

AI-driven TCAD Surrogate Modeling and Multi-Objective Optimization for Accelerated Technology Development

Derek Dou

Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
derek.dou@sandisk.com

Sudarshan Narayanan

Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
sudarshan.narayanan@sandisk.com

Qinghua Zhao

Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
qinghua.zhao@sandisk.com

Chao Qin

Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
chao.qin2@sandisk.com

Mohan Dunga

Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
mohan.dunga@sandisk.com

Abstract— We present a lightweight, technology-agnostic optimization framework that accelerates semiconductor device development by combining compact TCAD-trained surrogate models with the NSGA-II multi-objective genetic algorithm. Using small multi-layer perceptron models tailored to split simulation conditions, the framework quickly and accurately predicts device performance across large process spaces. NSGA-II then efficiently identifies Pareto-optimal trade-offs without requiring additional TCAD runs. Benchmark results show a 2–3× reduction in overall turnaround time and significantly lower engineering effort compared with conventional TCAD-only workflows, while maintaining strong agreement with full TCAD verification. The proposed approach provides a fast, scalable solution for rapid design-space exploration and device optimization.

Keywords—TCAD, semiconductor device design, multi-objective optimization, machine learning, surrogate models

I. INTRODUCTION

Technology Computer-Aided Design (TCAD) is a foundational tool for semiconductor technology development because it provides physics-based insight into device behavior before costly wafer experiments are performed. After a process and device simulation deck has been calibrated to silicon data, TCAD can be used to evaluate the impact of process splits, structural changes, and device-level trade-offs. However, calibration is only the first stage of the development cycle. In practice, once the baseline deck is available, engineers must still search a large process design space to identify combinations of implant conditions, spacer dimensions, and other process knobs that satisfy competing electrical targets. This optimization step often becomes the dominant bottleneck because full-flow simulations can require hours per split, while the number of possible process combinations grows rapidly with each additional parameter.

Conventional TCAD-based optimization is typically performed in a staged and manually guided manner. Engineers first select a limited design-of-experiments (DOE), inspect the resulting trends, narrow the search range, and then run additional

simulations around promising regions. Although this approach is physically interpretable, it is labor intensive and can miss globally favorable trade-offs when parameter interactions are nonlinear. The challenge is especially important for advanced CMOS and memory peripheral devices, where electrical metrics such as speed, leakage, short-channel control, and reliability margins are tightly coupled. A faster optimization method is therefore needed to convert calibrated TCAD data into actionable process guidance with minimal additional simulation cost.

Machine-learning-assisted TCAD workflows have recently attracted attention as a way to reduce the cost of device exploration. Prior studies have demonstrated that surrogate models can emulate expensive physics simulations and can be combined with optimization algorithms for design-space search [1], [2]. Nevertheless, many reported approaches rely on relatively heavy architectures, large datasets, or device-specific representations such as mesh-based graph inputs. These methods can be powerful, but they are not always convenient for day-to-day technology development, where the available dataset may be small, the optimization target may change frequently, and engineers need a framework that can be quickly retrained for new split conditions.

In this work, we introduce a lightweight and technology-agnostic optimization platform that combines compact multi-layer perceptron (MLP) surrogate models with the Non-Dominated Sorting Genetic Algorithm II (NSGA-II). Instead of attempting to replace TCAD entirely, the proposed method uses a moderate DOE dataset to learn the mapping between process parameters and selected device metrics. The trained surrogates then provide near-instant objective evaluations for NSGA-II, enabling rapid identification of Pareto-optimal process conditions. A NAND peripheral NMOS case study with seven process parameters and two competing objectives is used to demonstrate the method. The framework generates high-quality Pareto fronts within seconds, maintains strong agreement with TCAD verification, and reduces both

turnaround time and manual engineering effort compared with a conventional TCAD-only workflow.

II. METHODOLOGY

The proposed method combines lightweight surrogate modeling with multi-objective genetic optimization to accelerate TCAD-based device exploration. The key idea is to separate the expensive physics evaluation step from the repeated objective evaluations required by an optimizer. TCAD is used only to generate a representative DOE dataset and to verify final candidate solutions, while the computationally inexpensive surrogate models are used during the large-scale optimization loop.

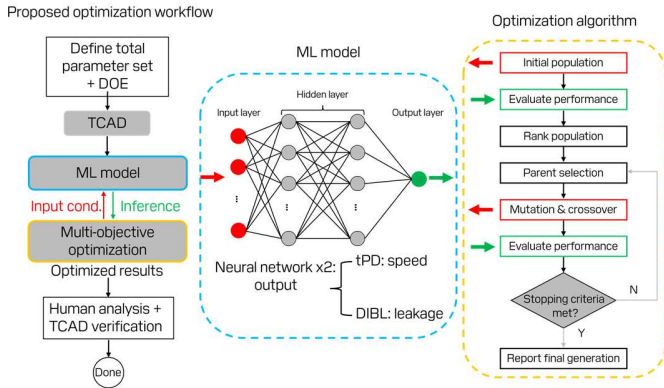


Fig. 1. Breakdown of optimization workflow. ML model and optimization algorithm are illustrated. The red and green arrows indicate the input/output directions.

As shown in Fig. 1, the input variables are process parameters selected by the device engineer, and the output variables are the device metrics used as optimization objectives. A compact MLP is adopted as the surrogate model because it provides sufficient nonlinear fitting capability while remaining simple to train and deploy. Separate MLPs are trained for each target metric rather than using a single multi-output network. This choice allows the model size, training behavior, and loss convergence to be tuned independently for each objective, which is useful when different metrics have different numerical scales or physical sensitivities.

Before training, the DOE data are divided into training and testing subsets, with 20% of the samples reserved for testing. The process parameters are normalized to improve numerical conditioning, and each target metric is modeled directly from the process-split vector. The surrogate model is trained using standard regression loss minimization, and its predictive quality is evaluated by comparing predicted and TCAD-simulated values on both training and held-out testing data. This train-test separation provides a practical check that the surrogate is capturing the underlying process-to-device relationship rather than only memorizing the DOE points.

The full workflow consists of three connected stages. First, a DOE is generated across the selected process window and full TCAD simulations are executed to obtain the target device metrics. Second, the DOE results are used to train the MLP surrogates, and NSGA-II is executed using the surrogate predictions as objective functions. Third, representative

Pareto-optimal candidates are selected for TCAD verification to confirm that the predicted trade-offs remain valid under the full physics model. NSGA-II is selected because it is well suited for problems with conflicting objectives and does not require gradients, which are generally unavailable for TCAD-based device metrics. Through non-dominated sorting and crowding-distance preservation, NSGA-II maintains a diverse set of candidate solutions along the Pareto front rather than collapsing to a single optimum.

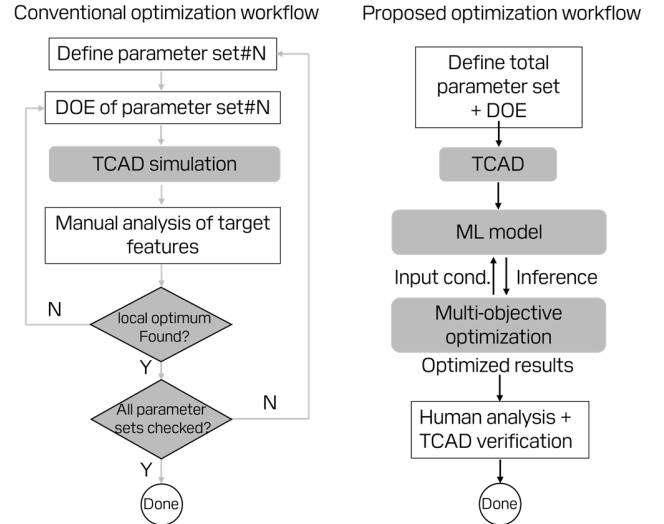


Fig. 2. Comparison between conventional (left) and proposed (right) optimization workflows.

Fig. 2 summarizes the contrast between the conventional and proposed workflows. In the conventional approach, engineers repeatedly inspect simulation trends, manually decide the next split direction, and launch additional TCAD runs until a satisfactory compromise is found. In the proposed approach, the DOE is used more efficiently: once the surrogate models are trained, the optimizer can evaluate thousands of virtual candidates at negligible cost. Human input is still essential for selecting meaningful process bounds, interpreting the final Pareto set, and choosing verification points, but the repetitive search and ranking steps are automated. This makes the framework practical for iterative development, where objectives and constraints may be updated several times during a project.

III. RESULTS AND DISCUSSION

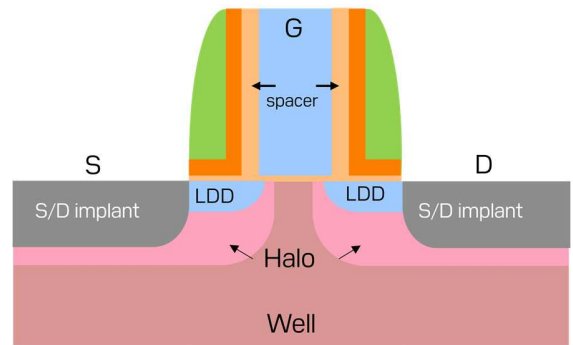


Fig. 3. Illustration of the device structure of interest.

The optimization study was conducted on the NAND peripheral NMOS device structure shown in Fig. 3. The selected process variables include offset spacer thickness and the dose and energy conditions for the lightly doped drain (LDD), halo, and source/drain (S/D) implants. These parameters are representative of knobs commonly adjusted during transistor optimization because they directly influence extension resistance, lateral junction placement, channel electrostatics, and short-channel behavior. As a result, they provide a meaningful test case for evaluating whether a surrogate-assisted method can capture nonlinear device trends across a realistic process window.

Two performance metrics were considered in this study. The first objective is propagation delay (tPD), extracted from a ring-oscillator-based configuration and used as an indicator of device speed. The second objective is drain-induced barrier lowering (DIBL), which is used as a leakage and electrostatic-control indicator. These objectives are naturally competing: process changes that improve drive current and speed may degrade short-channel control, while changes that suppress leakage can increase resistance or capacitance and therefore slow the device. This speed-leakage trade-off makes the problem well suited for Pareto-front analysis rather than single-objective optimization.

TABLE 1

DEVICE PARAMETERS WITH RANGES OF VARIATION

Parameters	Range (%) w.r.t POR
Offset spacer thickness	$\pm 25\%$
LDD energy & dose	$\pm 25\%$
Halo energy & dose	$\pm 25\%$
Source/Drain energy & dose	$\pm 50\%$

A full DOE consisting of 335 TCAD simulations was generated across the seven-parameter space, as summarized in Table 1. The DOE was intentionally kept moderate in size to reflect a realistic engineering constraint: although TCAD is highly informative, expanding the dataset indefinitely is not practical because each additional split consumes computing resources and queue time. The use of compact surrogates therefore targets the common development scenario in which a limited but representative set of calibrated simulations is available, and the goal is to extract as much optimization value from that dataset as possible.

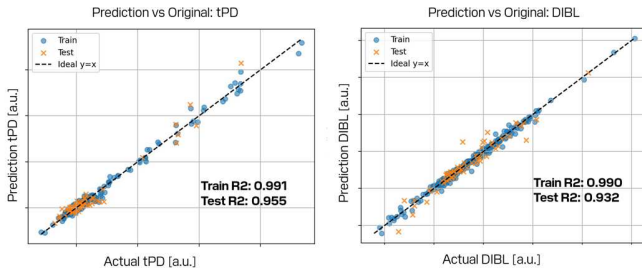


Fig. 4. Prediction vs. actual values for the two objectives: tPD (left) and DIBL (right).

Surrogate models were implemented using MLPs with three hidden layers of 128 neurons. This architecture was selected as

a balanced choice between fitting capability and computational simplicity. As shown in Fig. 4, the predicted values closely follow the TCAD-simulated values for both tPD and DIBL, including on the held-out testing data. The agreement indicates that the compact MLPs can capture the dominant nonlinear response of the device metrics over the selected process window. Using a learning rate of 0.001 and 150 training epochs, high and consistent R^2 values were obtained across both targets, supporting the use of the models as inexpensive objective evaluators during optimization.

The use of separate surrogate models also provides practical flexibility. If a new objective is added, such as threshold voltage, on-current, capacitance, or reliability-related margin, the corresponding model can be trained without changing the optimizer interface. Similarly, if one target metric requires additional data or a modified model architecture, it can be improved independently. This modularity is important for technology development, where optimization targets often evolve as device, circuit, and integration requirements are refined.

The NSGA-II optimizer was then executed using the trained MLPs. As illustrated in Fig. 5(a), the initial population spans a broad region of the tPD–DIBL space and rapidly converges toward the Pareto front within roughly 30 generations. The solution set stabilizes by 50 generations, indicating that the algorithm first performs broad exploration and then refines the non-dominated region. Because the objective evaluations are performed by the surrogate models, this convergence occurs without launching additional TCAD simulations during the search loop.

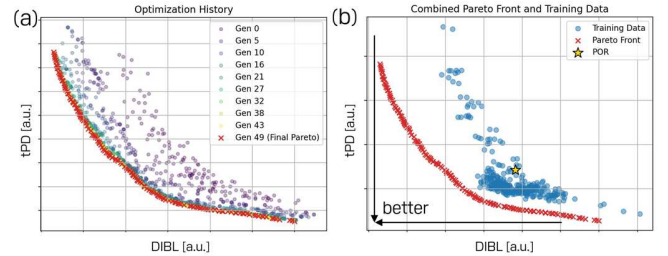


Fig. 5. (a) Optimization history of 50 generations. (b) combined pareto front and training data and POR reference.

To evaluate the overall improvement, the final Pareto front was plotted together with the DOE training points and the POR reference. As shown in Fig. 5(b), the Pareto set provides significantly improved speed–leakage trade-offs compared with both the original DOE samples and the POR device. The result demonstrates an important benefit of the proposed framework: even when the surrogate is trained only from the DOE points, NSGA-II can interpolate within the process window and identify combinations that were not explicitly simulated in the original dataset. These candidates can then be prioritized for TCAD verification instead of requiring engineers to manually infer the next split direction from scattered DOE trends.

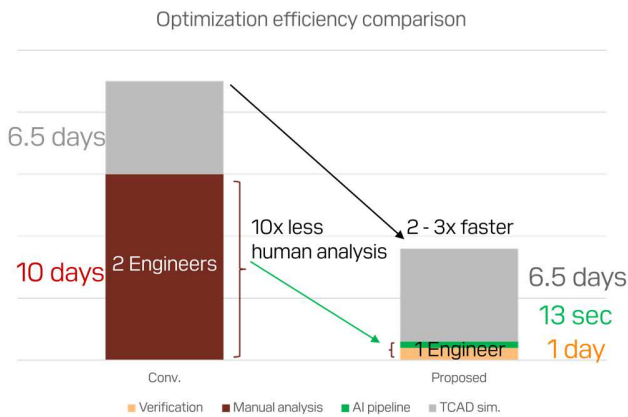


Fig. 6. Optimization workflow efficiency (days) comparison.

Notably, the full surrogate-assisted optimization, including MLP training and NSGA-II search, completes in only 13 seconds. This short runtime highlights the lightweight nature of the proposed pipeline and shows that the optimization stage is no longer the dominant schedule limiter once the DOE data have been generated. While both the conventional and proposed approaches share the same baseline TCAD simulation cost of approximately 6–7 days, the key difference lies in the post-simulation decision process. The conventional workflow requires roughly 10 additional days of manual analysis, trend interpretation, and iterative decision-making. As shown in Fig. 6, the proposed method reduces this human-driven portion to approximately one day, resulting in a 2–3 $\times$ overall turnaround-time reduction, 10 $\times$ lower human analysis effort, and a 2 $\times$ decrease in required engineering resources.

Beyond runtime reduction, the proposed approach improves workflow consistency. Manual optimization can depend strongly on the experience of individual engineers and may vary between projects. In contrast, a surrogate-assisted Pareto search provides a repeatable procedure for ranking candidate process

conditions under multiple objectives. The method does not eliminate the need for expert review; rather, it shifts engineering effort from repetitive screening to higher-value tasks such as defining the optimization problem, checking physical plausibility, selecting verification splits, and making final technology trade-off decisions. An excellent style manual for science writers is [7].

IV. CONCLUSION

We developed a fast, lightweight, and technology-agnostic multi-objective optimization pipeline using compact MLP surrogates and NSGA-II. Applied to a NAND peripheral NMOS device with seven process parameters and two competing objectives, tPD and DIBL, the framework produces high-quality Pareto-optimal solutions with a total optimization runtime of 13 seconds. The results show that a moderate TCAD DOE can be converted into an efficient surrogate-based search space, enabling rapid exploration of speed–leakage trade-offs without repeated manual TCAD iterations. Compared with a conventional TCAD workflow, the proposed method enables a 2–3 $\times$ reduction in overall turnaround time, a 10 $\times$ reduction in human analysis effort, and a 2 $\times$ reduction in required engineering resources. The approach is modular and can be extended to additional device metrics, process knobs, and technology platforms, making it a practical framework for accelerating CMOS device optimization in early and iterative technology development.

REFERENCES

- [1] T. Ma et al., "Fast Design Technology Co-Optimization Framework for Emerging Technology With Hierarchical Graph Embedding," *Int. Symp. Electron. Des. Autom. (ISEDA)*, 2024.
- [2] M. Ehteshamuddin et al., "Machine Learning-Assisted Multiobjective Optimization of Advanced Node Gate-All-Around Transistor for Logic and RF Applications," *IEEE Trans. Electron Devices*, vol. 71, no. 2, pp. 976–982, 2024.