Fast and efficient ResNN and Genetic optimization for PVT aware performance enhancement in digital circuits

Kushagra Agarwal*, Aryamaan Jain*, Deepthi Amuru, Zia Abbas
International Institute of Information Technology, Hyderabad
Email: {kushagra.agarwal, aryamaan.jain, deepthi.amuru}@research.iiit.ac.in, zia.abbas@iiit.ac.in

***Abstract*—This paper presents a fast and efficient optimization engine with multi-directional, multi-objective algorithms based on a robust transistor sizing approach to improve digital circuit performance. However, such optimization processes are highly simulator-dependent and computationally expensive tasks. Therefore, we propose developing machine learning-based reliable models considering process and operating variations to speed up the optimization procedure by running them on developed Residual Neural Network (ResNN) models instead of running expensive circuit simulations. Results on 22nm Metal Gate High-K digital cells show a reduction in delay and leakage up to 36.7% and 18.8%, respectively improving computational efficiency by several orders.**

***Index Terms*—Yield optimization, Genetic algorithm, Machine Learning, Leakage power, Propagation delay, CMOS, VLSI**

I. INTRODUCTION

For the past few decades, the IC (Integrated Circuit) industry has reinforced the electronic industry in designing and developing low-power, high-speed, complex, and compact devices at a reduced cost. The down-scaling of transistors is one of the phenomenal factors contributing to this development [1]. However, down-scaling to nanometer regime concerns increased static/leakage power dissipation and circuit sensitivity to process variations in manufacturing, challenging state-of-the-art circuit reliability [2]. The inter-die and intra-die variations in a chip are magnifying beyond the 45nm technology node, causing deviations in electrical characteristics of transistors manifested due to physical imperfections during manufacturing, leading to circuit performance deviation affecting the chip yield. The fluctuations in supply voltage and operating temperatures combined with process variations further deviate circuit performance from expected, increasing the threat of functional failures and timing mismatch. The efficient design of an IC confides in the circuit performance optimization in terms of power dissipation, operating speeds, and area. Therefore, PVT-aware circuit optimization with high yield has therefore became an interesting and urgent field of research in recent times.

Under this perception, we propose a PVT-aware multi-objective mathematical optimization engine based on Genetic Algorithm (GA) proficient in exploring the entire design space to find the optimal sizing of all devices in the standard cells to maximize the yield w.r.t power and speed specifications. Many experts have proposed algorithm-based optimization at various levels in the literature. However, such optimizations are simulation-dependent and computationally expensive tasks. At present, the speed up in analysis is one of the critical requirements of IC manufacturing with the highly demanding time to market scenarios on par with yield enhancement. The problem of improved speed of execution/runtime is possible with incorporating efficient Machine Learning (ML) techniques for circuit analysis and/or optimization. The novelty of this work is in assimilating the cutting-edge optimization algorithms running over the deep Residual Neural Network (ResNN) to improve the computational speed by several orders over simulator-dependent applications.

II. PREVIOUS WORKS

Many research groups and scientists have addressed the reliability optimization of VLSI circuits through algorithmic-based transistor sizing. Beg et al. [3] proposed an automatic optimized transistor sizing method in a feedback control system based on CMOS logic gates with small and large fan-in. Gate-level optimization approaches were mentioned in [4], [5]. Here, all transistors in a given logic gate are scaled by the same factor, and such top-down approaches are deeply circuit-specific, becoming highly complex for larger circuits. This paper addresses this problem with a GA-based bottom-up approach that estimates optimized transistor sizing in each standard cell searching across PVT-aware design space. Transistor-level leakage optimization with critical path delay as bound is proposed in [6]–[11]. In [6], Gupta et al. considered PVT variations for leakage optimization but only for corner cases. The paper optimizes critical path delay and power keeping bounds on both w.r.t nominal values. Abbas et al. proposed yield optimization of CMOS standard cells through transistor sizing at 40nm low-power (LP) technology in [8], [9]. Swarm Intelligence, Spider Monkey Optimization (SMO), are proposed in [10] for leakage optimization of 45nm LP applications. Neighborhood Cultivation Genetic Algorithm (NCGA) and Glowworm swarm optimization (GSO) are proposed in [11] for optimization of critical path delay with average leakage in bound for high performance (HP) applications and leakage power with critical path delay in bound for LP applications, considering all the PVT variations and aging ef-

* Equal Contribution

fects at 22nm MGK node. However, these works are dependent on license-dependent simulators (Synopsis-HSPICE, Mentor Graphics-ELDO, Cadence-Virtuoso and so on) with extensive computational time. Our work aims at providing substantial power-delay optimization at a significantly faster rate through machine-learning-based surrogate modeling.

III. THE PROPOSED METHODOLOGY

Process variations and operating variations can combine at lower technology nodes beyond 45nm to considerably differentiate the actual design from the intended design. The performance of the design can vary and be lower than the expected one. Similarly, the chip power can vary significantly higher than the nominal values due to the exponential dependence between process/ device parameters and transistor leakage. This translates into a reduced parametric yield and hence limits the number of shippable products. Therefore, the motivation for this work is to find an efficient surrogate multi-objective optimization engine, which could provide the final circuit sizing (W/L of all the transistors in the circuit under test) and yield optimization of the targeted circuits, which are robust enough against all variations mentioned above and fully functional as long as a circuit is in the giving operating conditions. The proposed ResNN running over optimization will improve the speed of the execution and be highly accurate.

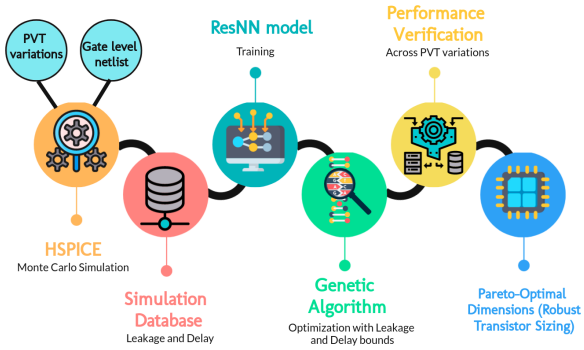

Fig. 1. The proposed PVT aware Pareto-optimal transistor sizing methodology

A. Modelling Delays and Leakages

The proposed methodology is as shown in Fig 1. The training data for the deep ResNN is generated as a vector of random values from the Gaussian distribution of each process parameter with 3σ variations in CMOS standard cells at 22nm High-K MGK through Predictive Technology Models (PTM) [12], [13], $X_p = [X_{p,1}X_{p,2}X_{p,3}...X_{p,k}]^T \epsilon R^k{}_{xp}$. 10 process parameters (PMOS and NMOS) - Channel Length, Transistor Width, Physical and Electrical equivalent of oxide thickness, nominal gate oxide thickness, Source/Drain junction depth, Channel doping concentration are considered in this work [14]. With these statistical distributions, random samples of temperature ranging from $-55°C$ to $125°C$ and supply voltage with a $\pm 10\%$ deviation from the nominal value (1.0V) are also included $X_r = [X_{r,1}X_{r,2}]^T \epsilon R_{xr}$ (operating

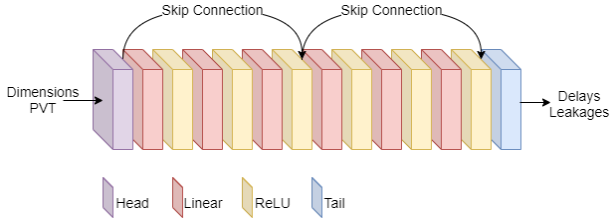

Fig. 2. A template of the neural network used to model delays and leakages given the dimensions and PVT values. Here, the value of network parameters used are block_size = 3 and num_blocks = 2.

variations). Design parameters - Channel length and width of each transistor in a cell (L and W respectively) are also varied in random within the bounds ($[22nm - 33nm]$ for L and $[44nm - 440nm]$ for W). Delay dependency on capacitive load is modeled as per [15]. Leakage power and propagation delay estimations of CMOS cells with PVT (Process including design parameters, supply voltage, temperature) variations are carried through HSPICE Monte-Carlo simulations.

In the next stage, the simulation standard cell database is given to train a machine learning model (ResNN) with residual blocks to predict delay and leakage values accurately for a given PVT variation. Further, genetic algorithm based optimization engine is built over ResNN to find the Pareto-optimal point with minimum delays and leakages by adjusting the gate dimensions (L and W of each transistor). Leakage and delay estimations for optimized gate dimensions are performed across the PVT to verify the nominal bounds (Nominal operating conditions (NOC) as Temperature = $25°C$ and Supply Voltage = 1.0V). The performance estimations were carried out on all the standard cells namely, AND2, AND3, NOR2, NOR3, NAND2, NAND3, XOR2 and FA.

The proposed model is capable of modeling more process parameters without increasing the modeling complexity. Abreast, it acts as a black box for modeling any technology node, including FinFETs.

B. Training the ResNN model

A surrogate model for PVT-aware leakage and delay estimation is developed employing residual neural networks (ResNN). ResNN is inspired by ResNet [16]. It has been shown that deeper networks perform better than their shallow counterparts to approximate complex functions, often resulting in low error values. Nevertheless, simply increasing the depth of a simple neural network results in failure to converge in many tasks due to the problem of vanishing gradients. However, our proposed ResNN model solves the vanishing gradients problem as the network learns the residuals rather than the output directly. Therefore, we chose ResNN to model the leakages and delays across the PVT variations. Let x represent the input parameters, f represent the learned function and y represent the output parameters. Then in residual networks, the output and gradients are yielded by

$$y = f(x) + x \implies \frac{dy}{dx} = \frac{df(x)}{dx} + 1$$

Here, $f(x) = y - x$ becomes a function to learn the residual. This makes the flow of gradient easier, thereby increasing the learning speed and accuracy.

We model the delays and leakages for each gate separately. The proposed ResNN architecture is demonstrated in Fig. 2. The network takes in the input parameters (PVT variations and gate dimensions) in the first stage. The head layer, which is a linear layer, projects this input to a lower or higher dimension, depending on the task. This projected input is then further sent to the body of the network which contains the residual connections. Finally, the tail of the network projects the vector from the body to the requisite delay or leakage dimensions.

The body of our network contains a series of residual layers - stacked linear and ReLU layers. We use two hyper parameters to tune our network for each gate,
(1) block_size - the size of the residual block or number of layers through which a skip connection is made and
(2) num_blocks - the number of blocks that a network has, which is the same as the number of skip connections. In Fig. 2, the block_size is kept at three with each skip connection consisting of 3 stacked *Linear + ReLU* blocks, and num_blocks is two with the network containing 2 skip connections in total. We used PyTorch to implement the proposed ResNN. Each gate level prediction model was trained for approximately 1000 epochs on a single RTX 2080 Ti GPU using Adam optimizer with learning rate set to 0.001.

C. Genetic Algorithm

The Genetic Algorithm (GA) is an evolutionary algorithm wherein random adjustments are made to existing solutions to produce more optimal ones. We used GA to find the optimal gate dimensions providing the lowest leakages and delays using the predictions from our trained ResNN model. We optimized transistor sizing (L and W values) in each standard cell while keeping all the other PVT at nominal (temperature - $25°C$, supply voltage - $1.0V$, process - nominal). This was done to ensure that the algorithm has a stable optimization goal and can find the best L and W values such that a consistent reduction in both the delay and leakage values can be achieved. A multi-objective fitness function was used to find the Pareto-optimal point, simultaneously reducing delays and leakages.

The sum of all delay and leakage combinations represents delay and leakage, respectively, in the fitness function. A population size of 100 was chosen, and the top 20% individuals (Elite Population) from the previous generation were carried forward to the next one. The rest 80% were subjected to random cross-overs and mutations, allowing incremental adjustments. The genetic algorithm was allowed to run till ten consecutive generations failed to improve the Pareto-optimal solution.

D. Performance verification across PVT variations

All the PVT were kept constant while running the genetic algorithm. Therefore, to ensure that the obtained gate dimensions are PVT-aware, we compared them with the initial sizing gate dimensions (estimated at nominal PVT) over 1000 iterations. The temperature and voltage values were randomly drawn from a uniform distribution with specified ranges in each iteration, whereas process parameters variations were sampled from a Gaussian with $\pm 3\sigma$ variance. The delay and leakage values for initial sizing and optimized sizing were then compared. A detailed description of results and observations are mentioned in the following section.

TABLE I
TRAINED RESNN MODEL PERFORMANCES FOR DELAY AND LEAKAGE ON THE HOLD OUT TEST SET. MSE FOR DELAY IS OF THE ORDER 10^{-24} AND FOR LEAKAGE IS 10^{-16}.

Gate	Delay			Leakage		
	R^2	MSE	MAPE	R^2	MSE	MAPE
AND2	0.998	0.615	1.786	0.997	9.194	1.504
AND3	0.997	1.220	2.488	0.998	9.215	1.222
NOR2	0.998	1.304	3.934	0.999	0.705	0.695
NOR3	0.998	2.121	3.365	0.999	0.620	0.583
NAND2	0.999	0.506	1.923	0.998	4.198	1.736
NAND3	0.998	0.438	1.983	0.999	10.246	0.669
XOR2	0.993	0.918	3.832	0.999	0.358	0.221
FA	0.913	233.596	5.285	0.991	275.399	2.397

TABLE II
PERFORMANCE VERIFICATION OF OPTIMAL SIZING OVER PVT VARIATIONS

Gate	Delay		Leakage	
	Average	Maximum	Average	Maximum
AND2	3.55%	17.60%	6.33%	7.63%
AND3	4.59%	19.10%.	7.66%	7.76%
NOR2	3.28%	17.51%	2.53%	2.94%
NOR3	3.96%	17.44%	7.18%	13.84%
NAND2	2.90%	13.52%	1.16%	4.77%
NAND3	8.63%	32.26%	6.36%	8.91%
XOR2	6.34%	17.38%	0.97%	1.38%
FA	31.30%	36.74%	7.22%	18.75%

IV. RESULTS

The performance of the proposed ResNN to approximate delays and leakages is measured through three evaluation metrics - coefficient of determination (R^2), mean squared error (MSE) and mean absolute percentage error (MAPE). The modeling results are tabulated in Table I. We were able to model all leakages and delays with high R^2 scores upto 0.99. The MSE for delay was of the order of magnitude 10^{-24} and for delay was of the order 10^{-16}. MAPE scores ranged between 1.786% and 5.285% for delays and between 0.221% and 2.397% for leakages, with the highest errors occurring for Full Adder due to its complex transistor network with 28 transistors. Increased training data would further reduce the errors in such complex cells. After training the ResNN model, genetic algorithm based multi-objective optimization was performed. The objective was to reduce both delay and leakage simultaneously for HP 22nm MGK standard cells. The percentage reduction achieved for delays and leakages in each cell are shown in Fig 3. The convergence of the

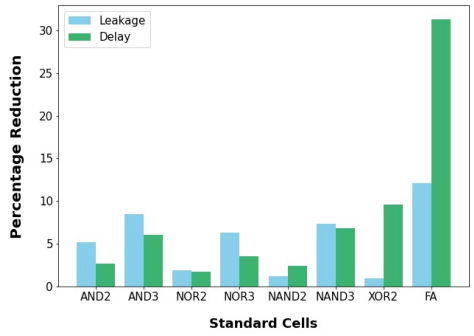

Fig. 3. % Reduction in leakage and delay through optimized transistor sizing

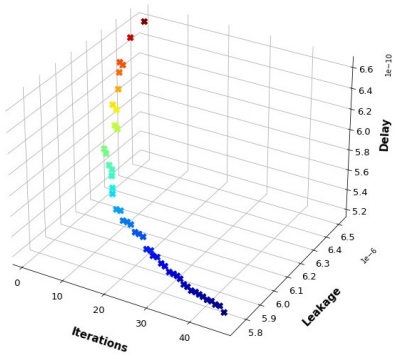

Fig. 4. The convergence of Genetic algorithm with iterations (generations) resulting in reduced delay and leakage for the Full Adder gate.

genetic algorithm with iterations, resulting in reduced delay and leakage for the Full Adder circuit is plotted in Fig. 4.The delay at average PVT values for optimal sizing was 31.33% reduced compared to initial sizing, and the leakage reduced by 12.13%.

Once the optimal sizing was achieved, performance verification was performed over PVT variations. Again optimal sizing delay and leakage were compared with their initial sizing counterparts and the results are listed in Table II. The average reduction in delay for Full Adder was 31.30% whereas the maximum was 36.74%. Similarly, the average reduction in leakage for FA was 7.22% with the maximum being 18.75%. The results indeed prove that the optimal sizing gate dimensions are robust to PVT variations, consistently performing better than the initial sizing. These robust standard cells can be further adapted in digital circuits resulting in PVT-aware optimization. Table III compares our work with previous works. The robustness of the proposed work is evident by reduction of leakage and delay for HP circuits at the same time as shown in table III with very minimal computational time compared to other proposed techniques.

V. Conclusion

In this paper we present a fast and efficient pipeline to achieve PVT aware performance improvement in digital circuits. The speedup achieved is attributed to our use of the trained ResNN model which can return delay and leakage values in a fraction of a second as a proxy for the time consuming HSPICE simulations. Genetic algorithm further optimizes the gate dimensions to return robust PVT aware Pareto-optimal transistor sizing.

TABLE III
Comparison of our work with previous methods

	[6]	[10]	[11]	[11]	Our work
Algorithm	ABC, PCO	SMO	GSO, NCGA	GSO, NCGA	GA
Technology node	45nm LP ptm	22nm/ 45nm LP ptm	22nm LP ptm	22nm HP ptm	22nm HP ptm
% optimi. leakage	45	58/64	50	NA	19
% optimi. delay	-5.1	NA	NA	43	37
Simulator	HSPICE	HSPICE	HSPICE	HSPICE	ResNN
Time (min)	∼90	∼30	∼30	∼300	∼2

References

[1] Carballo et al., "Itrs 2.0: Toward a re-framing of the semiconductor technology roadmap," in *2014 IEEE 32nd ICCD*, pp. 139–146, IEEE, 2014.

[2] S. Bhunia and S. Mukhopadhyay, *Low-power variation-tolerant design in nanometer silicon*. Springer, 2010.

[3] A. Beg, "Automating the sizing of transistors in cmos gates for low-power and high-noise margin operation," *IJCTA*, vol. 43, no. 11, pp. 1637–1654, 2015.

[4] K.Jeong et al., "Revisiting the linear programming framework for leakage power vs. performance optimization," in *2009 10th ISQED*, pp. 127–134, IEEE, 2009.

[5] F. Kashfi, S. Hatami, and M. Pedram, "Multi-objective optimization techniques for vlsi circuits," in *2011 12th ISQED*, pp. 1–8, IEEE, 2011.

[6] P. Gupta et al., "Pvt variations aware robust transistor sizing for power-delay optimal cmos digital circuit design," in *ISCAS*, pp. 1–5, IEEE, 2019.

[7] K. K. Kim and Y.-B. Kim, "A novel adaptive design methodology for minimum leakage power considering pvt variations on nanoscale vlsi systems," *IEEE transactions on VLSI systems*, vol. 17, no. 4, pp. 517–528, 2009.

[8] Z. Abbas and M. Olivieri, "Optimal transistor sizing for maximum yield in variation-aware standard cell design," *IJCTA*, vol. 44, no. 7, pp. 1400–1424, 2016.

[9] Z. Abbas, M. Olivieri, and A. Ripp, "Yield-driven power-delay-optimal cmos full-adder design complying with automotive product specifications of pvt variations and nbti degradations," *Journal of Computational Electronics*, vol. 15, no. 4, pp. 1424–1439, 2016.

[10] P. Saha, H. S. Kalluru, and Z. Abbas, "Transistor sizing based pvt-aware low power optimization using swarm intelligence," in *2021 34th VLSID and 2021 20th International Conference on Embedded Systems*, pp. 234–239, IEEE, 2021.

[11] H. Kalluru et al., "Algorithm driven power-timing optimization methodology for cmos digital circuits considering pvta variations," in *2021 IEEE ISCAS*, pp. 1–5, IEEE, 2021.

[12] B. Model, "http://www-device.eecs.berkeley.edu/bsim/."

[13] P. T. Model, "http://ptm.asu.edu/."

[14] D. Amuru et al., "Statistical variation aware leakage and total power estimation of 16 nm vlsi digital circuits based on regression models," in *VDAT*, pp. 565–578, Springer, 2019.

[15] D. Amuru, M. S. Ahmed, and Z. Abbas, "An efficient gradient boosting approach for pvt aware estimation of leakage power and propagation delay in cmos/finfet digital cells," in *2020 IEEE ISCAS*, pp. 1–5, 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.