The VLSI Design Evolution and Emerging Trends in CMOS, 3D Integration, and AI-Assisted EDA

Zhangyang Ding *

Department of Electrical and Computer Engineering, Michigan State University, 48823, Michigan, United States

* Corresponding Author Email: dingzha3@msu.edu

Abstract. The development of Very-Large-Scale Integration (VLSI) design has followed three connected elements which include continuous device scaling and architectural diversification and methodological intelligence growth. The three forces have successively expanded what designers can achieve through performance improvements and power efficiency gains and design complexity increases. The initial development period focused on scaling-based optimization which used transistor size reduction to follow Moore's and Dennard's laws for achieving better speed and density performance. The emergence of physical and power constraints led to a transition to three-dimensional integration and material breakthroughs including Fin Field-Effect Transistors (FinFETs) and gate-all-around devices and carbon nanotube transistors. Artificial intelligence entered the third phase to transform electronic design automation (EDA) through data-driven optimization and reliability assessment and design space exploration. The review demonstrates how VLSI foundations have developed through physical scaling to intelligent automation while showing the path toward post-Moore computing systems that combine advanced architectures with new materials and cognitive design approaches.

Keywords: VLSI design, FinFET, 3D integration, machine learning for EDA, post-Moore computing.

1. Introduction

Foundational theoretical studies in complementary metal–oxide–semiconductor (CMOS) technology established that transistor size reduction is the primary driver for enhanced performance and reduced power consumption. For decades, innovations such as strained silicon channels, high-k/metal gate integration, and Fin Field-Effect Transistors (FinFETs) enabled the continuation of Moore's Law well beyond the 22-nanometer technology node. However, as traditional planar scaling approached its physical boundaries—manifested by increased power density and significant gate leakage at sub-10 nanometer nodes—the industry was forced to pivot.

To address these limitations, researchers shifted focus toward novel device structures and integration methodologies. The self-aligned double-gate FinFET structure emerged to provide enhanced short-channel control, while 3D CMOS-over-CMOS stacking technology demonstrated that vertical integration could mitigate interconnect delays and facilitate heterogeneous logic-memory systems. Furthermore, the exploration of carbon-nanotube field-effect transistors (CNTFETs) has shown promise for superior energy performance in the post-silicon era.

Concurrent with these physical advancements, the domain of Electronic Design Automation (EDA) has undergone a paradigm shift. As physical scaling reached diminishing returns, design intelligence became a critical frontier. EDA has evolved into a data-driven discipline, leveraging machine learning (ML) techniques—such as deep learning, graph neural networks (GNNs), and reinforcement learning (RL)—to surpass human-made heuristics in layout quality and placement algorithms. This transition marks the emergence of Machine Learning for CAD (MLCAD). However, the integration of data-driven methods introduces new challenges regarding model stability and vulnerability to adversarial attacks, raising concerns about security and reliability in design flows.

Research Scope and Objectives Despite extensive literature on individual technologies, there is a need to synthesize how these disparate elements converge into a unified evolutionary path. This paper presents a systematic review of VLSI evolution, categorizing developments into three strategic phases: (1) physical scaling limits, (2) architectural diversification (specifically 3D and neuromorphic

computing), and (3) AI-assisted design automation. Methodologically, this review analyzes key literature ranging from device physics to algorithmic automation to demonstrate the industry's transition from "geometric scaling" to "intelligence-driven scaling."

The objective is to provide a comprehensive outlook on how uniting physical material advancements with cognitive automation systems will define the future of three-dimensional heterogeneous integration and post-Moore computing.

2. Historical Progression

The theoretical progress shows a shift from smaller transistor sizes to increased computational power through algorithmic intelligence which defeats physical barriers. The union of Machine Learning with neuromorphic systems and advanced device engineering establishes a new scientific field where learning methods now drive semiconductor progress instead of conventional size reductions. VLSI researchers need to create universal models which integrate device physics with architecture and design automation through established interdisciplinary principles.

2.1. Technology Scaling and the Limits of CMOS

The CMOS device miniaturization process under Moore's law and Dennard's scaling principles has delivered exponential growth in transistor density and performance and energy efficiency since its inception [1,2]. The traditional scaling approach decreased gate size and oxide thickness and supply voltage at equal rates which allowed power density to stay constant while transistor numbers doubled every two years. The nanometer scale brought physical barriers which caused quantum tunneling and short-channel effects to decrease traditional scaling advantages.

The semiconductor industry developed new device structures and materials to solve the emerging challenges in device manufacturing. The strained silicon transistor at the 90-nm node achieved better carrier mobility through the application of tensile and compressive strain techniques. The high-κ/metal-gate (HKMG) stack became available at the 45-nm node to reduce leakage current while preserving performance levels [2]. The FinFET technology introduced at 22 nm brought revolutionary electrostatic control and steep subthreshold slope performance which enabled Moore's law to continue beyond traditional planar boundaries [3]. The industry now seeks alternative computing approaches because traditional scaling methods no longer deliver efficient energy consumption and affordable manufacturing costs at 5 nm and beyond.

2.2. Beyond Planar Scaling: From FinFETs to Neuromorphic Architectures

The FinFET and gate-all-around (GAA) architectures follow Moore’s law by implementing structural advancements for its continuation [3]. The FinFET structure achieves better channel control through its three-dimensional gate design which enables operation at low voltage levels. The performance of systems becomes limited by variability and heat dissipation and interconnect delay when devices reach atomic size dimensions.

The neuromorphic processing field advances through Loihi chip development from Intel which implements brain-inspired computing through spiking neural networks (SNNs). The Loihi chip contains 128 neuromorphic processing units which allow users to program synaptic learning rules for executing parallel computations at reduced energy consumption compared to traditional von Neumann systems [10]. Theoretical models for neuromorphic computing use dynamical systems and local plasticity models to describe how neurons change state through time while learning occurs through weight adjustments in specific areas. The transition to architecture-level innovation demonstrates that physical scaling has its boundaries while requiring designers to unite algorithms with hardware components.

2.3. Machine Learning for Design Automation: Theoretical Underpinnings

The exponential growth of integrated circuit (IC) design complexity has become unavoidable because transistor scaling has reached its maximum point. The optimization of power-performance-area (PPA) tradeoffs and exploration of extensive design spaces remains difficult for traditional heuristic computer-aided design (CAD) methods. Machine learning (ML) uses data-based methods to automate parameter tuning through self-learning models which extract knowledge from past design experiences [6] [8].

Research into electronic design automation (EDA) machine learning applications identifies three essential methods for theoretical development:

1. Prediction Models: Supervised learning develops basic design models which replace costly simulation operations to estimate timing behavior and congestion and yield performance.

2. Decision Models: Reinforcement learning (RL) converts design optimization into a sequential decision system which allows agents to find best placement and routing actions through reward optimization [7].

3. Generative Models: Generative adversarial networks (GANs) and variational autoencoders (VAEs) generate new design examples which increase training data size and produce mask layouts that follow physical rules. Neural-network approaches have become the leading choice for IC design optimization through the implementation of graph neural networks (GNNs) and convolutional models for layout image processing [8].

The models demonstrate perfect compatibility with IC graph structures which results in both high performance and extensive design capabilities.

2.4. Reinforcement Learning in Physical Design

The core theoretical framework for design-space exploration emerged from reinforcement learning among various machine learning paradigms. The chip floor planning problem receives MDP formulation which uses states to store placement data and actions to place components and rewards to evaluate layout quality. Deep reinforcement learning agents trained on multiple chip designs achieve better placement results than human experts when they encounter new layouts according to Amirhossein et al. [7]. The research presents three essential theoretical advancements through its work. The research establishes chip floor planning as a complex contextual bandit problem which needs sequential decision-making to solve its high-dimensional structure. The research presents Edge-GNN architectures which transform circuit graph connections into useful state information for decision-making. The PPO algorithm enables stable gradient updates for large continuous action spaces through its implementation. The framework represents a fundamental change from rule-based optimization to experience-based policy development which enables autonomous VLSI design through intelligent decision-making.

2.5. Reliability and Security in ML for EDA

The development of EDA systems has advanced quickly but researchers continue to face difficulties when it comes to making models both reliable and secure. Xie et al. identify four main vulnerability areas which include data privacy breaches and model theft and adversarial attacks and fundamental system instability [9]. The implementation of ML in EDA creates additional security threats because models require access to confidential circuit information and they must handle diverse technology nodes and distributed training processes.

Theoretical EDA ML robustness requires three essential elements: developing formal methods to establish generalization limits for diverse data sets and implementing adversarial training for physical-design models and creating secure systems for distributed model training. The development of future EDA systems demands both precise predictions and proven system reliability and clear explanation capabilities.

3. Applications of Machine Learning and Emerging Architectures

3.1. Device-Level Scaling and Structural Innovation

Modern VLSI technology depends on the ongoing reduction of CMOS transistor sizes through complementary metal–oxide–semiconductor (CMOS) technology. Borkar established that deep-submicron scaling faces three major challenges which include power consumption and variability and interconnectivity problems because feature size reduction leads to increased leakage and longer delays [1]. Bohr and Young analyzed CMOS scaling patterns in detail to prove that Dennard scaling failed to work beyond the 130-nm technology node [2]. The authors demonstrated that process advancements through strained silicon and high-κ/metal gate (HKMG) and FinFET technologies became essential to sustain Moore's law progress.

Hisamoto et al. developed the FinFET which operates as a self-aligned double-gate MOSFET that scales down to 20 nm while transitioning from planar to three-dimensional device structures [3]. The new device structure delivered enhanced electrostatic control and minimized short-channel effects while creating a structural framework to sustain Moore's law when traditional scaling became impossible. The combination of through-silicon via (TSV) and monolithic inter-tier via (MIV) technologies enables vertical Moore's law extension according to Fenouillet-Beranger et al. [4]. The transmission electron microscopy (TEM) cross-section image in [4] Figure 1 demonstrates how two transistor layers from different tiers can be precisely aligned for monolithic integration.

The research by Karthik et al. evaluated power-efficient design approaches for CMOS and carbon-nanotube FET (CNTFET) circuits [5]. The research demonstrated that nanocarbon devices achieve better leakage reduction and switching energy efficiency which indicates a transition toward energy-focused design approaches. The development of new devices and structural designs created a base for post-Dennard scaling which demonstrates that future advancements need device physics optimization alongside circuit technique development and system architecture enhancement.

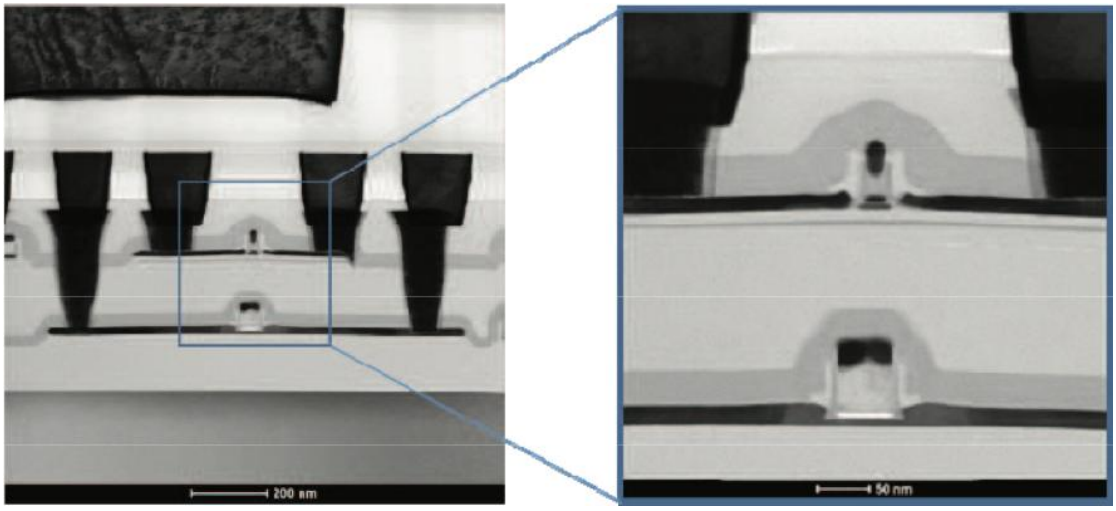

Figure 1. TEM cross-section of the 3D sequential structure up to M2 line. Nanometric top and bottom transistors alignment is observed. (Data from: [4])

3.2. Architectural Intelligence and Neuromorphic Computing

The advancement of computing efficiency depends on architectural innovation because conventional CMOS technology has reached its atomic size limits. The performance of AI and high-performance workloads suffers from three major limitations which stem from traditional von Neumann architectures: memory bandwidth constraints and synchronization delays and data transfer power consumption. The development of hardware-based learning systems has led to the creation of architectural intelligence as a solution for these performance barriers.

The Loihi neuromorphic manycore processor from Intel's 14-nm process represents a well-developed example of this concept which Davies et al. introduced in their work [10]. The Loihi chip

contains 128 spiking cores which connect through an asynchronous network-on-chip according to Figure 2 of [10]. The neuromorphic core of Loihi contains multiple neuron and synapse compartments which perform local event-driven computation and the on-chip microcode enables users to define programmable learning rules and plasticity dynamics. The asynchronous communication system reduces clocking expenses while enabling extensive parallel message transmission which allows neurons to activate only during event occurrences that match biological neural activity patterns.

The architectural benefits of this paradigm receive support from quantitative assessment results. The pre-silicon performance and energy consumption data from Table 1 of [10] demonstrate that Loihi achieves two to three orders of magnitude better energy efficiency than digital processors for executing machine-learning operations. The physical integration of computation and memory substrates leads to substantial reductions in data transfer energy consumption. The Loihi system performs online learning through local synaptic updates which eliminates the need for external retraining because it supports dynamic weight matrix modifications. The design of Loihi implements essential principles from temporal coding and local plasticity theories. The precise timing of spikes in temporal coding provides a more extensive information capacity than amplitude-based methods while spike-timing-dependent plasticity (STDP) plasticity rules enable on-chip learning through pre- and post-synaptic activity causal relationships.

The system demonstrates that intelligence and learning can develop from hardware components without requiring additional energy-consuming data movement between processing and memory units. Neuromorphic computing establishes a new fundamental approach which replaces transistor size reduction with performance enhancement through adaptive learning systems. The approach uses bio-inspired designs with adaptive learning systems to achieve exponential improvements in energy efficiency and computational density instead of pursuing smaller transistors. The evolution of Moore's law transforms from a geometric measurement to an intelligence-based assessment which evaluates VLSI progress through real-time learning and adaptation capabilities.

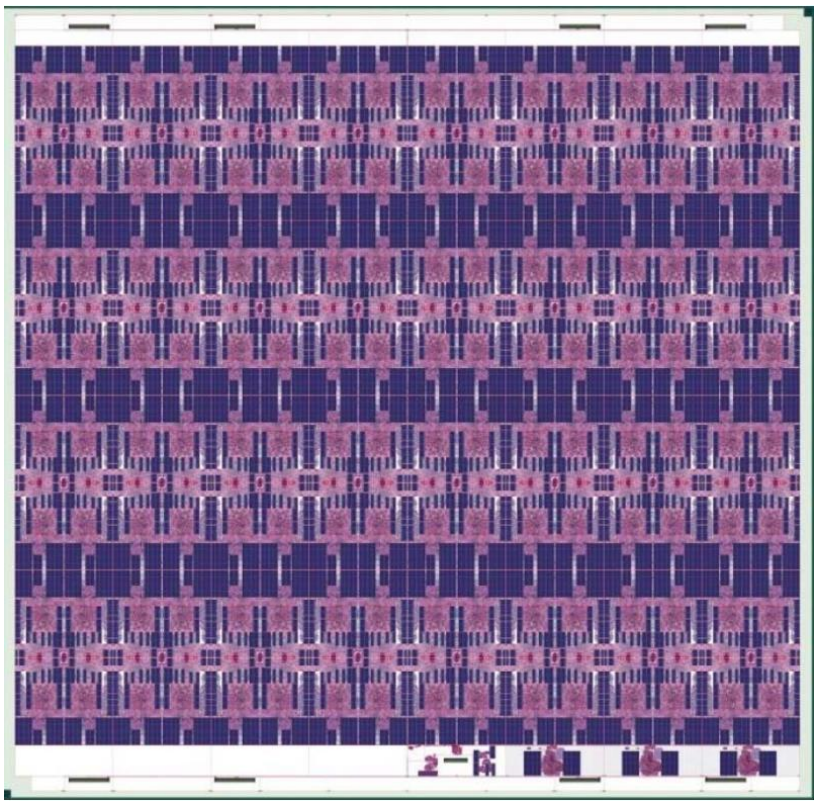

Figure 2. Loihi chip plot. (Data from: [10])

Table 1. Loihi pre-silicon performance and energy measurements.

Measured Parameter	Value at 0.75 V
Cross-sectional spike bandwidth per tile	3.44 Gspike/s
Within-tile spike energ	1.7 pJ
Within-tile spike latency	2.1 ns
Energy per tile hop (E-W / N-S)	3.0 pJ / 4.0 pJ
Latency per tile hop (E-W / N-S)	4.1 ns / 6.5 ns
Energy per synaptic spike op (min)	23.6 pJ
Time per synaptic spike op (max)	3.5 ns
Energy per synaptic update (pairwise STDP)	120 pJ
Time per synaptic update (pairwise STDP)	6.1 ns
Energy per neuron update (active / inactive)	81 pJ / 52 pJ
Time per neuron update (active / inactive)	8.4 ns / 5.3 ns
Mesh-wide barrier sync time (1-32 tiles)	113-465 ns

3.3. Machine Learning–Driven Electronic Design Automation

The exponential increase in design complexity makes traditional heuristic methods in computer-aided design (CAD) unable to handle extensive design exploration and multi-objective optimization.

Hasan et al. and Rapp et al. conducted extensive reviews of machine learning (ML) applications in electronic design automation (EDA), which organized ML techniques into synthesis, placement and routing, and verification categories [6][8]. The three main design paradigms that improve design scalability and productivity, according to their research, include deep learning, graph neural networks (GNNs), and reinforcement learning (RL).

Design-space exploration, from a foundational perspective, uses reinforcement learning as its core theory. Amirhossein et al. developed a Markov decision process (MDP) framework for chip floor planning, which they trained using proximal policy optimization (PPO) [7]. The Edge-GNN encoder established topological connections between circuit elements, producing manufacturable floorplans from heterogeneous designs within short time periods. The research in [7] shows that RL-based design frameworks scale better with larger pre-training datasets, achieving higher placement quality and faster convergence (Figure 3).

Xie et al. demonstrated that ML applications in EDA face two essential problems regarding model dependability and security, including data privacy breaches, model theft, adversarial attacks, and poor design transferability [9]. The authors suggested two solutions—federated learning and adversarial robustness training—since ML success in EDA requires both high accuracy and dependable, reproducible results. These research findings demonstrate that ML serves as a fundamental principle for developing future design automation systems that will evolve EDA into adaptive, data-driven, autonomous systems.

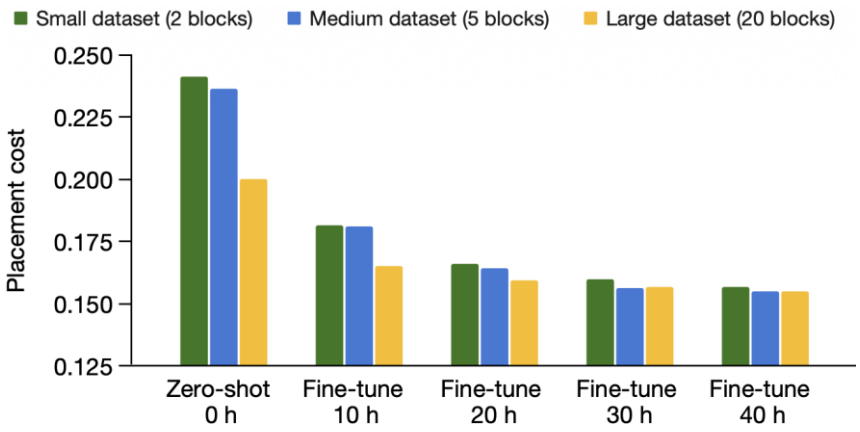

Figure 3. Effect of pre-training dataset size. (Data from: [7])

4. Future Outlook

The development of VLSI design will depend on the ongoing combination of physical advancements with cognitive automation systems. The development of Beyond-FinFET devices and gate-all-around architectures needs to be optimized with back-end interconnects and power delivery networks to solve short-channel effects and power constraints that previous scaling research identified [1] [2] [3]. The transition of monolithic 3D integration from laboratory proof-of-concept to design-kit readiness will need fundamental advancements in thermal control systems and yield prediction models and CAD optimization methods [4]. The development of carbon-nanotube field-effect transistors (CNTFETs) for energy-centric scaling will lead to better heterogeneous CMOS–CNTFET integration and precise cross-technology power–performance–area (PPA) modeling [5]. The development of neuromorphic computing systems needs to move from experimental prototypes to functional systems that serve specific workloads. The development of future chips based on Loihi's event-driven and plasticity-enabled principles will include mixed-signal interfaces and enhanced on-chip learning rules and compiler tools that convert temporal coding into practical applications for edge inference and continual learning [10]. The ML4EDA community needs to shift from individual case studies to develop open and reproducible benchmarks for EDA applications [6][8]. The development of robust transferable learning models requires researchers to create shared datasets and establish standardized metrics and cross-node generalization frameworks. Physical design optimization through reinforcement learning requires better theoretical foundations and improved policy abstraction methods and hybrid approaches with traditional optimization techniques [7]. The development of trustworthy automation systems requires immediate attention because security and reliability issues including model theft and adversarial attacks and poor generalization need federated training systems with formal verification for timing and routing and congestion predictors [9]. The success of end-to-end intelligence-driven co-design depends on the simultaneous learning of device physics with 3D partitioning and architecture templates and EDA optimization policies under physical and manufacturing restrictions. The new paradigm of scaling uses learning capacity as its measurement instead of traditional geometric growth which will lead to semiconductor advancement through hardware-algorithm collaboration [1]–[10].

5. Conclusion

The research examines IC technology development through physical scaling to intelligent automation by studying three connected aspects: device-level advancements, architecture-level intelligence, and machine learning-based design approaches. It investigates IC scaling progress from CMOS miniaturization to 3D stacking, neuromorphic processing, and AI-based EDA tools to show that modern scaling theory now includes both computational and cognitive dimensions.

The development of integrated circuit technology shows that each scaling period advances through physical breakthroughs and computational system improvements. The first period of CMOS scaling was characterized by transistor miniaturization, driven by the work of Borkar and Bohr and Young, who extended Moore’s and Dennard’s principles. However, as dimensions reached the nanometer scale, quantum tunneling, leakage, and interconnect delays emerged, making classical scaling models increasingly ineffective.

The physical barriers of device scaling led to the development of FinFETs and strained silicon and high-κ/metal-gate transistors which achieved scaling through material and geometric enhancements. The development of three-dimensional integration and low-power circuit design methods brought new structural and energy-based scaling methods which worked alongside traditional physical scaling techniques. The introduction of architecture-level intelligence brought about a new technological period. The Loihi neuromorphic processor proved that learning and parallel processing methods can achieve energy efficiency without requiring smaller transistors thus creating a brain-inspired computing system that connects device physics to algorithmic design. The rising complexity of modern VLSI systems demands optimization frameworks which use data-driven and learning-based

methods for optimization. The implementation of artificial intelligence through reinforcement learning for chip floor planning and machine learning-driven EDA surveys demonstrates how to replace conventional heuristic design methods with intelligent self-optimizing systems. The industry transition demonstrates that physical scaling has transformed into computational and cognitive scaling which depends on learning as its fundamental progress mechanism. The advancement of semiconductor technology depends on two fundamental elements which include transistor development and intelligent algorithm implementation for learning from large design spaces and performance prediction and process node adaptation. The semiconductor industry now operates under an "intelligence-driven scaling" model which unites physics with architecture and machine learning to establish new development standards.

Despite these accplishments, the development of AI and data center operations still faces multiple essential obstacles. The main obstacle to support big AI operations and data center workloads stems from energy efficiency problems. Xie et al. emphasize that ML models used for design reliability and security need to generate results which are both verifiable and interpretable. The hardware–algorithm co-design process needs to develop into single frameworks which unite design targets with physical boundaries and learning-based objectives. A unified framework which connects device physics to architectural abstraction and machine intelligence will establish a new VLSI development phase where learning functions as the core force behind Moore's law advancement in the intelligent era.

References

[1] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23 – 29, Jul.–Aug. 1999.

[2] M. T. Bohr and I. A. Young, "CMOS scaling trends and beyond," *IEEE Micro*, vol. 37, no. 6, pp. 20 – 29, Nov.–Dec. 2017.

[3] D. Hisamoto *et al.*, "FinFET—a self-aligned double-gate MOSFET scalable to 20 nm," *IEEE Transactions on Electron Devices*, vol. 47, no. 12, pp. 2320 – 2325, Dec. 2000.

[4] C. Fenouillet-Beranger *et al.*, "Recent advances in 3D VLSI integration," *IEEE Transactions on Electron Devices*, vol. 67, no. 12, pp. 5475 – 5483, Dec. 2020.

[5] K. A. Karthik, S. M. Akshay, and A. S. Roy, "An overview of low-power VLSI design methods for CMOS and CNTFET-based circuits," *IEEE Access*, vol. 10, pp. 83597 – 83612, Aug. 2022.

[6] M. Z. Hasan, M. Li, A. K. Mishra, and T. Srikanthan, "Machine learning for electronic design automation: A survey," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 26, no. 5, pp. 1 – 46, Aug. 2021.

[7] A. Amirhossein *et al.*, "A graph placement methodology for fast chip design," *Nature*, vol. 594, pp. 207 – 212, Jun. 2021.

[8] M. Rapp *et al.*, "MLCAD: A survey of research in machine learning for CAD," *IEEE* Transactions *on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 41, no. 10, pp. 3162 – 3179, Oct. 2022.

[9] Z. Xie, J. Pan, C. Chang, J. Hu, and Y. Chen, "The dark side: Security and reliability concerns in machine learning for EDA," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 42, no. 4, pp. 1171 – 1186, Apr. 2023.

[10] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82 – 99, Jan.–Feb. 2018.