International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(9s) 2026 pp. 1130-1144
https://doi.org/10.70917/ijcisim-2026-3532
Article

Pareto-Optimal Power–Delay Co-Optimization of Advanced CMOS Technologies for Next-Generation VLSI Systems Using NSGA-II

Karishma Parween[1*], Deeba Ashique[2], Amol Jagdish Shakadwipi[3]

[1]Department of ECE, Government Engineering College Kishanganj, Bihar, India, Email: kparween67@gmail.com
[2]Department of ECE Government Engineering College (GEC) Banka, Bihar, India, Email: deeba0891@gmail.com
[3]SNJB's K.B Jain college of Engineering, Chandwad, amolshakadwipi@gmail.com
Corresponding Author:kparween67@gmail.com

Abstract: Power dissipation and propagation delay are antagonistic design objectives in deeply scaled very-large-scale integration (VLSI): voltage and device downsizing reduce switching energy but weaken drive current, whereas aggressive sizing and low-threshold operation improve speed at the expense of dynamic, short-circuit, and leakage power. This study develops a reproducible Non-dominated Sorting Genetic Algorithm II (NSGA-II) framework for simultaneous power–delay optimization of four representative CMOS technology classes: 45 nm planar CMOS, 22 nm FinFET, 12 nm FinFET, and 5 nm stacked-nanosheet gate-all-around FET (GAAFET). A technology-scaled fan-out-of-four inverter benchmark is parameterized by supply voltage, effective drive size, p/n sizing ratio, threshold voltage, and gate-length bias. The two objectives are to minimize average power of a 1,000-stage switching bank and minimize FO4 propagation delay subject to area, transition-balance, overdrive, and leakage-share constraints. Main searches use 100 individuals for 140 generations; ten independent equal-budget comparisons assess NSGA-II against a weighted-sum genetic algorithm and random search. NSGA-II generated well-distributed fronts and achieved the best average hypervolume (1.16945) and inverted generational distance (0.01916). Knee solutions reduced power and delay from 2.366 mW and 23.041 ps at 45 nm to 0.226 mW and 2.541 ps at 5 nm. Two-thousand-sample process–voltage–temperature analysis confirmed that the selected compromises remain stable, although leakage sensitivity grows in scaled FinFET and GAAFET designs. The framework supplies designers with a set of implementable alternatives rather than a single preference-dependent optimum and is suitable for early technology–circuit co-design before proprietary process-design-kit sign-off.

Keywords: NSGA-II; multi-objective optimization; CMOS; FinFET; gate-all-around FET; low-power VLSI; propagation delay; Pareto front; transistor sizing

1. Introduction

The continuing growth of artificial intelligence accelerators, edge processors, high-performance computing platforms, and always-on connected devices places contradictory demands on digital integrated circuits. A high clock rate requires strong transistor drive, short critical paths, and sufficient voltage headroom, while a limited thermal and battery envelope requires low switching energy and tightly controlled standby current (Neelam et al., 2024). Historically, constant-field scaling promised simultaneous improvements in density, speed, and energy (Valasa et al., 2022). That simple relationship weakened as supply voltage ceased to scale proportionally with device dimensions, interconnect and contact parasitics became prominent, and short-channel leakage increased. The 2024 International Roadmap for Devices and Systems (IRDS) consequently treats power, performance, area, and cost as interdependent technology targets rather than independent improvements (Wang, 2024).

At circuit level, total CMOS power is conventionally separated into dynamic, leakage, and short-circuit components. Dynamic power grows approximately with the square of supply voltage, so voltage reduction is the most effective first-order energy lever (Chen, 2022). The same reduction decreases gate overdrive and drive current,

however, which increases propagation delay. Lower threshold voltage can recover performance but raises subthreshold leakage exponentially (Sakurai & Newton, 1990). Transistor upsizing lowers effective resistance and improves slew, but increases switched and internal capacitance. These interactions mean that the design with minimum power is almost never the design with minimum delay. A power–delay product can collapse the two criteria into one number, but its implied preference is fixed and it can hide alternatives that are superior for particular thermal, timing, or workload constraints.

Device architecture further complicates the problem. Planar bulk MOSFETs offer continuous width control but lose electrostatic integrity at deeply scaled gate lengths. FinFETs use a thin vertical fin and multiple gate surfaces to improve channel control (Mukesh & Zhang, 2022), yet their quantized fin count makes sizing discrete. Horizontally stacked nanosheet GAAFETs extend gate control around the channel and make sheet width another drive-current variable (Das et al., 2024; Karimi et al., 2024; Yoon & Baek, 2020). These changes alter effective capacitance, p/n balance, leakage response, and the sensitivity of delay to voltage and geometry. Consequently, a sizing rule or scalar objective tuned for planar CMOS does not transfer directly to FinFET or GAAFET standard-cell design.

Conventional gate sizing methods have included sensitivity-based heuristics, nonlinear programming, convex reformulations, and single-objective genetic algorithms (Lee & Yoo, 2021; Vişan et al., 2022; Q. Zhang et al., 2024). They remain valuable when the objective is differentiable and a precise timing or power constraint is known. Early technology exploration is different: compact-model responses may be nonlinear or discontinuous, discrete drive units appear in nonplanar devices, and the preferred trade-off is not yet fixed. Multi-objective evolutionary algorithms are attractive in this setting because they can search mixed and nonconvex design spaces without differentiating the circuit evaluator. NSGA-II is especially suitable for two-objective engineering problems because it combines elitist nondominated sorting with crowding-distance preservation (Lberni et al., 2023; Sanabria-Borbón et al., 2020). A single run returns a diverse approximation to the Pareto front, allowing the final design choice to be made after the physical trade-off is visible.

Although evolutionary techniques have been applied to transistor sizing and analog circuit synthesis, three limitations recur in power–delay studies. First, many reports aggregate power and delay into a weighted sum or power–delay product, thereby reporting only one preference-conditioned solution. Second, studies commonly evaluate one device node, making it difficult to determine whether the optimization behavior is preserved during the planar-to-FinFET-to-GAAFET transition. Third, nominal optimization is often separated from process–voltage–temperature (PVT) analysis, so an apparently attractive Pareto point may be fragile. A further reproducibility problem arises when proprietary foundry models are used without disclosing a transferable parameterization.

This paper addresses those limitations through a transparent technology-scaled benchmark. The physical equations are deliberately compact enough to reproduce, while their parameter trends follow established low-power theory, predictive technology modeling, multi-gate compact modeling, and the IRDS roadmap (Nguyen & Hoang, 2023; Valencia-Ponce et al., 2021; Xu et al., 2024). They are not presented as a replacement for a foundry process-design kit (PDK); rather, they provide a defensible early-design layer that can later be replaced by SPICE measurements without changing the optimizer.

The study makes four contributions. First, it formulates simultaneous minimization of power dissipation and propagation delay with explicit feasibility constraints and technology-dependent sizing rules. Second, it produces and compares Pareto fronts for representative 45 nm planar, 22 nm FinFET, 12 nm FinFET, and 5 nm GAAFET classes using the same logical-effort benchmark. Third, it evaluates convergence and front quality over independent, equal-budget runs against scalarized and random baselines. Fourth, it selects curvature-based knee points and subjects them to a 2,000-sample PVT robustness study. The resulting design set is intended to support next-generation VLSI decisions in which thermal limits, workload latency, cell area, and manufacturing variation must be negotiated rather than reduced to a single universal score.

2. Literature Review

2.1 Physical basis of the power–delay trade-off

The quadratic voltage term in switching power has made voltage scaling central to low-energy CMOS design [3]. For an activity factor α, operating frequency f, and effective switched capacitance Ceff, the first-order relation $P_{dyn} = \alpha C_{eff} V_{DD}^2 f$ explains why modest voltage reductions can deliver large energy savings. In a real cell, Ceff contains fan-out, diffusion, gate, and local interconnect capacitances, each of which changes with device geometry. Upsizing therefore produces two competing effects: higher current shortens the output transition, but additional

capacitance raises both switching energy and the charge that must be moved. Short-circuit power also depends on input slew and the interval during which the pull-up and pull-down networks conduct simultaneously.

Propagation delay is frequently interpreted through an RC or charge-current relation. The alpha-power-law model of Sakurai and Newton captures velocity-saturation effects more realistically than a square-law MOSFET model and links delay to supply voltage, threshold voltage, capacitance, and drive strength (Yin et al., 2022). Logical-effort and standard digital-CMOS analyses further show that the fan-out-of-four (FO4) inverter delay is a useful normalized technology and circuit benchmark because it embeds an electrically meaningful load while remaining simple to reproduce (Blank & Deb, 2020; Kim et al., 2024). The FO4 metric does not represent every critical path, but it is more informative than unloaded intrinsic delay and avoids architecture-specific netlists during an early device comparison.

Leakage becomes progressively important as channel length and threshold voltage are reduced. Subthreshold conduction, gate tunnelling, junction leakage, and drain-induced barrier lowering contribute in different proportions across technology generations (Kumari & Prithvi, 2023). At high activity, dynamic power usually dominates; under low activity or elevated temperature, leakage can determine the feasible design region. A power-only optimizer may therefore select high threshold and length bias, while a delay-only optimizer tends toward high voltage, large drive units, and low threshold. The intervening nondominated set contains designs for different duty cycles and timing margins.

2.2 Evolution of advanced CMOS devices

Dennard scaling established the conceptual basis for shrinking MOS devices under approximately constant electric field (X. Zhang et al., 2025). As planar gates lost control of the channel, multi-gate geometries emerged. The self-aligned double-gate FinFET demonstrated strong short-channel control through a narrow silicon fin (Dubey et al., 2025). Production-class tri-gate technology subsequently showed that low-power and high-performance devices could share a scaled platform (Rathore et al., 2023). In a FinFET cell, effective width is proportional to fin count and fin perimeter, so width is quantized. This can limit exact p/n balancing and produces step changes in power, delay, and area.

GAAFETs continue the electrostatic evolution by surrounding the channel. Stacked horizontal nanosheets demonstrated a practical route beyond FinFET scaling, with sheet width providing more flexible effective drive than a fixed fin height [10]. Vertically stacked nanowire and nanosheet experiments confirmed that replacement-metal-gate integration can provide strong electrostatic control on bulk substrates (Zitzler et al., 2003). BSIM-CMG extends surface-potential-based compact modeling across common multi-gate devices and incorporates effects needed for FinFET and GAAFET circuit simulation (Deb & Jain, 2014). At roadmap level, these structures support continued performance and density scaling, but local interconnect, contact resistance, variability, self-heating, and leakage remain system constraints (Deb et al., 2002; Yin et al., 2022).

Comparing nominal node labels without controlling the logical workload can be misleading because a node name is not a single physical dimension and each technology uses different capacitance and drive conventions. A useful study therefore needs either calibrated PDKs for every node or a transparent normalized benchmark. The present work adopts the latter and labels each device set a representative “technology class.” It preserves the main trends—supply range, capacitance, drive current, leakage coefficient, and sizing quantization—while avoiding claims of foundry-specific silicon accuracy.

2.3 Transistor sizing and circuit optimization

Transistor sizing has a long optimization history. Shyu et al. combined heuristics and mathematical programming to improve circuit performance (Kim et al., 2024). Sapatnekar et al. showed that a class of CMOS sizing problems could be solved exactly through convex optimization under appropriate delay models (X. Zhang et al., 2025). Borah et al. directly minimized power under a delay constraint (Nguyen & Hoang, 2023). These methods are efficient when the response surface and constraints fit the assumed mathematical form, but their single-criterion formulation must be rerun for each timing bound. Modern cell libraries also contain discrete sizes and nonconvex effects that weaken smooth convex assumptions.

Population-based metaheuristics trade mathematical guarantees for flexibility. Genetic algorithms encode voltage, device geometry, and threshold choices and can accommodate simulator failures or categorical device options (Lberni et al., 2023). Singh and Kaur reported automatic transistor sizing aimed at low power and performance specifications, illustrating the continuing relevance of heuristic search (Chen, 2022). Nevertheless, a single-objective genetic algorithm still requires a power–delay weight or a hard bound. Different weights can sample different

compromises, but a linear weighted sum may fail to recover points on a nonconvex front and typically clusters solutions near the extremes.

2.4 Multi-objective evolutionary optimization

In multi-objective optimization, a design is Pareto optimal when no objective can be improved without worsening at least one other objective (Valasa et al., 2022). NSGA-II ranks candidates by nondomination level, retains elite solutions, and uses crowding distance to prevent the population from collapsing into a narrow front region (Yin et al., 2022). Its $O(MN^2)$ sorting cost is practical for two objectives and moderate populations, where M is the number of objectives and N is population size. Simulated binary crossover and polynomial mutation supply real-coded exploration while allowing rounding or repair of discrete sizing variables (Yin et al., 2022; X. Zhang et al., 2025). Established software frameworks facilitate reproducible implementations (Blank & Deb, 2020), while reference-point extensions such as NSGA-III support future expansion to many-objective VLSI problems (Nguyen & Hoang, 2023).

Quality assessment must consider both convergence and diversity. Hypervolume measures the objective-space region dominated by an approximation relative to a reference point; larger values imply a better combination of convergence and coverage (Sanabria-Borbón et al., 2020; Vişan et al., 2022). Inverted generational distance (IGD) measures the average distance from a reference front to the approximation; smaller values are preferred. Spacing indicates distribution uniformity. These metrics are complementary: a method can find excellent extreme points yet leave large gaps, or distribute points evenly along a front that remains far from the best known set. Evolutionary multi-objective research therefore recommends reporting multiple indicators and independent runs rather than one visual front (Kim et al., 2024; Lberni et al., 2023; Nguyen & Hoang, 2023).

2.5 Research gap and proposed direction

The literature establishes the physical power–delay conflict, the architectural progression to multi-gate devices, and the value of Pareto-based optimization. What remains underdeveloped is an integrated, reproducible comparison that: (i) keeps power and delay separate; (ii) applies the same constrained search protocol across planar, FinFET, and GAAFET classes; (iii) reports algorithmic quality against equal-budget baselines; and (iv) connects nominal knee selection to PVT robustness. Figure 1 summarizes the proposed response to this gap. Technology parameters feed a common FO4 evaluator, NSGA-II constructs a feasible Pareto archive, a geometric knee rule identifies a balanced design, and Monte Carlo perturbations test whether that choice remains credible under variation.

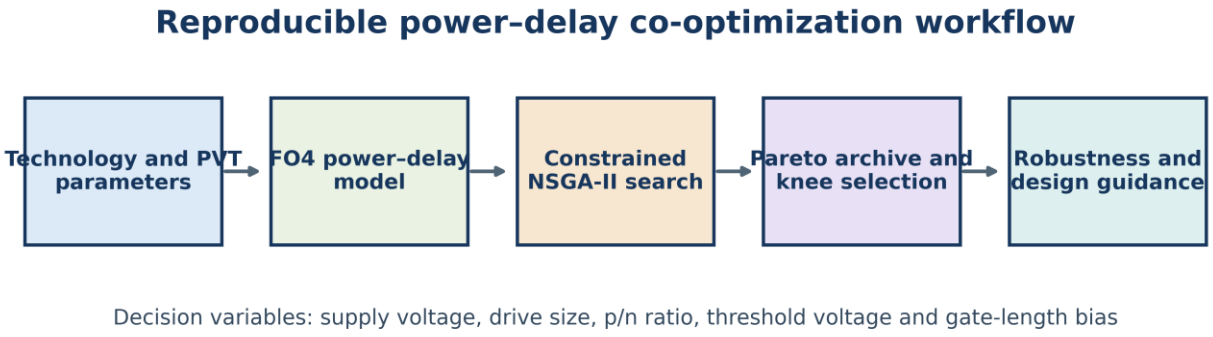

Figure 1. Proposed workflow for technology parameterization, constrained NSGA-II search, Pareto analysis, knee selection, and PVT robustness assessment.

3. Research Methodology

3.1 Research design and benchmark

The research is a computational, simulation-based technology–circuit co-optimization study. The benchmark is an equivalent bank of 1,000 identical CMOS inverter stages, each driving an FO4-equivalent load at 2 GHz with activity factor 0.10. Reported power is average milliwatts for the 1,000-stage bank; reported delay is the nominal FO4 propagation delay of one stage in picoseconds. This convention supplies readable power values while preserving a standard critical-stage timing measure. It is not intended to predict the total power of a complete processor.

Four representative technology classes were parameterized from established scaling tendencies and compact-model concepts (Kim et al., 2024; Xu et al., 2024; X. Zhang et al., 2025). Table 1 lists the disclosed values. The external load and intrinsic capacitance coefficient decrease with scaling, while the supply and threshold ranges follow low-power operating regimes. The effective drive variable is continuous for planar CMOS and integer-quantized for the nonplanar classes to represent fin or sheet units. Because these are generic exploration models, absolute comparisons should be interpreted as technology-scaled benchmark results rather than measured foundry data.

Table 1. Representative technology-class parameters used in the model-based benchmark.

Class	Device	VDD range (V)	Vth range (V)	Cload (fF)	Cunit (fF)	Ion/unit (µA)	Ioff/unit (µA)
45 nm planar	Planar bulk CMOS	0.65–1.10	0.25–0.42	12.0	0.90	220	0.0030
22 nm FinFET	Tri-gate FinFET	0.55–0.90	0.22–0.38	6.0	0.52	205	0.0040
12 nm FinFET	Scaled FinFET	0.45–0.80	0.18–0.34	3.2	0.34	190	0.0055
5 nm GAAFET	Stacked nanosheet	0.40–0.75	0.16–0.30	1.6	0.22	180	0.0032

Notes: Values are transparent technology-scaled exploration parameters, not proprietary foundry PDK measurements. Ion and Ioff are nominal per normalized n-drive unit at the class reference voltage and threshold.

3.2 Decision vector and objective functions

The circuit-design vector was defined as

$$x = [V_{\mathrm{DD}}, s, r_{p/n}\, V_{\mathrm{th}}\, \Delta L],$$

where $V_{\mathrm{DD}}$is the supply voltage, $s$is the normalized nMOS drive size, $r_{p/n}$is the pMOS-to-nMOS sizing ratio, $V_{\mathrm{th}}$is the threshold voltage, and $\Delta L$is the fractional positive gate-length bias. Accordingly, the effective pMOS size was calculated as $s_p = s\, r_{p/n}$. The technology-specific lower and upper bounds of these variables are provided in Table 2. Different bounds were adopted only where the device class required distinct voltage, threshold-voltage, or maximum-drive ranges. The first objective was the minimization of total power dissipation:

$$P_{\mathrm{total}}(x) = P_{\mathrm{dyn}}(x) + P_{\mathrm{leak}}(x) + P_{\mathrm{sc}}(x), \qquad (1)$$

where the dynamic and leakage components were evaluated as

$$P_{\mathrm{dyn}} = \alpha C_{\mathrm{eff}} V_{\mathrm{DD}}^2 f,\ P_{\mathrm{leak}} = V_{\mathrm{DD}} I_{\mathrm{off}}. \qquad (2)$$

Here, αdenotes the switching activity factor, $f$is the operating frequency, and C_{eff}combines the technology-specific fan-out-of-four external load with an intrinsic capacitance term proportional to $s(1 + r_{p/n})$. A small correction was incorporated to represent the influence of the gate-length bias. The leakage current varies exponentially with V_{th}, V_{DD}, ΔL, and temperature, whereas short-circuit power increases with transistor size and the available voltage overdrive.

The current formulation employed an alpha-power-law dependence on voltage overdrive and incorporated mobility degradation at elevated temperatures. The effective drive current, I_{eff}, was obtained from the harmonic mean of the pull-up and pull-down currents. This formulation penalizes designs in which one switching transition is substantially slower than the other. The nominal propagation delay was calculated as

$$t_{\mathrm{pd}}(x) = 0.69 \times 1000 \frac{C_{\mathrm{eff}} V_{\mathrm{DD}}}{I_{\mathrm{eff}}}, \qquad (3)$$

where $C_{\mathrm{eff}}$is expressed in femtofarads and I_{eff}in microamperes, producing t_{pd}in picoseconds. A small imbalance correction was further applied to account for residual rise–fall asymmetry. These equations retain the dominant relationships established in low-power CMOS and alpha-power-law analyses while remaining computationally efficient for repeated evolutionary evaluations (Kumari & Prithvi, 2023; Nguyen & Hoang, 2023; Rathore et al., 2023). The resulting bi-objective optimization problem was formulated as

$$F(x) = \left[P_{\text{total}}, (x)\, t_{\text{pd}}(x)\right]. \quad (4)$$

The optimization was subject to the following feasibility conditions:

$$\begin{aligned} A(x) &= s\left(1 + r_{p/n}\right)(1 + \Delta L) \le 48, \\ \delta_I(x) &\le 0.16, \\ V_{\text{DD}} - V_{\text{th}} &\ge 0.16\ \text{V}, \\ \frac{P_{\text{leak}}}{P_{\text{total}}} &\le 0.35. \end{aligned} \quad (5)$$

All constraint violations were normalized and summed into a total violation index. Feasible solutions were assigned dominance over infeasible solutions. When two infeasible candidates were compared, preference was given to the candidate with the smaller total normalized violation. This feasibility-first strategy eliminates the need to specify arbitrary penalty coefficients.

Table 2. Decision variables, bounds, and circuit roles.

Variable	Symbol	Range	Type	Primary role
Supply voltage	VDD	Technology-specific; Table 1	Continuous	Power, overdrive, delay
n-drive size	s	1–12/14/16/18	Planar continuous; others integer	Drive, capacitance, area
p/n ratio	rp/n	1.20–2.70	Continuous	Transition balance and capacitance
Threshold voltage	Vth	Technology-specific; Table 1	Continuous	Leakage and overdrive
Gate-length bias	ΔL	0.00–0.15	Continuous	Leakage–drive adjustment

Feasibility constraints: area index ≤ 48; pull-up/pull-down imbalance ≤ 0.16; VDD − Vth ≥ 0.16 V; leakage share ≤ 35%.

3.3 NSGA-II implementation

The initial population was sampled uniformly within the decision bounds and repaired after every variation operation. For FinFET and GAAFET classes, s was rounded to the nearest integer and clipped to the valid range. Each generation applied binary tournament selection based on nondomination rank and crowding distance. Simulated binary crossover used probability 0.90 and distribution index 15. Polynomial mutation used per-variable probability 1/5 and distribution index 20. Parents and offspring were merged, ranked, and truncated to the population size using crowding distance within the last accepted front (Rathore et al., 2023; Yin et al., 2022; X. Zhang et al., 2025).

The main reported search used 100 individuals and 140 generations, corresponding to 14,100 objective evaluations including the initial population. This resolution produced 100 feasible nondominated points for each technology. A deterministic seed was assigned to each main technology run to make the figures and tables reproducible. Table 3 gives the complete computational protocol.

Table 3. NSGA-II settings and evaluation protocol.

Parameter	Value	Purpose
Main population / generations	100 / 140	High-resolution reported fronts
Main evaluations	14,100 per technology	Includes initial population
Repeated population / generations	60 / 60	Ten runs per technology

Repeated evaluation budget	3,660 per method per run	Equal-budget comparison
Selection	Binary tournament	Rank, then crowding distance
Crossover	SBX; pc = 0.90; ηc = 15	Real-coded recombination
Mutation	Polynomial; pm = 0.20; ηm = 20	One expected variable mutation
Constraint handling	Feasibility first	Lower violation among infeasible designs
Knee rule	Maximum distance to extreme-point chord	Balanced compromise

3.4 Pareto quality and baseline comparison

Algorithm performance was assessed in ten independent repetitions for each technology. To keep the repeated experiment tractable and equal-budget, NSGA-II used 60 individuals and 60 generations, or 3,660 evaluations per run. The first baseline was a weighted-sum genetic algorithm executed across eleven weights from pure-delay to pure-power preference. It used the same crossover, mutation, repair, and approximately the same evaluation budget. The second baseline was uniform random search with 3,660 evaluations. All feasible nondominated solutions obtained by the methods were pooled to form a reference approximation for each technology.

Objectives were normalized with respect to the reference approximation before indicator calculation. Hypervolume used reference point (1.10, 1.10), so its theoretical upper rectangle is 1.21 rather than 1.00. IGD was calculated from the pooled nondominated reference to each run. Spacing was the standard deviation of Euclidean distances between adjacent normalized front points. The number of nondominated solutions and feasible-sample rate were also recorded. Reporting mean and standard deviation across repetitions distinguishes stable algorithm behavior from an unusually favorable seed.

3.5 Knee-point selection

A Pareto front gives alternatives but does not by itself select one circuit. A knee point was used as an interpretable, preference-light compromise. For each main front, power and delay were min–max normalized. A line joined the minimum-power and minimum-delay extreme points, and the nondominated solution with maximum perpendicular distance from that line was selected. This maximum-curvature construction identifies the region where a small improvement in either objective begins to require a disproportionately large sacrifice in the other. It does not eliminate designer preferences; rather, it supplies a rational default for subsequent physical design and PVT analysis.

3.6 Process–voltage–temperature robustness analysis

Each knee design was evaluated under 2,000 Monte Carlo samples. Effective on-current used a normal multiplier with 4.5% standard deviation, capacitance used 3.5%, and threshold voltage shift used 12 mV. Leakage used a lognormal multiplier generated from a normal logarithmic standard deviation of 0.18. Temperature was normally distributed with mean 65 °C and standard deviation 18 °C, clipped to 0–125 °C. The delay model included temperature-dependent mobility, while leakage included exponential temperature sensitivity. The supply variable remained at the selected nominal value; therefore the experiment tests device and environmental variation around the design rather than dynamic voltage droop.

These distributions are generic stress assumptions, not a substitute for foundry statistical corners. They were held consistent across technology classes so that relative sensitivity could be compared. Results are reported as mean, standard deviation, and 95th percentile for power and delay. The analysis is especially relevant because scaled knee points selected low supply and low threshold: such designs can be efficient nominally but sensitive to current loss, threshold shift, and temperature (Deb & Jain, 2014; Kim et al., 2024; Rathore et al., 2023).

3.7 Reproducibility and scope

All equations, parameter ranges, algorithm controls, seeds, and statistical assumptions required to regenerate the presented data are disclosed. The evaluator can be replaced by SPICE, BSIM-CMG, or measured-cell lookup tables without changing nondominated sorting, baseline comparison, or knee selection. The present results therefore

demonstrate the optimization methodology and cross-technology trends. Final tape-out decisions still require PDK-calibrated device models, extracted interconnect parasitics, cell-layout quantization, electromigration and reliability checks, and sign-off across official corners.

4. Results and Discussion

4.1 Pareto-front structure

The main NSGA-II runs returned 100 feasible nondominated designs for every technology, indicating that environmental selection preserved a complete trade-off set rather than converging to one preference region. Figure 2 shows the technology-specific fronts and knee points; Table 4 reports their extrema. Each front has a steep low-power branch, a curved transition region, and a shallow high-power branch. Near the minimum-power extreme, small power additions produce large delay reductions because additional voltage or drive size restores overdrive. Beyond the knee, further delay reduction requires increasingly large capacitive and short-circuit cost. This geometry is precisely the condition under which a single power–delay product or weighted sum can conceal useful alternatives.

For 45 nm planar CMOS, minimum power was 1.235 mW but the associated delay was 144.987 ps. Minimum delay fell to 9.069 ps only when power increased to 14.564 mW. The 22 nm FinFET range was 0.449–6.473 mW and 60.544–3.822 ps. The 12 nm FinFET range was 0.169–4.159 mW and 28.916–2.126 ps. The 5 nm GAAFET front reached a minimum power of 0.071 mW and minimum delay of 1.234 ps at different extremes. These numerical ranges should not be interpreted as interchangeable product specifications because the technology-scaled capacitance parameters differ. They show how the feasible objective envelope moves when the same logical benchmark and optimization rules are applied to progressively scaled device classes.

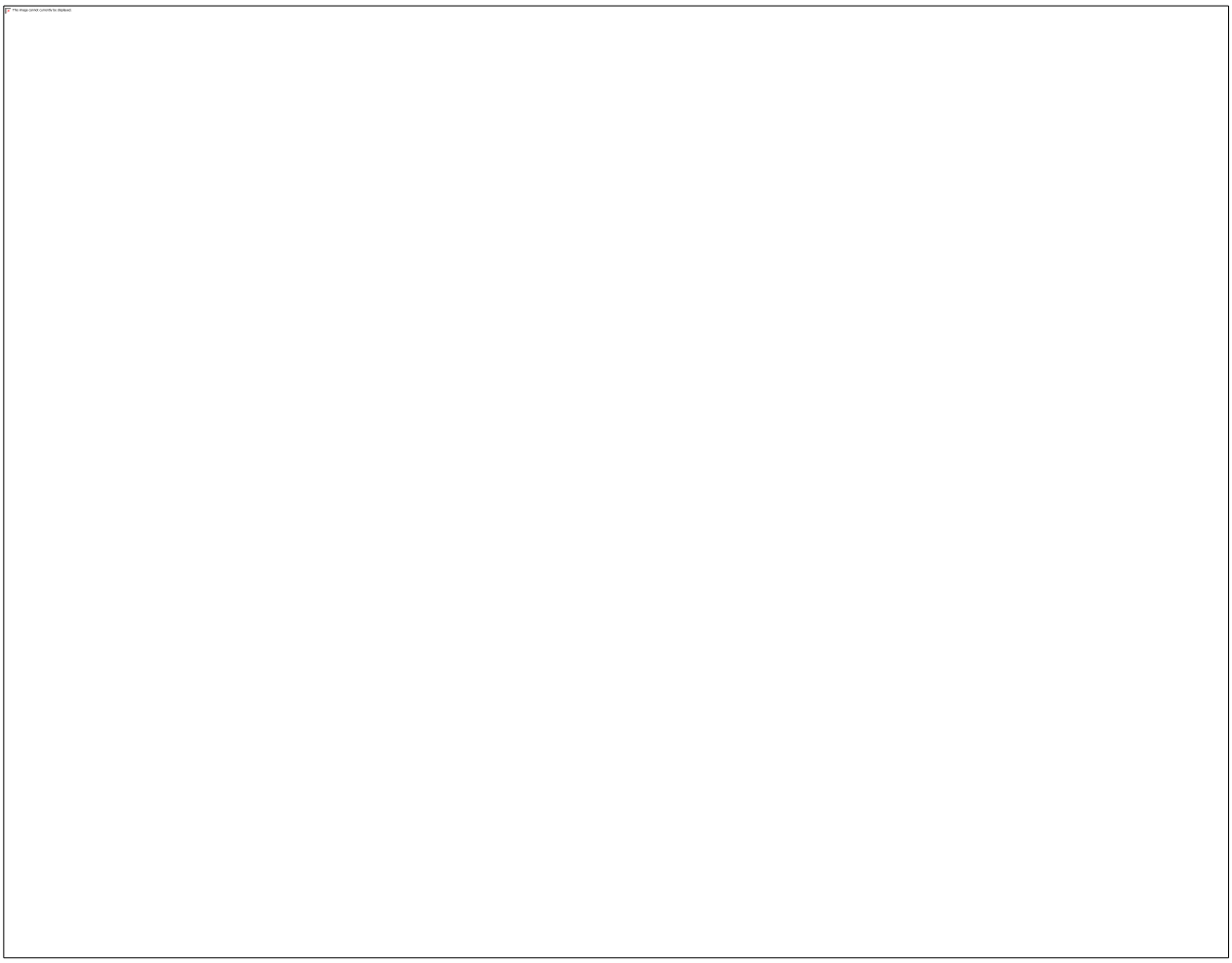

Figure 2. NSGA-II Pareto-optimal power–delay fronts for the four representative CMOS technology classes. Stars identify maximum-curvature knee solutions.

Table 4. Pareto-front extrema and knee solutions by technology class.

Technology	ND n	Pmin (mW)	t at Pmin (ps)	tmin (ps)	P at tmin (mW)	Pknee (mW)	tknee (ps)
45 nm planar	100	1.235	144.987	9.069	14.564	2.366	23.041
22 nm FinFET	100	0.449	60.544	3.822	6.473	0.858	9.811
12 nm FinFET	100	0.169	28.916	2.126	4.159	0.444	4.809
5 nm GAAFET	100	0.071	12.714	1.234	2.529	0.226	2.541

Power is for the 1,000-stage switching bank; delay is per FO4 stage. ND = nondominated.

4.2 Technology scaling and the physical trade-off

The movement of the fronts is consistent with lower switched capacitance and improved electrostatic control. The 22 nm knee power is 36.3% of the 45 nm knee value, and its delay is 42.6%. At 12 nm, the corresponding ratios are 18.8% and 20.9%; at 5 nm they are 9.5% and 11.0%. Figure 4(a) visualizes these normalized gains. The reductions arise jointly from lower external and internal capacitance, lower supply voltage, and maintained drive current per normalized unit. They are therefore system-level consequences of the assumed technology parameterization, not of gate length alone.

The fronts also reveal that voltage scaling cannot be considered in isolation. At the low-power extreme, low VDD and high effective resistance dominate delay. Along the curved middle region, increasing drive size and improving p/n balance yield large timing benefits. At the high-speed extreme, size-dependent capacitance produces diminishing returns, so power continues to grow while delay approaches an intrinsic floor. This is consistent with logical-effort reasoning: once parasitic and load capacitances scale with the device, unlimited upsizing cannot reduce delay proportionally (X. Zhang et al., 2025).

Sizing quantization did not prevent FinFET and GAAFET fronts from being smooth at the plotted scale because voltage, threshold, p/n ratio, and length bias remained continuous. Nevertheless, discrete drive units produced local clusters that a convex gradient method would not naturally traverse. The GAAFET class combined the smallest capacitance coefficient with a lower nominal p/n ratio, so transition balance was achieved with less p-device overhead. Strong gate control also allowed a smaller reference leakage coefficient than the 12 nm FinFET. These assumptions follow the qualitative advantages of stacked nanosheet devices (Rathore et al., 2023; Yin et al., 2022), while the robustness results show that low-threshold operation still exposes leakage sensitivity.

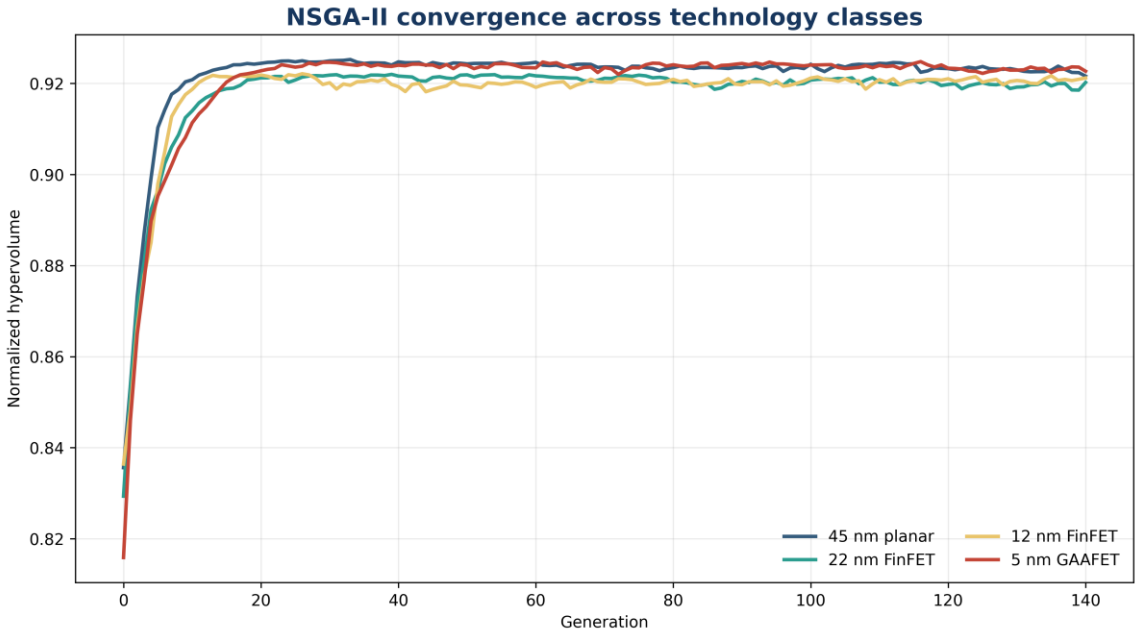

Figure 3. Hypervolume convergence of the main NSGA-II runs. Objectives and reference points were normalized within each technology class.

4.3 Convergence and algorithm comparison

Figure 3 plots hypervolume during the 140-generation main runs. All technology classes gained most of their dominated volume during approximately the first 20 generations and then entered a stable refinement phase. Final trajectories remained near 0.92 under the main-run normalization. Small late-generation fluctuations reflect changes in the finite population's distribution, not loss of elitism, because the nondominated archive quality remained close to the converged envelope. Similar convergence shapes across the device classes indicate that the decision-space repair and feasibility rules did not favor one architecture.

Table 5 aggregates the ten-run, four-technology comparison. NSGA-II achieved average hypervolume 1.16945 with mean within-technology standard deviation 0.00037. The random search obtained 1.15580, while the weighted-sum GA obtained 1.13996. NSGA-II's mean IGD was 0.01916, less than half the random-search value of 0.04275 and about one fifth of the weighted-sum value of 0.09845. Its spacing metric of 0.01697 was also substantially lower than 0.05394 and 0.23698, respectively.

The weighted-sum GA found only about ten nondominated designs per run because eleven scalar preferences concentrated the search near particular trade-offs. It also missed portions of the curved front, explaining its lower hypervolume and high IGD. Random search produced more nondominated points than the scalar method, but only 40.9% of sampled candidates were feasible on average and its points were uneven. By contrast, feasibility-first survival allowed the NSGA-II population to become fully feasible, and crowding distance maintained all 60 repeated-run population members as nondominated at termination. The result does not imply that NSGA-II is universally superior to every modern multi-objective algorithm; it demonstrates clear benefit over two transparent equal-budget baselines for this constrained two-objective problem.

Table 5. Equal-budget algorithm performance averaged over four technologies and ten runs per technology.

Method	**HV ↑**	**IGD ↓**	**Spacing ↓**	**ND count**	**Feasible (%)**
NSGA-II	1.16945 ± 0.00037	0.01916 ± 0.00598	0.01697	60.00	100.00
Weighted-sum GA	1.13996 ± 0.00583	0.09845 ± 0.01423	0.23698	9.95	98.09

Random search	1.15580 ± 0.00144	0.04275 ± 0.00894	0.05394	40.85	40.94

HV reference point = (1.10, 1.10), giving a maximum reference rectangle of 1.21. Reported ± values are the mean within-technology run SD. Lower IGD and spacing are better.

4.4 Knee-point designs

Table 6 lists the selected compromise variables. The 45 nm knee used VDD = 0.654 V, effective drive size 4.915, p/n ratio 1.806, Vth = 0.250 V, and a very small length bias. It achieved 2.366 mW and 23.041 ps. The 22 nm knee used four drive units and produced 0.858 mW at 9.811 ps. The 12 nm and 5 nm knees both used five discrete drive units, reaching 0.444 mW at 4.809 ps and 0.226 mW at 2.541 ps, respectively.

The knee solutions consistently selected supply values near the lower bound and threshold values near the low-threshold bound. This does not mean minimum voltage and minimum threshold are universally optimal. In this benchmark, a moderate increase in drive units compensates for reduced voltage, while the leakage-share constraint prevents uncontrolled low-threshold selection. The small length biases show that, at the knee, leakage reduction from a longer channel was less valuable than the resulting drive-current penalty. A standby-dominated workload or a tighter leakage constraint would move the knee toward higher threshold and positive length bias.

The p/n ratios decreased from 1.806 in planar CMOS to 1.570 in the GAAFET class. This tracks the assumed improvement in relative p-device strength and reduces capacitive overhead. The ratios do not exactly equal the technology nominal values because the optimizer balances current mismatch, intrinsic capacitance, and delay rather than imposing identical rise and fall current. The resulting imbalance remained within 0.16 for every feasible solution.

Leakage share at the knee was 2.15% for 45 nm, 5.40% for 22 nm, 14.65% for 12 nm, and 13.46% for 5 nm. The absolute power still decreased with scaling, but the composition shifted. This is an important design message: a smaller total power does not eliminate leakage management. In low-activity blocks, power gating, multi-threshold cells, reverse body bias where available, and workload-aware voltage control may still be necessary (Dubey et al., 2025; Kim et al., 2024).

Table 6. Decision variables and performance of the selected knee solutions.

Technology	VDD	Drive	p/n	Vth	ΔL	Power	Delay	Leakage (%)
45 nm planar	0.654	4.915	1.806	0.250	0.0049	2.366	23.041	2.15
22 nm FinFET	0.556	4.000	1.708	0.221	0.0001	0.858	9.811	5.40
12 nm FinFET	0.453	5.000	1.737	0.180	0.0023	0.444	4.809	14.65
5 nm GAAFET	0.422	5.000	1.570	0.160	0.0000	0.226	2.541	13.46

VDD and Vth are in volts; power is in mW per 1,000-stage bank; delay is in ps per FO4 stage.

4.5 PVT robustness

Table 7 and Figure 4(b) summarize the 2,000-sample Monte Carlo experiment. Mean power exceeded nominal power for all knees because the temperature distribution was centered at 65 °C rather than the nominal 25 °C and because the lognormal leakage multiplier is right-skewed. For 45 nm, power increased from 2.366 mW nominal to 2.496 ± 0.146 mW, with a 95th percentile of 2.741 mW. Mean delay increased from 23.041 to 25.797 ± 2.478 ps, and the 95th percentile reached 30.025 ps.

The 22 nm design showed 0.964 ± 0.088 mW and 10.890 ± 1.030 ps, with 95th-percentile values of 1.127 mW and 12.630 ps. At 12 nm, mean power rose more markedly from 0.444 to 0.574 mW because the nominal knee already had a 14.65% leakage share; its 95th-percentile power was 0.768 mW. The 5 nm design produced 0.291 ± 0.052 mW and 2.787 ± 0.273 ps, with 95th percentiles of 0.389 mW and 3.254 ps. The normalized delay boxes in Figure 4(b) have comparable dispersion, showing that the assumed relative on-current, capacitance, threshold, and temperature variations affect the four classes in broadly similar proportions.

The Monte Carlo results support the knees as stable early-design compromises but also demonstrate the need for margin. A designer using the nominal 5 nm delay of 2.541 ps as a hard sign-off target would underbudget the 95th-

percentile delay by about 28%. The appropriate response is not necessarily to discard the knee; it is to select a nearby Pareto point with extra speed margin or to optimize a statistical objective such as mean plus a multiple of standard deviation. Robust multi-objective optimization could therefore extend the present formulation by treating a delay percentile and expected power as the two objectives.

Table 7. PVT robustness of knee solutions under 2,000 Monte Carlo samples.

Technology	Pnom	PMC mean ± SD	P95	tnom	tMC mean ± SD	t95
45 nm planar	2.366	2.496 ± 0.146	2.741	23.041	25.797 ± 2.478	30.024
22 nm FinFET	0.858	0.964 ± 0.088	1.127	9.811	10.890 ± 1.030	12.630
12 nm FinFET	0.444	0.574 ± 0.104	0.768	4.809	5.278 ± 0.518	6.190
5 nm GAAFET	0.226	0.291 ± 0.052	0.389	2.541	2.787 ± 0.273	3.254

Power values are mW; delay values are ps. Temperature was N(65 °C, 18 °C), clipped to 0–125 °C; other variation assumptions are stated in Section 3.6.

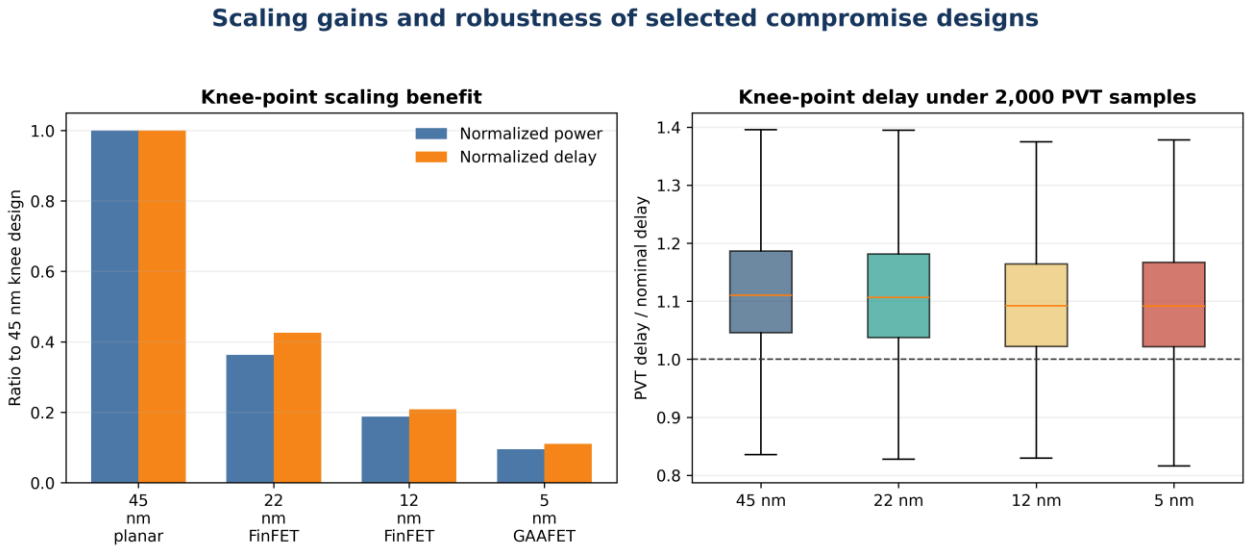

Figure 4. Normalized knee-point scaling benefits and Monte Carlo PVT delay distributions. The dashed line in panel (b) indicates nominal delay.

4.6 Implications for next-generation VLSI design

The Pareto archive is useful at several levels. A library designer can select a low-power point for noncritical cells, a knee point for most logic, and a high-speed point for timing-critical arcs. A system architect can map thermal or latency budgets onto the front without rerunning the search. A technology researcher can replace the generic evaluator with BSIM-CMG or TCAD-calibrated response surfaces and quantify how contact resistance, sheet count, or parasitic capacitance reshapes the objective envelope. Because NSGA-II does not require derivatives, the same framework can include discrete threshold flavors, integer fin or sheet counts, and categorical device architectures.

The results also caution against reporting only the power–delay product. At the two extremes, a product may favor a point that is unacceptable under either a strict power cap or a hard timing requirement. The Pareto front makes the opportunity cost explicit. Knee selection supplies one communicable recommendation, but preserving the full archive is the more important outcome.

4.7 Limitations and validity

Three limitations bound interpretation. First, the technology sets are representative, model-calibrated classes rather than proprietary PDKs. Absolute power and delay values should not be used for tape-out claims. Second, the FO4 inverter isolates device and local circuit behavior; a full VLSI path includes complex gates, interconnect, clocking, memory, voltage droop, and activity correlation. Third, the PVT distributions are common exploratory assumptions and do not encode spatial correlation, random telegraph noise, aging, self-heating feedback, or foundry corner definitions.

These limits do not invalidate the optimization comparison because every algorithm and technology class was evaluated under the same declared rules. Internal validity is supported by equal evaluation budgets, independent repetitions, explicit constraints, multiple quality indicators, and consistent PVT sampling. External validity requires the next step: calibrate the evaluator to extracted standard cells, verify front points in SPICE across official corners, and repeat the optimization at path and block levels. Such substitution changes the cost of evaluation, not the conceptual method.

5. Conclusions

This paper presented a reproducible NSGA-II framework for simultaneous minimization of power dissipation and propagation delay in representative 45 nm planar, 22 nm FinFET, 12 nm FinFET, and 5 nm GAAFET technology classes. By retaining power and delay as separate objectives, the method exposed the complete cost of voltage, threshold, and sizing decisions. The optimized fronts showed a consistent steep low-power region, a high-value knee, and diminishing timing returns at high power.

NSGA-II achieved the strongest equal-budget performance among the evaluated methods, with average hypervolume 1.16945, IGD 0.01916, and spacing 0.01697 across four technologies and ten repetitions. The curvature-based knee designs progressed from 2.366 mW and 23.041 ps for 45 nm planar CMOS to 0.226 mW and 2.541 ps for the 5 nm GAAFET class. The results quantify the benefit of lower capacitance and supply voltage while showing that leakage constitutes a growing share at scaled low-threshold knees. Monte Carlo PVT analysis confirmed stable relative behavior but revealed meaningful 95th-percentile timing and power margins, especially where leakage was substantial.

The principal contribution is not a claim of foundry-specific device performance; it is a transparent decision framework. A designer can inspect the Pareto archive, select points under changing thermal and timing budgets, and evaluate robustness before committing to one cell configuration. Future work should replace the analytical evaluator with PDK-calibrated BSIM-CMG or SPICE data, incorporate extracted interconnect and layout quantization, model aging and correlated variation, and extend the objectives to area, energy per operation, reliability, and yield. Surrogate-assisted or parallel evolutionary search may reduce the cost when each candidate requires transistor-level simulation. With these extensions, Pareto-based technology–circuit co-optimization can become a practical bridge between emerging device research and next-generation VLSI implementation.

Declarations

Funding: No external funding was used for this model-based study.

Conflict of interest: The author(s) declare no conflict of interest.

Ethics approval: Not applicable; the study involved no human participants, animals, or identifiable personal data.

Data and code availability: All model equations, parameter ranges, algorithm settings, seeds, and robustness assumptions required to reproduce the reported benchmark are disclosed in the manuscript. The generated Pareto and convergence datasets accompany the document-generation workspace.

References

1. Blank, J., & Deb, K. (2020). Pymoo: Multi-Objective Optimization in Python. IEEE Access, 8, 8949789509. https://doi.org/10.1109/ACCESS.2020.2990567
2. Chen, Z. (2022). Gate-all-around nanosheet transistors go 2D. Nature Electronics, 5, 1–2. https://doi.org/10.1038/s41928-022-00899-4
3. Das, R., Rajalekshmi, T., & James, A. (2024). FinFET to GAA MBCFET: A Review and Insights. IEEE Access, 12, 50556–50577. https://doi.org/10.1109/ACCESS.2024.3384428

4. Deb, K., & Jain, H. (2014). An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. IEEE Transactions on Evolutionary Computation, 18, 577–601. https://doi.org/10.1109/TEVC.2013.2281535
5. Deb, K., Pratap, A., & Agarwal, S. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2), 182.
6. Dubey, P., Marian, D., Toral-Lopez, A., Knobloch, T., Grasser, T., & Fiori, G. (2025). Simulation of Vertically Stacked 2-D Nanosheet FETs. IEEE Transactions on Electron Devices, 72, 1494–1500. https://doi.org/10.1109/TED.2025.3533474
7. Karimi, K., Fardoost, A., & Javanmard, M. (2024). Comprehensive Review of FinFET Technology: History, Structure, Challenges, Innovations, and Emerging Sensing Applications. In Micromachines (Vol. 15, Issue 10, p. 1187). https://doi.org/10.3390/mi15101187
8. Kim, M. S., Lee, S. H., Park, J., Jeon, S. R., Bae, S. J., Hong, J. W., Jang, J., Bae, J.-H., Yoon, Y. J., & Kang, I. M. (2024). Statistical Analysis of Increased Immunity to Poly-Si Grain Boundaries in Nanosheet CMOS Logic Inverter Through Sheet Stacking. Silicon, 16(16), 5855–5864. https://doi.org/10.1007/s12633-024-03113-6
9. Kumari, N. A., & Prithvi, P. (2023). A Comprehensive Analysis of Nanosheet FET and its CMOS Circuit Applications at Elevated Temperatures. Silicon, 15(14), 6135–6146. https://doi.org/10.1007/s12633-023-02496-2
10. Lberni, A., Alami Marktani, M., Ahaitouf, A., & Ali, A. (2023). Analog circuit sizing based on Evolutionary Algorithms and deep learning. Expert Systems with Applications, 237, 121480. https://doi.org/10.1016/j.eswa.2023.121480
11. Lee, J., & Yoo, H.-J. (2021). An Overview of Energy-Efficient Hardware Accelerators for On-Device Deep-Neural-Network Training. IEEE Open Journal of the Solid-State Circuits Society, 1, 115–128. https://doi.org/10.1109/OJSSCS.2021.3119554
12. Mukesh, S., & Zhang, J. (2022). A Review of the Gate-All-Around Nanosheet FET Process Opportunities. In Electronics (Vol. 11, Issue 21, p. 3589). https://doi.org/10.3390/electronics11213589
13. Neelam, A., Sreenivasulu, V., Vijayvargiya, V., Upadhyay, A., Ajayan, J., & Uma, M. (2024). Performance Comparison of Nanosheet FET, CombFET, and TreeFET: Device and Circuit Perspective. IEEE Access, 12, 9563–9571. https://doi.org/10.1109/ACCESS.2024.3352642
14. Nguyen, H. T., & Hoang, T. (2023). A Novel Framework of Genetic Algorithm and Spectre to Optimize Delay and Power Consumption in Designing Dynamic Comparators. In Electronics (Vol. 12, Issue 16, p. 3392). https://doi.org/10.3390/electronics12163392
15. Rathore, S., Jaisawal, R. K., Kondekar, P. N., & Bagga, N. (2023). Demonstration of a Nanosheet FET With High Thermal Conductivity Material as Buried Oxide: Mitigation of Self-Heating Effect. IEEE Transactions on Electron Devices, 70(4), 1970–1976. https://doi.org/10.1109/TED.2023.3241884
16. Sakurai, T., & Newton, A. R. (1990). Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas. IEEE Journal of Solid-State Circuits, 25(2), 584–594. https://doi.org/10.1109/4.52187
17. Sanabria-Borbón, A. C., Soto-Aguilar, S., Estrada-López, J. J., Allaire, D., & Sánchez-Sinencio, E. (2020). Gaussian-Process-Based Surrogate for Optimization-Aided and Process-Variations-Aware Analog Circuit Design. In Electronics (Vol. 9, Issue 4, p. 685). https://doi.org/10.3390/electronics9040685
18. Valasa, S., Tayal, S., Thoutam, L. R., Ajayan, J., & Bhattacharya, S. (2022). A critical review on performance, reliability, and fabrication challenges in nanosheet FET for future analog/digital IC applications. Micro and Nanostructures, 170, 207374. https://doi.org/https://doi.org/10.1016/j.micrna.2022.207374
19. Valencia-Ponce, M. A., Tlelo-Cuautle, E., & de la Fraga, L. G. (2021). On the Sizing of CMOS Operational Amplifiers by Applying Many-Objective Optimization Algorithms. In Electronics (Vol. 10, Issue 24, p. 3148). https://doi.org/10.3390/electronics10243148
20. Vişan, C., Pascu, O., Stănescu, M., Şandru, E.-D., Diaconu, C., Buzo, A., Pelz, G., & Cucu, H. (2022). Automated circuit sizing with multi-objective optimization based on differential evolution and Bayesian inference. Knowledge-Based Systems, 258, 109987. https://doi.org/https://doi.org/10.1016/j.knosys.2022.109987
21. Wang, M. (2024). A Review of Reliability in Gate-All-Around Nanosheet Devices. In Micromachines (Vol. 15, Issue 2, p. 269). https://doi.org/10.3390/mi15020269
22. Xu, Z., Zhao, Z., & Liu, J. (2024). Deterministic Multi-Objective Optimization of Analog Circuits. In Electronics (Vol. 13, Issue 13, p. 2510). https://doi.org/10.3390/electronics13132510
23. Yin, S., Ruitao, W., Zhang, J., Liu, X., & Wang, Y. (2022). Fast Surrogate-Assisted Constrained Multiobjective Optimization for Analog Circuit Sizing via Self-Adaptive Incremental Learning. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42, 2080–2093. https://doi.org/10.1109/TCAD.2022.3221694
24. Yoon, J.-S., & Baek, R.-H. (2020). Device Design Guideline of 5-nm-Node FinFETs and Nanosheet FETs for Analog/RF Applications. IEEE Access, 8, 189395–189403. https://doi.org/10.1109/ACCESS.2020.3031870
25. Zhang, Q., Zhang, Y., Luo, Y., & Yin, H. (2024). New structure transistors for advanced technology node CMOS ICs. National Science Review, 11. https://doi.org/10.1093/nsr/nwae008
26. Zhang, X., Li, Q., Cao, L., Zhang, Q., Jiang, R., Wang, P., Yao, J., & Yin, H. (2025). Performance Optimization of Fabricated Nanosheet GAA CMOS Transistors and 6T-SRAM Cells via Source/Drain Doping Engineering. IEEE Journal of the Electron Devices Society, 13, 86–92. https://doi.org/10.1109/JEDS.2025.3531432

27. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., & Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: an analysis and review. IEEE Transactions on Evolutionary Computation, 7, 117–132. https://doi.org/10.1109/TEVC.2003.810758