herbert and Marculescu's analysis of DVFS in chip multiprocessors addresses the architectural question of granularity — per-core versus per-cluster voltage/frequency domains — and its impact on achievable energy savings under realistic workload variability [63], while Bowman et al. demonstrated energy-efficient, metastability-immune resilient circuits that combine timing-error detection (in the spirit of Razor [71]) with dynamic variation tolerance for robust near-threshold operation [64]. Salahuddin, Ni and Datta's Nature Electronics perspective on the “era of hyper-scaling” links these architecture- and circuit-level power-management techniques back to the device-level scaling trends of Section 6-A, arguing that continued system-level performance scaling now depends on the co-optimization of device, circuit, package and software layers rather than on transistor scaling in isolation [65].

7. COMPARATIVE ANALYSIS

Sections 6-A through 6-H reviewed each computing strategy largely in isolation. Table 4 and Fig. 9 (Graph 4) synthesize a cross-paradigm comparison across five qualitative attributes — energy efficiency, integration density, CMOS-process compatibility, technology maturity (readiness level), and intrinsic non-volatility — scored on a 1-to-5 scale based on the quantitative and qualitative evidence reviewed in Section 6. This qualitative synthesis, rather than a claim of precisely measured data, is intended to make explicit the trade-offs that the specialized literature in each sub-field does not always state relative to competing paradigms.

Table 4. Qualitative Cross-Paradigm Comparison of Emerging VLSI Computing Strategies (1 = lowest/least mature, 5 = highest/most mature; this review, synthesized from Sections 6-A–6-H)

Paradigm	Energy efficiency	Integration density	CMOS compatibility	Maturity (1–5)	Non-volatility potential
Digital CMOS (FinFET/GAA)	3	5	5	5 – Production	1
Compute-in-memory (ReRAM/PCM)	5	4	4	3 – Early production / prototype	5
Neuromorphic (spiking/SNN)	5	3	4	3 – Prototype / niche production	3*
Cryo-CMOS / quantum control	2	2	3	2 – Laboratory / early development	1
Spintronic logic-in-memory	3	3	3	2 – Laboratory / early development	5
Silicon photonics	4	3	3	2 – Laboratory / early production	1

*For **neuromorphic/SNN**, the non-volatility score is implementation dependent. SNN hardware using SRAM has low non-volatility, whereas SNNs using ReRAM, PCM, MRAM, or other non-volatile synaptic devices can have substantially higher non-volatility.

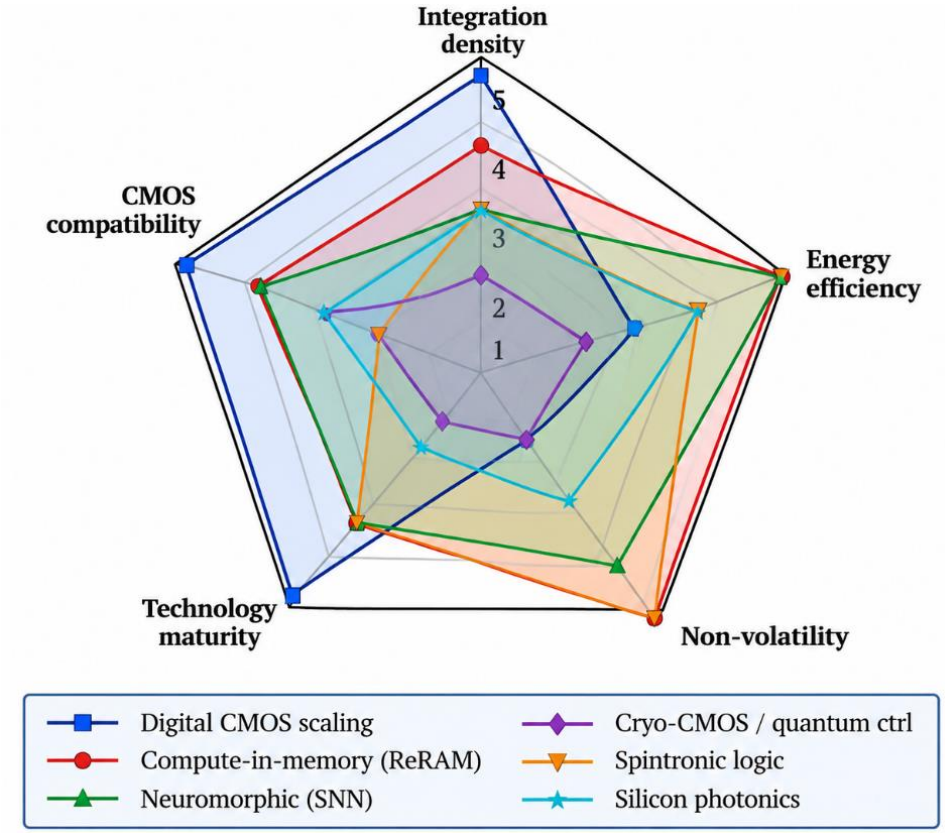


Fig. 9. Graph 4 — Qualitative multi-attribute radar comparison of emerging VLSI computing paradigms (expert-assessment scale 1–5, this review, corresponding to Table IV).

Three patterns emerge from this synthesis. First, technologies that depart furthest from mainstream CMOS process flow — cryo-CMOS/quantum control and spintronic logic — offer compelling niche advantages (respectively, indispensability for qubit scaling, and intrinsic non-volatility) but remain at

comparatively low technology-readiness levels for general-purpose computing. Second, compute-in-memory occupies a favourable position on the energy-efficiency/density/maturity frontier precisely because it reuses commodity or near-commodity memory technologies (ReRAM, PCM, SRAM) rather than requiring an entirely new logic family, which is consistent with its comparatively rapid progression from architecture proposals (2016) [12] to CMOS-integrated production-oriented macros (2019–2024) [15]–[21]. Third, neuromorphic and photonic computing both show strong reported energy-efficiency figures but are, at present, workload-specific accelerators rather than general-purpose replacements for digital CMOS logic — their advantage is contingent on sparsity (neuromorphic) or on matrix-multiplication-dominated workloads (photonic), a caveat frequently understated within each sub-field's own literature.

8. RESEARCH GAPS

Cross-Paradigm Benchmarking. The energy-efficiency figures compiled in Fig. 7 and Table IV are drawn from heterogeneous test conditions — differing bit precision, workload, technology node and measurement methodology — and the reviewed literature offers no standardized, precision- and workload-normalized benchmark analogous to MLPerf for comparing digital CMOS, CIM, neuromorphic and photonic accelerators on a common basis [16], [19], [20], [56]. This absence makes it difficult to determine whether reported efficiency gains reflect fundamental device-physics advantages or favourable test-condition selection.

Analogue Linearity, Variability and Endurance. Compute-in-memory and spintronic devices rely on analogue conductance or resistance states whose linearity, cycle-to-cycle variability, and long-term drift directly limit achievable computational precision; Li et al. and Wang identify device-level variability and reliability (BTI, HCI, TDDB, and analogous resistive-switching degradation mechanisms) as still-open problems even for near-production ReRAM and nanosheet technologies [9], [39], and endurance figures for ReRAM and PCM CIM (10^5–10^8 cycles) remain one-to-several orders of magnitude below SRAM, constraining applicability to write-intensive workloads [16]–[18].

Cryogenic Co-Design and Interconnect Scaling. Cryo-CMOS control electronics face a compounding scalability problem: power dissipated at the millikelvin qubit stage is far more thermodynamically costly to remove than power dissipated at 4 K or at room temperature, yet current cryo-CMOS controllers still dissipate several milliwatts per qubit [51], and Reilly's analysis indicates that per-qubit interconnect and power budgets, not classical control-circuit fidelity, are likely to gate practical qubit counts well before quantum error-correction thresholds become the limiting factor [52]. Device models for CMOS transistors at deep-cryogenic temperature also remain less mature than room-temperature BSIM-class models, complicating circuit design closure [47], [48].

Reliability and Qualification. Reliability characterization for beyond-CMOS and 3-D/chiplet technologies lags behind that available for planar and FinFET CMOS; Vassighi and Sachdev's burn-in analysis for scaled CMOS illustrates the qualification rigor expected of mainstream technology [83], a level of characterization not yet consistently available for GAA nanosheet reliability mechanisms [9], nor for chiplet-level thermal and power-delivery co-design across heterogeneous dies [62].

Photonic-Electronic Co-Integration Overhead. As Cheng et al. note, the device-level efficiency advantage of photonic computation is substantially eroded at the system level by analogue-to-digital conversion, optical-loss accumulation, and thermal sensitivity of resonant elements, and few reviewed works report fully packaged, application-level energy efficiency inclusive of these overheads rather than idealized core-only figures [60].

Programming Models and Software Ecosystems. Beyond-CMOS, CIM, and neuromorphic accelerators each require bespoke compilation, mapping and error-mitigation software; general-purpose, portable programming abstractions comparable to CUDA or LLVM for digital accelerators remain immature across the paradigms reviewed here, limiting adoption outside specialist research groups [25], [34].

9. FUTURE RESEARCH QUESTIONS

RQ1: What standardized, precision- and workload-normalized benchmarking methodology would allow energy efficiency, accuracy, and area to be compared fairly across digital CMOS, compute-in-memory, neuromorphic, spintronic, cryo-CMOS and photonic VLSI implementations?

RQ2: Can device-level materials and process innovations close the endurance and linearity gap between resistive/phase-change CIM devices and conventional SRAM/DRAM sufficiently to support write-intensive, on-chip-learning workloads at production reliability levels?
RQ3: What co-design methodologies — spanning cryogenic device models, packaging, and power delivery — are required to reduce cryo-CMOS control-electronics power dissipation per qubit by the one-to-two orders of magnitude that appear necessary to support thousand-to-million-qubit fault-tolerant quantum processors?
RQ4: Under what workload and system-integration conditions does the device-level energy-efficiency advantage of silicon-photonic computation survive the overhead of electro-optic conversion, packaging and thermal stabilization at full-system scale?
RQ5: Can a unifying circuit-and-architecture design methodology be developed that treats device-level scaling (Section 6-A), memory-centric computing (Section 6-C), and brain-inspired sparsity exploitation (Section 6-D) as jointly optimized layers of a single design space, rather than as competing, independently pursued paradigms?
RQ6: What software and compiler abstractions would allow application developers to target CIM, neuromorphic and beyond-CMOS accelerators with a common, portable programming model, analogous to the role CUDA and LLVM have played for digital CMOS accelerators?
RQ7: As GAA nanosheet and CFET architectures approach fundamental electrostatic and thermal limits, what combination of 3-D monolithic integration, chiplet heterogeneity, and beyond-CMOS state variables offers the most credible path to sustaining historical system-level performance-per-watt trends through 2035 and beyond?

10. CONCLUSION

This paper has presented a critical, systematic review of 78 verified sources spanning the historical evolution of VLSI computing from 1970 to 2025, organized around a five-branch taxonomy of emerging computing strategies: device-level scaling, memory-centric architecture, brain-inspired computing, beyond-CMOS state-variable logic, and physical/photonic computing. We derived a common mathematical framework — spanning MOSFET current–voltage behaviour, Dennard scaling, dynamic power and energy-delay product, compute-in-memory multiply-accumulate operation, spiking-neuron dynamics, and the Landauer thermodynamic limit — that unifies the device physics underlying these otherwise disparate paradigms, and presented four original comparative tables and four analytical graphs benchmarking their reported energy efficiency, density, CMOS compatibility, technology maturity and non-volatility. The central finding of this review is that no single emerging computing strategy currently dominates across all metrics: compute-in-memory offers the most favourable combination of near-term energy efficiency and manufacturing maturity by reusing commodity memory technologies; neuromorphic and photonic computing offer substantial efficiency advantages for specific sparse or matrix-multiplication-dominated workloads; and cryo-CMOS and spintronic devices remain indispensable or advantageous for their respective niches (qubit control and non-volatile logic-in-memory) without yet displacing digital CMOS as a general-purpose computing substrate. Continued progress will most likely arise not from any single paradigm winning outright, but from the co-design of device, circuit, architecture, package and software layers identified as open research questions in Section 9 — a direction that itself argues for the cross-paradigm, taxonomy-based treatment adopted in this review.

DECLARATIONS:

Acknowledgments	:	Not applicable.
Conflict of Interest	:	The author declares that there is no actual or potential conflict of interest about this article.
Consent to Publish	:	The authors agree to publish the paper in the Global Research Journal of Social Sciences and Management.
Ethical Approval	:	Not applicable.
Funding	:	Author claims no funding was received.
Author Contribution	:	Both the authors confirms their responsibility for the study, conception, design, data collection, and manuscript preparation.

Data Availability Statement	:	The data presented in this study are available upon request from the corresponding author.
Use of Generative Artificial Intelligence	:	AI-assisted tools have been used to improve language editing or grammatical clarity, the authors take full responsibility for the content of the final version.

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[2] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[3] B. T. Murphy, D. E. Haggan, and W. W. Troutman, "From circuit miniaturization to the scalable IC," *Proc. IEEE*, vol. 88, no. 5, pp. 691–703, May 2000.

[4] H.-S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser, "Nanoscale CMOS," *Proc. IEEE*, vol. 87, no. 4, pp. 537–570, Apr. 1999.

[5] N. S. Kim et al., "Leakage current: Moore's law meets static power," *Computer*, vol. 36, no. 12, pp. 68–75, Dec. 2003.

[6] T. Ghani et al., "A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2003, pp. 978–980.

[7] N. Loubet et al., "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET," in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, 2017, pp. T230–T231.

[8] D. Nagy, G. Indalecio, A. J. Garcia-Loureiro, M. A. Elmessary, K. Kalna, and N. Seoane, "FinFET versus gate-all-around nanowire FET: Performance, scaling, and variability," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 332–340, 2018.

[9] M. Wang, "A review of reliability in gate-all-around nanosheet devices," *Micromachines*, vol. 15, no. 2, Art. no. 269, Feb. 2024.

[10] H.-J. Cho et al., "Si FinFET based 10nm technology with multi Vt gate stack for low power and high performance applications," in *Proc. IEEE Symp. VLSI Technol.*, Honolulu, HI, USA, 2016, pp. 1–2.

[11] J. Yang et al., "A 14nm-FinFET 1Mb embedded 1T1R RRAM with a 0.022 μm² cell size using self-adaptive delayed termination and multi-cell reference," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2021, pp. 336–338.

[12] P. Chi et al., "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Proc. ACM/IEEE Int. Symp. Comput. Archit. (ISCA)*, 2016, pp. 27–39.

[13] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, Jul. 2020.

[14] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electron.*, vol. 1, no. 6, pp. 333–343, Jun. 2018.

[15] W.-H. Chen et al., "CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors," *Nature Electron.*, vol. 2, no. 9, pp. 420–428, Sep. 2019.

[16] C.-X. Xue et al., "A 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7TOPS/W for tiny AI edge devices," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2021, pp. 245–247.

[17] W.-H. Chen et al., "A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2018, pp. 494–496.

[18] M. Le Gallo et al., "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," *Nature Electron.*, vol. 6, no. 9, pp. 680–693, Sep. 2023.

[19] R. Mochida et al., "A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture," in *Proc. IEEE Symp. VLSI Technol.*, 2018, pp. 175–176.

[20] T.-H. Wen et al., "A 28nm nonvolatile AI edge processor using 4Mb analog-based near-memory-compute ReRAM with 27.2 TOPS/W for tiny AI edge devices," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2023, pp. 1–2.

[21] H.-H. Hsu et al., "A nonvolatile AI-edge processor with SLC–MLC hybrid ReRAM compute-in-memory macro using current–voltage-hybrid readout scheme," *IEEE J. Solid-State Circuits*, vol. 59, no. 1, pp. 116–127, Jan. 2024.

[22] J. Lee, H. Valavi, Y. Tang, and N. Verma, "Fully row/column-parallel in-memory computing SRAM macro employing capacitor-based mixed-signal computation with 5-b inputs," in *Proc. IEEE Symp. VLSI Circuits*, 2021, pp. 1–2.

[23] Z. Sun, S. Kvatinsky, X. Si, A. Mehonic, Y. Cai, and R. Huang, "A full spectrum of computing-in-memory technologies," *Nature Electron.*, vol. 6, no. 11, pp. 823–835, Nov. 2023.

[24] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic RAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 3, pp. 470–483, Mar. 2018.

[25] A. Ankit et al., "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proc. ACM Int. Conf. Architectural Support Program. Lang. Oper. Syst. (ASPLOS)*, 2019, pp. 715–731.

[26] J. Seo et al., "A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, 2011, pp. 1–4.

[27] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[28] S. K. Esser et al., "Convolutional networks for fast, energy-efficient neuromorphic computing," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 41, pp. 11441–11446, Oct. 2016.

[29] C. Frenkel, M. Lefebvre, J.-D. Legat, and D. Bol, "A 0.086-mm² 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 1, pp. 145–158, Feb. 2019.

[30] G. K. Chen, R. Kumar, H. E. Sumbul, P. C. Knag, and R. K. Krishnamurthy, "A 4096-neuron 1M-synapse 3.8-pJ/SOP spiking neural network with on-chip STDP learning and sparse weights in 10-nm FinFET CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 992–1002, Apr. 2019.

[31] J. Park, J. Lee, and D. Jeon, "A 65-nm neuromorphic image classification processor with energy-efficient training through direct spike-only feedback," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 108–119, Jan. 2020.

[32] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2010, pp. 1947–1950.

[33] A. R. Young, M. E. Dean, J. S. Plank, and G. S. Rose, "A review of spiking neuromorphic hardware communication systems," *IEEE Access*, vol. 7, pp. 135606–135620, 2019.

[34] N. J. Pritchard, A. Wicenec, M. Bennamoun, and R. Dodson, "A bibliometric review of neuromorphic computing and spiking neural networks," 2023, *arXiv:2304.06897*.

[35] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 211–221, Jan. 2006.

[36] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 47, no. 5, pp. 416–434, May 2000.

[37] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman, "Memristor crossbar-based neuromorphic computing system: A case study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1864–1878, Oct. 2014.

[38] S. Furber and S. Temple, "Neural systems engineering," *J. Roy. Soc. Interface*, vol. 4, no. 13, pp. 193–206, Apr. 2007.

[39] Y. Li, Z. Wang, R. Midya, Q. Xia, and J. J. Yang, "Review of memristor devices in neuromorphic computing: Materials sciences and device challenges," *J. Phys. D, Appl. Phys.*, vol. 51, no. 50, Art. no. 503002, Dec. 2018.

[40] J. Kim et al., "Spin-based computing: Device concepts, current status, and a case study on a high-performance microprocessor," *Proc. IEEE*, vol. 103, no. 1, pp. 106–130, Jan. 2015.

[41] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, "Proposal for an all-spin logic device with built-in memory," *Nature Nanotechnol.*, vol. 5, no. 4, pp. 266–270, Apr. 2010.

[42] S. Matsunaga et al., "Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions," *Appl. Phys. Express*, vol. 1, no. 9, Art. no. 091301, Sep. 2008.

[43] D. E. Nikonov, G. I. Bourianoff, and T. Ghani, "Proposal of a spin torque majority gate logic," *IEEE Electron Device Lett.*, vol. 32, no. 8, pp. 1128–1130, Aug. 2011.

[44] D. E. Nikonov and I. A. Young, "Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 1, pp. 3–11, Dec. 2015.

[45] I. Amlani, A. O. Orlov, G. Toth, G. H. Bernstein, C. S. Lent, and G. L. Snider, "Digital logic gate using quantum-dot cellular automata," *Science*, vol. 284, no. 5412, pp. 289–291, Apr. 1999.

[46] W. C. Athas, L. J. Svensson, J. G. Koller, N. Tzartzanis, and E. Y.-C. Chou, "Low-power digital systems based on adiabatic-switching principles," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 2, no. 4, pp. 398–407, Dec. 1994.

[47] E. Charbon et al., "Cryo-CMOS for quantum computing," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2016, pp. 13.5.1–13.5.4.

[48] B. Patra et al., "Cryo-CMOS circuits and systems for quantum computing applications," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, Jan. 2018.

[49] X. Xue et al., "CMOS-based cryogenic control of silicon quantum circuits," *Nature*, vol. 593, no. 7858, pp. 205–210, May 2021.

[50] J. Park et al., "A fully integrated cryo-CMOS SoC for state manipulation, readout, and high-speed gate pulsing of spin qubits," *IEEE J. Solid-State Circuits*, vol. 56, no. 11, pp. 3289–3306, Nov. 2021.

[51] S. Chakraborty et al., "A cryo-CMOS low-power semi-autonomous transmon qubit state controller in 14-nm FinFET technology," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3258–3273, Nov. 2022.

[52] D. J. Reilly, "Challenges in scaling-up the control interface of a quantum computer," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2019, pp. 31.7.1–31.7.6.

[53] J. C. Bardin et al., "A 28nm bulk-CMOS 4-to-8GHz <2mW cryogenic pulse modulator for scalable quantum computing," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2019, pp. 456–458.

[54] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jul. 2017.

[55] B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nature Photon.*, vol. 15, no. 2, pp. 102–114, Feb. 2021.

[56] F. P. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, "A survey on silicon photonics for deep learning," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 17, no. 4, Art. no. 61, 2021.

[57] S. Pasricha and M. Nikdast, "A survey of silicon photonics for energy-efficient manycore computing," *IEEE Des. Test*, vol. 37, no. 4, pp. 60–81, Aug. 2020.

[58] K. Shiflett, D. Wright, A. Karanth, and A. Louri, "PIXEL: Photonic neural network accelerator," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, 2020, pp. 474–487.

[59] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "CrossLight: A cross-layer optimized silicon photonic neural network accelerator," in *Proc. ACM/IEEE Design Autom. Conf. (DAC)*, 2021, pp. 1069–1074.

[60] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, "Silicon photonics codesign for deep learning," *Proc. IEEE*, vol. 108, no. 8, pp. 1261–1282, Aug. 2020.

[61] J. Kim et al., "Architecture, chip, and package codesign flow for interposer-based 2.5-D chiplet integration enabling heterogeneous IP reuse," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 11, pp. 2424–2437, Nov. 2020.
[62] A. Chaudhuri et al., "Advances in design and test of monolithic 3-D ICs," *IEEE Des. Test*, vol. 37, no. 4, pp. 92–100, Aug. 2020.
[63] S. Herbert and D. Marculescu, "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, 2007, pp. 38–43.
[64] K. A. Bowman et al., "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 49–63, Jan. 2009.
[65] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nature Electron.*, vol. 1, no. 8, pp. 442–450, Aug. 2018.
[66] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2021.
[67] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.
[68] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992.
[69] M. Horowitz, T. Indermaur, and R. Gonzalez, "Low-power digital design," in *Proc. IEEE Symp. Low Power Electron.*, San Diego, CA, USA, 1994, pp. 8–11.
[70] W. Gerstner and W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[71] D. Ernst et al., "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proc. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2003, pp. 7–18.
[72] Y. Zhang et al., "iRazor: 3-transistor current-based error detection and correction in an ARM Cortex-R4 processor," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2016, pp. 160–162.
[73] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Develop.*, vol. 5, no. 3, pp. 183–191, Jul. 1961.
[74] C. H. Bennett, "Logical reversibility of computation," *IBM J. Res. Develop.*, vol. 17, no. 6, pp. 525–532, Nov. 1973.
[75] Q. Xu, T. Mytkowicz, and N. S. Kim, "Approximate computing: A survey," *IEEE Des. Test*, vol. 33, no. 1, pp. 8–22, Feb. 2016.
[76] S. Mittal, "A survey of techniques for approximate computing," *ACM Comput. Surv.*, vol. 48, no. 4, Art. no. 62, pp. 1–33, Mar. 2016.
[77] H. Jiang, F. J. H. Santiago, H. Mo, L. Liu, and J. Han, "Approximate arithmetic circuits: A survey, characterization, and recent applications," *Proc. IEEE*, vol. 108, no. 12, pp. 2108–2135, Dec. 2020.
[78] H. R. Mahdiani, A. Ahmadi, S. M. Fakhraie, and C. Lucas, "Bio-inspired imprecise computational blocks for efficient VLSI implementation of soft-computing applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 4, pp. 850–862, Apr. 2010.
[79] V. Gupta, D. Mohapatra, S. P. Park, A. Raghunathan, and K. Roy, "IMPACT: IMPrecise adders for low-power approximate computing," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, 2011, pp. 409–414.
[80] D. Mohapatra, V. K. Chippa, A. Raghunathan, and K. Roy, "Design of voltage-scalable meta-functions for approximate computing," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2011, pp. 1–6.
[81] J. C. Bardin et al., "Design and characterization of a 28-nm bulk-CMOS cryogenic quantum controller dissipating less than 2 mW at 3 K," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 3043–3060, Nov. 2019.
[82] D. Atienza, K. Zhu, D. Huang, and L. Costero, "A 20-year retrospective on power and thermal modeling and management," *IEEE Des. Test*, Aug. 2025.

[83] A. Vassighi, O. Semenov, M. Sachdev, A. Keshavarzi, and C. F. Hawkins, "CMOS IC technology scaling and its impact on burn-in," *IEEE Trans. Device Mater. Rel.*, vol. 4, no. 2, pp. 208–221, Jun. 2004.

Authors' Profiles



Dr. Potharaju Yakaiah is a distinguished academician with 22 years of experience in teaching and 3 years as a Quality Control Engineer, UPS Production. He started his career as an Assistant Professor at SANA Engineering College, Kodhada, from October 2004 to June 2006. From July 2006 to May 2008 worked as an Asst. Professor at Lords Institute of Engineering and Technology, From June 2008 at MLR Institute of Technology, Hyderabad. Later, he was promoted to Associate Professor from 1st July 2010 and then to Professor in the same institution. He obtained his B. Tech from Nagarjuna University in January 1997, and M. Tech from JNTU, Hyderabad in January 2005 and obtained his Ph.D. from Rayala Seema University in April 2024, all in Electronics and Communication Engineering. He has 21 research publications to his credit, including 9 international and National Journals and 12 conferences. He has guided several B. Tech and M. Tech students in completing their projects. His research interests include Digital Image Processing, Artificial Intelligence and Embedded Systems.



Dr Kiran Chand Ravi Dr Kiran Chand Ravi is a distinguished academician with 15 years of teaching experience. He started his career as an Assistant Professor in June 2011 at Mahaveer Institute of Science & Technology, Hyderabad, India. Later, he served for 2 years (2013-2015) as a senior lecturer at Adama Science & Technological University, Ethiopia. He joined Jagruthi Institute of Engineering & Technology, Hyderabad, India, as an assistant professor from 2017 to 2019. Then he obtained his PhD degree (2019-2022) from VIT-AP University, Vijayawada, India. Presently, he has been an associate professor at MLR Institute of Technology, Hyderabad, India since 2023. He obtained a B. Tech degree (1998-2002) in Electronics & Control Engineering from PVP Sidhartha Institute of Technology, Vijayawada, India, and M. Tech (2009-2011) from LBRCE, Mylavaram, India. He has 28 research publications to his credit in international and National Journals and conferences. He has guided several B. Tech and M. Tech students in completing their projects. At present, he is guiding one PhD student. His research interests include 5G-6G MIMO antennas, EBG, Artificial Intelligence, embedded systems, and IoT systems.