Fast Machine Learning Based Prediction for Temperature Simulation Using Compact Models

Mohammadamin Hajikhodaverdian*, Sherief Reda†, Ayse K. Coskun*
* Boston University - (aminhaji, acoskun)@bu.edu
† Brown University - sherief_reda@brown.edu

***Abstract*—As transistor densities increase, managing thermal challenges in 3D IC designs becomes more complex. Traditional methods like finite element methods and compact thermal models (CTMs) are computationally expensive, while existing machine learning (ML) models require large datasets and a long training time. To address these challenges with the ML models, we introduce a novel ML framework that integrates with CTMs to accelerate steady-state thermal simulations without needing large datasets. Our approach achieves up to $70\times$ speedup over state-of-the-art simulators, enabling real-time, high-resolution thermal simulations for 2D and 3D IC designs.** [1]

***Index Terms*—Thermal simulation, Compact thermal models (CTM), Machine learning, 3D IC**

I. Introduction

Thermal simulation is critical in integrated circuit (IC) design, particularly with the shift toward 3D architectures. As transistor densities increase, heat dissipation becomes a significant challenge, leading to hot spots and thermal gradients that can degrade chip performance, reliability, and lifespan. Therefore, thermal analysis during chip design is an essential task [1]. Traditional methods for thermal analysis, such as finite element methods (FEM), offer high accuracy. However, these computational methods require substantial memory, time, designs, and fine-granularity simulation of 3D architectures. Compact Thermal Models (CTMs) provide a more efficient alternative. While CTMs significantly reduce computational requirements, they rely on numerical solvers. This makes CTMs computationally expensive for applications with dynamic workloads or frequently changing designs.

Recent advancements in machine learning (ML) have introduced alternative methods for thermal simulation. ML models can predict temperature distributions directly, offering significant speedups compared to numerical solvers. However, existing ML approaches suffer from critical limitations. These models require large datasets for training, rely on complex architectures like convolutional neural networks (CNNs) or graph neural networks (GNNs), and often lack adaptability to new floorplans or power distributions without retraining [2].

This work addresses these limitations by designing a lightweight ML framework that integrates with CTMs. Our approach leverages the linear nature of the heat conduction equation to develop a simple, physically informed model using linear regression. By embedding the problem's physics directly into the ML framework, we minimize the need for large datasets and retraining. Furthermore, we introduce a window-based scalability technique to handle large grid sizes efficiently, preserving accuracy while reducing computational overhead. This method is robust to changes in floorplans and architectures, requiring only a few additional samples for adaptation. Our contributions are summarized below:

- A linear regression model that aligns directly with the steady-state thermal conduction problem, eliminating the need for complex architectures or extensive training data.
- A method for splitting large grids into smaller overlapping windows, ensuring efficient computation and accurate predictions for high-resolution simulations.

We achieve up to $70\times$ speedup over state-of-the-art simulators like PACT [3] while maintaining accuracy across diverse IC designs, including challenging 3D architectures. By combining the efficiency of CTMs with the adaptability of ML, this framework offers a practical solution for fast, accurate thermal simulation, making real-time analysis in advanced IC designs.

II. Related Work

Machine learning (ML)-based methods have emerged as promising alternatives for thermal simulation. Some studies have used infrared (IR) imaging for thermal predictions, but these are limited to post-silicon stages and require costly equipment [4]. Simulation-based ML approaches, such as convolutional neural networks (CNNs) and graph neural networks (GNNs), enable pre-silicon analysis by predicting temperature distributions directly from power maps. However, these methods demand large datasets and retraining for new designs, increasing complexity and computational overhead [2], [5].

Our work bridges the gap between CTMs and ML-based methods by introducing a lightweight, physically-informed ML framework. Unlike prior approaches, our method leverages the inherent linearity of the thermal conduction problem, requiring minimal training samples and avoiding retraining. Furthermore, our window-based scalability technique ensures efficiency for large-scale simulations without sacrificing accuracy.

III. Steady-state Temperature Prediction using Machine Learning and CTMs

In this work, we address the computational limitations of traditional thermal solvers and rethink the need for complex ML models by focusing on the steady-state thermal conduction problem's inherent linearity. The relation between power (P) and temperature (T) in CTM, which is based on the duality

[1]This research was partially funded by the NSF CCF 2131127 grant

978-3-9826741-0-0/DATE25/© 2025 EDAA

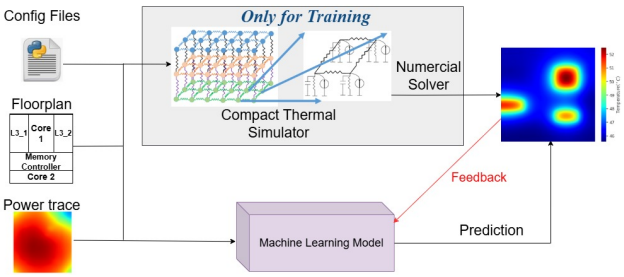

Fig. 1. Framework of the designed method: The machine learning model is trained using power traces and temperature maps from the compact thermal simulator. Once trained, the model begins bypassing the simulator, directly predicting temperature.

of thermal and electric characteristics, is governed by a linear equation, $GT = P$, where G is the thermal conductance matrix. Instead of using a numerical solver to solve the linear equation for every new P, as traditional compact thermal simulators (CTS) do, we suggest estimating G^{-1} with machine learning models. Since $T = G^{-1}P$, having G^{-1} makes it easy to calculate the temperature for any P. A simple idea would be to compute the inverse of G directly, but as the grid size grows, the number of elements in the G matrix exponentially increases, making the explicit matrix inversion computationally expensive.

In our workflow (Fig. 1), input data includes configuration files, floorplans, and power traces. The process has two phases: training and inference. During training, CTS solves the thermal conduction equation from randomly selected power maps to generate accurate temperature maps and train the ML model as the G matrix remains consistent within the same architecture. Once the model meets a predefined error threshold (e.g., mean squared error), it transitions to inference, bypassing CTS's numerical solver. The main advantage of the ML model is the capability to predict temperatures quickly and without numerical simulations.

While the solution using linear regression is straightforward, with a large grid size (e.g., 512×512), the number of model parameters would increase exponentially, leading to memory issues and increased computation. As a scalable solution for handling larger grid sizes, we design a windowing technique that splits the grid into smaller windows (e.g., 64×64). This technique remains effective since the conductance matrix is often uniform across a design. Using the windowed method, we can achieve high accuracy and fast training and inference even for fine-resolution simulations.

IV. Evaluation

Our method was evaluated on 2D and 3D IC designs with varying grid sizes and architectural complexities, demonstrating its scalability, accuracy, and efficiency compared to the state-of-the-art PACT simulator. The total power of 2D designs and each active layer in 3D designs does not exceed $140W$, and the footprint of all designs is $16.8mm \times 14.64mm$.

A. Performance Speed-up

Fig. 2 illustrates the speed-up achieved by our ML framework over PACT across different configurations. For 2D designs,

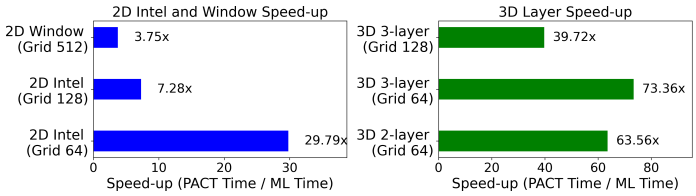

Fig. 2. Comparing the speedup of our method over all different simulations, we consistently improve the simulation time.

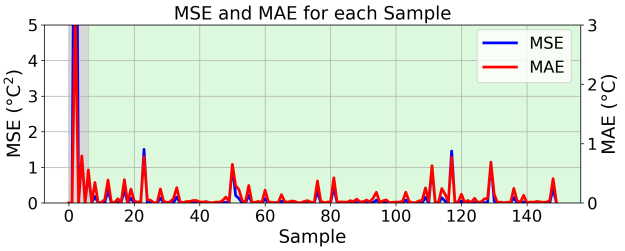

Fig. 3. Mean squared error (MSE) and mean absolute error (MAE) error for 3-layer 3D architecture with a grid size of 128×128 over 140 different samples

Intel i7 $6950\times$ processor block level floorplan, our method achieved a speed-up of up to $29.79\times$ for a 64×64 grid. The window-based approach maintained $3.75\times$ speed-up for larger grids of 512×512, respectively. Similarly, for 3D designs, a monolithic 3D architecture with one computation layer and one to two memory layers, our framework achieved up to $73.36\times$ speed-up for a 3-layer architecture with a 64×64 grid, showcasing its efficiency in handling complex 3D scenarios.

B. Accuracy

Fig. 3 shows the accuracy of our approach for a 3D 3-layer design. The gray area in the figure corresponds to the training phase, with only 6 samples used for training. The method keeps the error low for a 3D architecture. It maintains mean absolute error (MAE) on inference not higher than $1°C$, showing it effectively generalizes with a small number of training samples while keeping the prediction quality and robustness of our model even under steep temperature gradients.

References

[1] Y. Sun, C. Zhan, J. Guo, Y. Fu, G. Li, and J. Xia, "Localized thermal effect of sub-16nm finfet technologies and its impact on circuit reliability designs and methodologies," in *2015 IEEE International Reliability Physics Symposium*, 2015, pp. 3D.2.1–3D.2.6.

[2] L. Chen, W. Jin, and S. X.-D. Tan, "Fast thermal analysis for chiplet design based on graph convolution networks," in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2022, pp. 485–492.

[3] Z. Yuan, P. Shukla, S. Chetoui, S. Nemtzow, S. Reda, and A. K. Coskun, "Pact: An extensible parallel thermal simulator for emerging integration and cooling technologies," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 4, pp. 1048–1061, 2021.

[4] J. Lu, J. Zhang, and S. X.-D. Tan, "Real-time thermal map estimation for amd multi-core cpus using transformer," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–7.

[5] R. Ranade, H. He, J. Pathak, N. Chang, A. Kumar, and J. Wen, "A thermal machine learning solver for chip simulation," in *Proceedings of the 2022 ACM/IEEE Workshop on Machine Learning for CAD*, ser. MLCAD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 111–117. [Online]. Available: https://doi.org/10.1145/3551901.3556484