Fast SoC thermal simulation with physics-aware U-Net

Yu-Sheng Lin* **Li-Song Lin*** **Chin-jui Chang*** Ting-Yu Lin Ya-Wen Yu
Shih-Hong Pan Wei Cheng Lee Kai-En Yang Yi-Chen Lin Tai-Yu Chen
Jason Yeh

Mediatek Inc.

{bob-ys.lin, li-song.Lin, chin-jui.chang} @mediatek.com
{tim-ty.lin, yw.yu, henry.pan, wei-cheng.lee} @mediatek.com
{ds_kyan_ke.yang, yi-chen.lin, kidd.chen, jason.yeh} @mediatek.com

Abstract

Fast thermal simulation for System on Chip (SoC) plays a crucial role in integrated circuit (IC) design industry, particularly as power density escalates with increasing computational requirements. It is imperative to assess thermal performance comprehensively during the design phase, utilizing a rapid and precise thermal simulator to expedite design iterations. In this paper, we introduce a fast, physics-aware thermal simulator that draws inspiration from Fourier's law and the Fourier-Biot equation, which correspond to the first and second derivatives of the temperature map. Consequently, the learning objective evolves from merely translating images to approximating natural phenomena such as the thermal gradient and thermal laplacian. By replacing the image-based loss with thermal-aware loss, the proposed model achieves lower prediction error, higher data efficiency, and more physically accurate behavior. The present model demonstrates a significant improvement, achieving a 34% reduction in Maximum Temperature Error (MTE), showcasing the potential for integrating physics-aware learning into SoC thermal design.

1 Introduction

The growing demand for high performance in mobile, 5G and AI computing applications is increasing the criticality and challenge of thermal management design. Among these, SoC thermal design stands as the most critical factor [4, 9].

High temperatures can lead to CPU throttling or overheating, which in turn results in decreased device performance and poor user experiences; this issue becomes even more critical with the advent of 3D stacked chiplets [6]. Furthermore, the complexities of SoC Interlecture Property (IP) placement design, which involves various target IPs and multiple physical constraints such as thermal, IR and timing, lead to an extensive design of experiments (DOE). Additionally, the IC industry faces significant time constraints, and conventional thermal simulation methods using Computational Fluid Dynamics (CFD) tools are highly time-consuming. Typically, it takes dozens of minutes to a few hours to perform steady state thermal simulation with CFD tools. In response to these challenges, there is an urgent need for a method that is able to provide immediate feedback from power input to temperature output. Accelerating the simulation process allows for the assessment of a wide variety of floorplan candidates, thereby fostering innovative, thermally-aware floorplan design explorations.

*Equal contribution

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

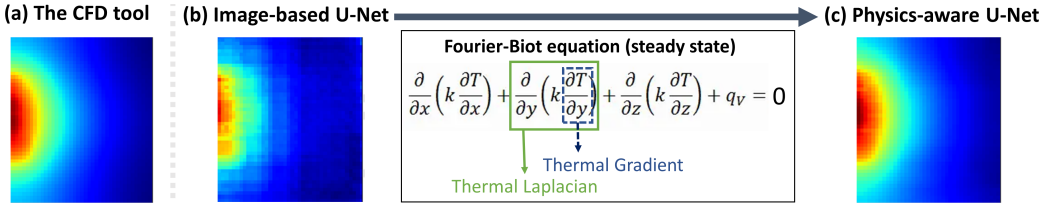

Figure 1: The proposed physics-aware model adopts thermal gradient and thermal laplacian, the key concepts in the heat conduction equation, to achieve the simulation with higher fidelity.

Thermal analysis, in general, can be executed either through empirical experimentation or computational simulations. Within the realm of mobile SoC design, simulations are predominantly employed to attain a more optimized thermal design. Nonetheless, for the exploration of thermally critical floorplan placement designs, examining each individual floorplan scenario and its corresponding power setting often necessitates considerable computational time. Recently, Deep Neural Networks (DNNs) have been used to accelerate CFD tools. The associated research works can be broadly divided into two primary categories: generic-physics models and task-specific ones. Generic-physics models, including neural ODEs [2] and FNO [5], leverage a universal framework underpinned by DNNs to address differential equations. However, these generic models require meticulous provision of domain-specific assumptions by the user, such as boundary conditions, prior to the initiation of the training process. In stark contrast, task-specific models only require paired input and output data for certain application scenario, and the corresponding model can be trained. The work in [7] proposes a thermal solver that handles constant and distributed heat transfer coefficients. The EDGe [3] interprets full-chip thermal simulation as image translation, thereby converting the power map into the temperature map with the U-Net [8] architecture. However, task-specific models may inadvertently overlook the inherent physical nature of a certain task. For example, the EDGe model is trained with the image-based MSE loss only, a fitting loss that does not take into account the continuous relations between two adjacent pixels (cf. Figure 1(b)). As a result, the enhancement of their performance is largely dependent on the availability of extensive training data.

In summary, generic-physics models are dedicatedly crafted by experts, whereas task-specific models usually leverage less domain knowledge and require more training data. In this paper, we integrate physical constraints into task-specific models, resulting in a model that is both accurate and data-efficient. We introduce a novel steady-state thermal simulator for immediate power-to-temperature mapping, offering a speedup of over 100 times compared to the CFD commercial tool.[2] This speedup allows exploration towards optimal floorplan placement. The accuracy of the proposed model is amplified by employing the domain-specific loss design combined with U-Net modification. As a result, the proposed model can achieve lower loss, higher data efficiency, and more realistic physical behavior.

2 Methods

We review the image-based loss $\mathcal{L}_{MSE}$ used in EDGe, derive the physics-aware U-Net[3] from the thermal governing equations, and then illustrate the overall learning objective $\mathcal{L}_{energy}$. In conventional image-to-image translation, the de facto loss function is the mean-squared error in pixel space Ω (1). The following loss terms implicitly assume and thus omit the summation over a batch of samples for simplicity, where $\hat{T}$ is the predicted temperature map and T is the ground truth temperature map.

$$\mathcal{L}_{MSE} = \Sigma_{e\in\Omega}(\hat{T}_e - T_e)^2 \tag{1}$$

As shown in Figure 1(b), the temperature contour generated by the U-Net with image-based loss is not as smooth as the one from the benchmark CFD tool (Figure 1(a)), with some contour discontinuities observed. In physics, the temperature relationship between adjacent grids can be described by

[2]The table of speed comparison is on the appendix section.

[3]We use the EDGe, which is based on U-Net, as the backbone of the proposed model. (github link).

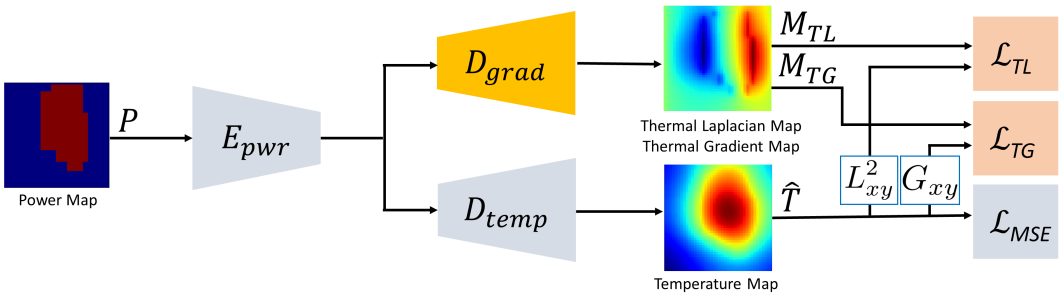

Figure 2: The proposed model decodes thermal gradient and thermal laplacian map, and apply physics-aware loss besides the objective for image translation.

Fourier's law (2), which states that the rate of heat transfer is proportional to the negative temperature gradient.

$$q_v = -k\nabla T \tag{2}$$

where q_v is the energy generation rate per unit volume; k is the thermal conductivity of the material; ∇T is the temperature gradient. In three dimensions, the equation becomes (3), where ρ and c are the density and specific heat of the material, respectively; q is the power source.

$$\frac{\partial}{\partial x}(k\frac{\partial T}{\partial x}) + \frac{\partial}{\partial y}(k\frac{\partial T}{\partial y}) + \frac{\partial}{\partial z}(k\frac{\partial T}{\partial z}) + q = \rho c\frac{\partial T}{\partial t} \tag{3}$$

The Fourier-Biot equation (3) is a general heat conduction equation that describes the energy conservation property in rectangular coordinates. For the proposed method, which targets at steady state conditions, $\partial T/\partial t$ is zero. The governing equations inspire a physics-aware network architecture (Figure 2) that emulates the properties of the heat conduction equation. The proposed model encodes the power map P with the encoder E_{pwr}, and predicts the temperature and derivative infomation with two decoders. The temperature decoder D_{temp} predicts the temperature map $\hat{T}$, and the Sobel and Laplacian operators, G_{xy} and L^2_{xy}, are convolved with $\hat{T}$ to get the thermal gradient map $\hat{T} * G_{xy}$ and the thermal laplacian map $\hat{T} * L^2_{xy}$. The thermal gradient decoder D_{grad} outputs the first and second derivatives of the temperature map, M_{TG} and M_{TL}. To learn the derivatives in the governing equation, novel loss terms $\mathcal{L}_{TG}$ and $\mathcal{L}_{TL}$ are added. The thermal gradient and thermal laplacian loss terms are given by (4) and (5), where kernel operations by G_{xy} give N-channel tensors that represent gradient maps in N directions; $*$ is the convolution operator; γ_1 and γ_2 are hyperparameters.

$$\mathcal{L}_{TG} = \Sigma_{e\in\Omega}\Sigma_{n=1}^{N}\{[(\hat{T} * G_{xy})_{e,n} - (T * G_{xy})_{e,n}]^2 + \gamma_1 \times [M_{TG,e,n} - (T * G_{xy})_{e,n}]^2\} \tag{4}$$

$$\mathcal{L}_{TL} = \Sigma_{e\in\Omega}\{[(\hat{T} * L^2_{xy})_{e} - (T * L^2_{xy})_{e}]^2 + \gamma_2 \times [M_{TL,e} - (T * L^2_{xy})_{e}]^2\} \tag{5}$$

In combination, the final loss function is defined in (6), where α, β are hyperparameters.

$$\mathcal{L}_{energy} = \mathcal{L}_{MSE} + \alpha\mathcal{L}_{TG} + \beta\mathcal{L}_{TL} \tag{6}$$

3 Experiments

The dataset consists of 1950 paired images of power maps and temperature maps. We split the training set of size 1500 and the testing set of size 450. For the experiments of data efficiency (Figure 3), subsets are sampled from the training set for training. All the models shown in the ablation study (Table 1) are trained with a subset of 1000 samples. Following the setting in EDGe, each pixel in the image represents an area of size $0.25mm \times 0.25mm$, and the size of each image is 60 pixels $\times$ 60 pixels. Each power map contains randomly spawned rectilinear shapes that are filled with power values sampled from a uniform distribution, and the ground truth temperature maps are parsed from the simulation results of Ansys-Icepak [1]; to prevent overfitting, data augmentations such as random

Table 1: Ablation study of the physics-aware loss. Evaluation of prediction performance using MSE and MTE, with MSE_{TG} and MSE_{TL} indicating physics fidelity.

Error \Model		Image-based	Physics-aware	
Metrics	Definition	$\mathcal{L}_{MSE}$	$\mathcal{L}_{MSE} + \mathcal{L}_{TG}$	$\mathcal{L}_{MSE} + \mathcal{L}_{TG} + \mathcal{L}_{TL}$
MSE	$(\hat{T} - T)^2$	0.82	0.72 (-12%)	**0.58 (-29%)**
MTE	$\max(\lvert\hat{T} - T\rvert)$	2.24	1.87 (-16%)	**1.48 (-34%)**
MSE_{TG}	$[(\hat{T} - T) * G_{xy}]^2$	0.15	0.09 (-40%)	**0.06 (-60%)**
MSE_{TL}	$[(\hat{T} - T) * L^2_{xy}]^2$	0.32	0.17 (-47%)	**0.08 (-75%)**

rotation and random flip are applied during training. The image-based baseline model applies the U-Net and loss of EDGe, and the physics-aware model refers to the modified U-Net trained with $\mathcal{L}_{energy}$, batch size of 64, and 720 epochs at a learning rate of 0.002 with the Adam optimizer. We investigate the effectiveness of the proposed method, the data efficiency, and the learned physical phenomona. The initial two rows in Table 1 demonstrate the superior performance of the proposed physics-aware U-Net over the image-based baseline, as evidenced by an approximate 30% reduction in loss. The MSE (Mean Squared Error) is used to measure the quality of the generated image. The MTE (Maximum Temperature Error) is employed to identify the most significant pixel error in a temperature prediction. Metrics in Table 1 are pixel-averaged per prediction, with mean values reported over the testing set. We scrutinize the data efficiency of the proposed model by reducing the training set size. Given the lengthy data collection process with CFD tools, a model with equivalent error rates using less data is preferred. Comparing testing errors of models trained with 250 and 500 samples, as depicted in Figure 3, it can be concluded that the proposed model achieves similar error rates with half the amount of the data, demonstrating its high data efficiency, further emphasizing the benefits of embedding physical insights into model training, especially when there is a paucity of data. In Figure 4, We visualize the thermal gradient map to assess if the proposed model has learned realistic physical phenomena. The thermal gradient contour of the physics-aware model exhibits smoother characteristics when contrasted with the image-based model. The MSE_{TG}, and MSE_{TL} in Table 1 represent the MSE of the predicted thermal gradient and thermal laplacian maps. The smaller errors of the proposed models suggest more realistic simulations than the image-based model.

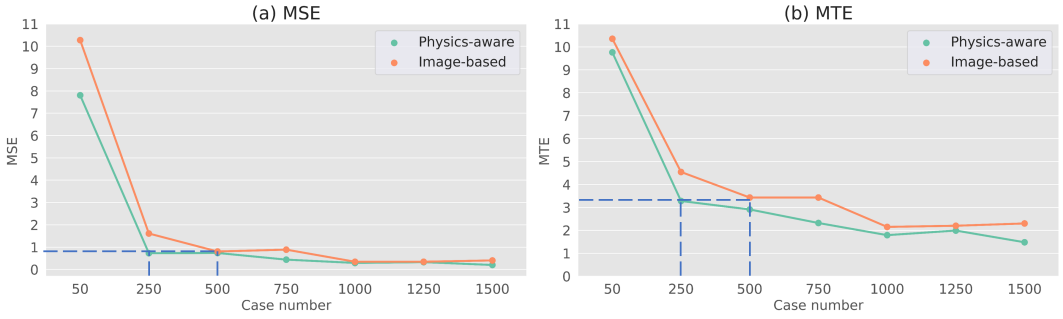

Figure 3: Averaged testing errors for models trained with varying sizes of training sets.

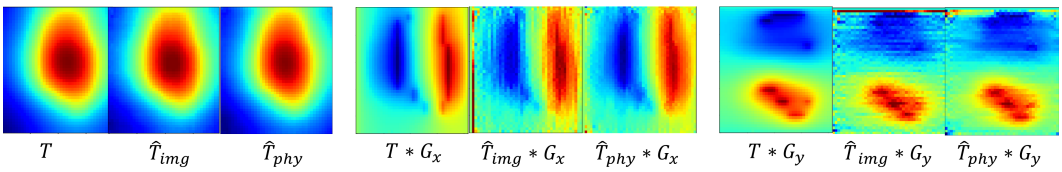

Figure 4: Visualization of the learned thermal gradient maps. The $\hat{T}_{img}$ and $\hat{T}_{phy}$ represent the temperature predictions made by the image-based model and the physics-aware model, respectively. The symbol G signifies the Sobel kernel, with the subscript denoting the dimensionality.

4 Conclusions

This paper introduces a physics-aware thermal simulator for SoC design, which attains enhanced prediction accuracy, data efficiency, and fidelity to physical behavior. The proposed model, based on the U-Net architecture and inspired by Fourier's law and the Fourier-Biot equation, achieves a 34% reduction in MTE, underscoring the potential of integrating physics-aware learning into thermal simulation. Furthermore, the SoC thermal simulator can achieve a runtime reduction of approximately 100 times compared to the conventional CFD tool. The results open up new avenues for research directions and have far-reaching implications for the semiconductor industry, particularly in the context of increasing computational demands and power density. Future research will expand this model to other simulations and optimize its performance, paving the way for advanced physics-aware machine learning models for thermal simulation and beyond.

Acknowledgements

This research is supported by MediaTek Inc. Special thanks are given to Cheng-Che Lee and Wu Yu Hsuan for their contributions to this work. Additional gratitude is extended to the team at MediaTek Inc. for their support and guidance.

References

[1] Ansys-icepak.

[2] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[3] Vidya A. Chhabria, Vipul Ahuja, Ashwath Prabhu, Nikhil Patil, Palkesh Jain, and Sachin S. Sapatnekar. Thermal and ir drop analysis using convolutional encoder-decoder networks. In *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 690–696, 2021.

[4] Sheng-Liang Kuo, Chi-Wen Pan, Pei-Yu Huang, Chien-Tse Fang, Shin-Yu Hsiau, and Tai-Yu Chen. An innovative heterogeneous soc thermal model for smartphone system. In *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 384–391, 2018.

[5] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2020.

[6] Rahul Mathur, Chien-Ju Chao, Rossana Liu, Nikhil Tadepalli, Pranavi Chandupatla, Shawn Hung, Xiaoqing Xu, Saurabh Sinha, and Jaydeep Kulkarni. Thermal analysis of a 3d stacked high-performance commercial microprocessor using face-to-face wafer bonding technology. In *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, pages 541–547, 2020.

[7] Rishikesh Ranade, Haiyang He, Jay Pathak, Norman Chang, Akhilesh Kumar, and Jimin Wen. A thermal machine learning solver for chip simulation. In *Proceedings of the 2022 ACM/IEEE Workshop on Machine Learning for CAD*, MLCAD '22, page 111–117, New York, NY, USA, 2022. Association for Computing Machinery.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[9] Jongkyu Yoo, Taekeun An, Chigwan Oh, Youngsang Cho, Heeseok Lee, Yunhyeok Im, Minkyu Kim, and Minsu Kim. Thermal-aware optimization of soc floorplan with heterogenous multi-cores. In *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, pages 1–6, 2022.

Appendix

Table 2: Comparative analysis of the CFD Tool and the proposed model

Model	Node/Pixel	Runtime/case	Speedup	CPU cores	GPU
CFD tool (Ansys-Icepak)	1.3M nodes	$\approx$ 30 mins	1x	4	None
The proposed model	3600 pixels	$\approx$ 3 ms	**> 100x**	4	RTX 2080 Ti x 1