Fast Thermal-driven 3D Fixed-outline Floorplanning By Learning-based Thermal Analysis

Yikai Liu, Jindong Zhou, Jiayi Li, Pingqiang Zhou
School of Information Science and Technology, *ShanghaiTech University*
Shanghai, China
{liuyk2023,zhoujd,lijy22023,zhoupq}@shanghaitech.edu.cn

2025 IEEE 16th International Conference on ASIC (ASICON) | 979-8-3315-3917-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ASICON66040.2025.11326217

***Abstract*—Higher integration density in 3D ICs brings severe thermal issues which potentially degrades the performance and reliability of a 3D chip. To address this problem, thermal-aware optimization is necessary at early floorplanning stage. In this work, we proposed a novel thermal-driven 3D fixed-outline floorplanning algorithm with learning-based fast thermal evaluation method. To integrate the neural network model into the floorplanning framework, we also proposed an adaptive cost function. Experimental results shows that compared with thermal simulation tool *HotSpot* [1], our method can speed up the thermal analysis by about 3700× with high accuracy. Leveraging the accuracy and efficiency of deep neural network, the floorplanner can effectively reduce the peak temperature of the chip.**

***Index Terms*—3D-IC, Floorplan, Thermal Analysis, Deep Neural Networks**

I. Introduction

Three Dimensional (3D) IC technology develops rapidly recently with the maturity of process and manufacturing [2]. While enjoying the advantages of small package area and high bandwidth, the 3D systems are suffering from severe thermal problems brought by the die stacking [3]–[5], as power density increases linearly with the number of stacked tiers.

Many investigations have been conducted to alleviate the thermal problems of Integrated Circuits (IC) [6] in physical design, by arranging the layout of the circuit modules to reduce the peak temperature. For 3D circuits, thermal-aware floorplanning and placement have also been explored [7]–[9]. The general idea is to adjust the locations of modules or macros according to the results of thermal analysis after each iteration.

In our work, we focus on 3D floorplanning. During the iterative process of 3D fixed-outline floorplanning, *it is complex and time consuming to do 3D thermal analysis*. Accurate evaluation methods like Finite Element Method (FEM) and classical thermal analysis tool *HotSpot* [1] involve complex matrix solving, which consumes massive amount of time. When it comes to a multi-tier 3D circuit, the analysis time is even longer. However, typical optimization process requires thousands of iterations, thus makes accurate thermal analysis impractical. To address this problem, existing methods usually use simplified approximation thermal models to guide the optimization, including green function method [10], power-blurring method (PB) [11], [12], or directly using the power features of each block like vertical-heat-flow method (VHF) [13]–[15] as an indicator of the thermal performance.

This work is supported by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 24JD1402500.

In recent years, deep learning has shown its efficiency in fitting complex functions with powerful feature extraction and regression capabilities. Many researches have adopted deep learning to accelerate the thermal evaluation [16]–[19] of a 2D or 3D IC. Results show that a well-trained model can greatly accelerate the expensive thermal analysis process and accurately predict a temperature distribution.

In our work, we accelerate the thermal-driven 3D floorplanning process by incorporating learning-based thermal analysis method. The contribuitons of this work include:

- We develop a thermal-aware fixed-outline floorplan framework for 3D-IC, leveraging sequence pair as the floorplan representation and simulated annealing as the optimization method.
- We use deep neural networks to estimate the final steady-state temperature distribution from 3D-IC layout and power distribution in each iteration of 3D floorplanning.
- We evaluate the two-tier 3D-IC on GSRC benchmarks. The results show that our method can reduce the maximum temperature by 11.26K on the largest case. The learning-based model can accelerate the thermal evaluation by about 3700× compared with *HotSpot* simulation.

II. Our Work

A. Learning-based Thermal Analysis

To avoid the huge time cost of traditional thermal evaluation methods like *HotSpot* or FEM while achieving high accuracy, we propose a novel neural network based regression model to evaluate the steady-state temperature distribution of a 3D chip. In order to capture the relationship between power distribution and temperature effectively, we choose U-Net [20] architecture as the backbone of our model.

1) Temperature distribution prediction: For each tier, the input floorplan is first converted into a power map P. We divide the floorplan region into $N \times N$ grids, and the power at grid (i, j) is calculated by

$$p_{ij} = \sum_{b \in B} Overlap_{ij}(b) \times PowerDensity(b)$$

where B is the set of all blocks, $Overlap_{ij}(b)$ is the overlap area between grid (i, j) and the block b. Then, the power map is fed into the trained model and the final output temperature

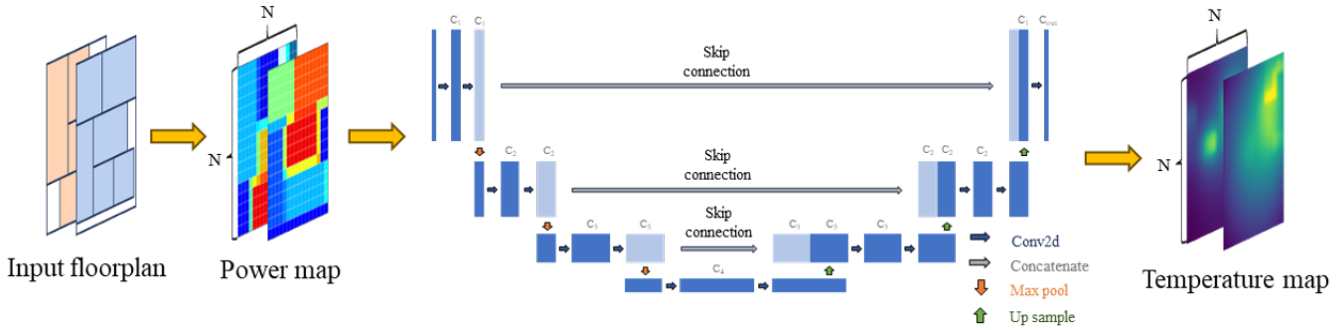


Fig. 1. The proposed thermal analysis flow and the U-Net network architecture .

map size is $N \times N$. The proposed flow and the network structure of our U-Net is shown in Fig. 1. It's worth noting that the model size is determined by the intermediate channel number $[C_1, C_2, C_3, C_4]$. To enable the thermal evaluation in the optimization process, a trade-off is made between the size and accuracy of the model which will be discussed in Section III.

2) Dataset Generation: In order to train the U-Net model for thermal analysis, we employ *HotSpot* to generate the groundtruth temperature distribution of a given 3D chip. The chip structure is shown in Fig. 2. In this work, we modeled

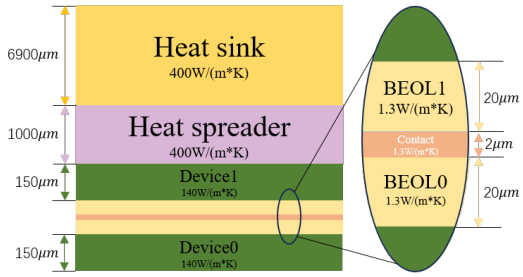


Fig. 2. 2-tier face-to-face chip structure.

a 2-tier face-to-face 3D system. The whole structure has seven layers: Device0, BEOL0, Contact, BEOL1, Device1, heat spreader and a heat sink.

One key procedure in dataset generation is to produce random legal floorplans. We develop an algorithm based on the slicing tree representation [21], where the floorplan can be viewed as the result of recursively applying horizontal and vertical cuts to a rectangle. The detail of the recursive process is described by algorithm 1, where the strategy set S = {CutLeft, CutRight, CutTop, CutBottom, BisectHorizontal, BisectVertical}, CutX means discard the corresponding part of current node. Fig. 3 gives an example of how these strategies are applied to a node. min_ratio controls the minimal block area and k controls the average block size. A larger k will cause the bisection stop earlier and lead to a larger average block size.

Algorithm 1 recursive_bisection

Input: node, S;

area = node.width $\times$ node.height
ratio = area / total_chip_area
prob = $\max(\frac{\text{ratio-min\_ratio}}{\text{1-min\_ratio}}, 0)^k$
if random() $<$ prob **then**
  strategy = random_choice(S)
  new_nodes = apply_strategy(node, strategy)
  for new_node in new_nodes **do**
    recursive_bisection(new_node, S)
    node.children.append(new_node)

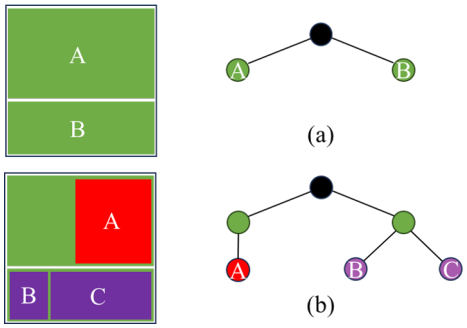


Fig. 3. An example of floorplan generation. (a) The original floorplan and tree representation. (b) The floorplan after applying CutLeft strategy on tree node A and BisectHorizontal on tree node B.

B. Thermal-driven Fixed-outline Floorplanning

The thermal-aware fixed-outline floorplanning problem can be formulated as follows. Given a set of M blocks $B = \{b_i | 1 \leq i \leq M\}$ with width w_i and height h_i, the goal is to find the position x_i, y_i and the orientation o_i of each block, such that $0 \leq x_i \leq W - w_i$, $0 \leq y_i \leq H - h_i$ and the blocks should not overlap with each other. Here W, H are the width and height of the chip. In the meanwhile, the wirelength and peak temperature should be optimized.

In this work, we use a simulated-annealing based method with multi-tier sequence pair representation to solve the problem. The cost function of simulated-annealing method can be written as

$$cost = WL + \lambda C_T + \mu C_A$$

where WL is the total HPWL of the circuit, C_T is the thermal cost and C_A is the area cost. Here, we use the total exceeding width and height as the penalty to the fixed-outline constraint, i.e. $C_A = \max(0, W_f - W) + \max(0, H_f - H)$. W_f and H_f are the width and height of the bounding rectangle of the floorplan. λ, μ control the penalty strength.

However, special caution should be taken when adding the temperature cost C_T. In our learning-based method, the peak temperature can only be obtained when the floorplan satisfies the fixed-outline constraint. However, this constraint is often violated during the early stage of the annealing process. Thus, the actual cost function should be:

$$cost = \begin{cases} WL + \lambda C_T & \text{if constraint is met} \\ WL + \lambda C_{T0} + \mu C_A & \text{otherwise} \end{cases} \quad (1)$$

As a result, a proper value of the constant C_{T0} needs to be assigned before getting the first legal floorplan.

In fact, the value of C_{T0} cannot be chosen arbitrarily. As shown in Fig. 4, a high C_{T0} will enforce the solver to search only within feasible solutions, while a low C_{T0} may fail to find a solution if the peak temperature of feasible floorplan is higher. This is mainly caused by the "discontinuity" of the cost function in the vicinity of feasible region boundary. To address the problem, we use an adaptive cost function to alleviate the discontinuity:

$$cost = \begin{cases} WL + \lambda(C_T - C_{init}), & \text{if constraint is met} \\ WL + \mu C_A, & \text{otherwise} \end{cases} \quad (2)$$

Instead of adding λC_{T0} to the cost of unsatisfied case, we decrease the cost of the satisfied case by λC_{init}, where C_{init} is the peak temperature of the first feasible solution during the optimization process.

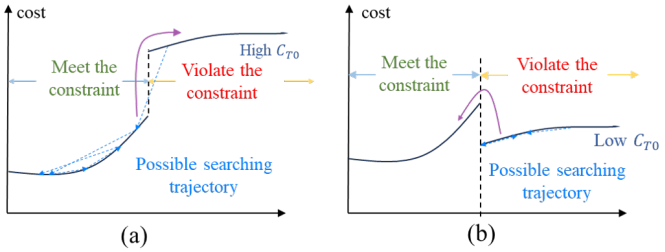


Fig. 4. Improper C_{T0} choice. (a) A high C_{T0} will cause the solver to only search within feasible solutions. (b) A low C_{T0} might fail to find a feasible solution.

III. Experimental Results

A. Experiment setup

We implement the proposed floorplanning algorithm in C/C++ on a Linux environment using an Intel i9-13900K processor. The neural network model is trained with PyTorch on an RTX 4070 GPU. We evaluate our algorithm using the GSRC [22] benchmark suite. For each benchmark, we reserve 10% of the total block area as whitespace. The width and height of the 2-tier chip are calculated as:

$$W = H = \sqrt{\frac{\text{total\_block\_area} \times (1 + 0.1)}{2}}.$$

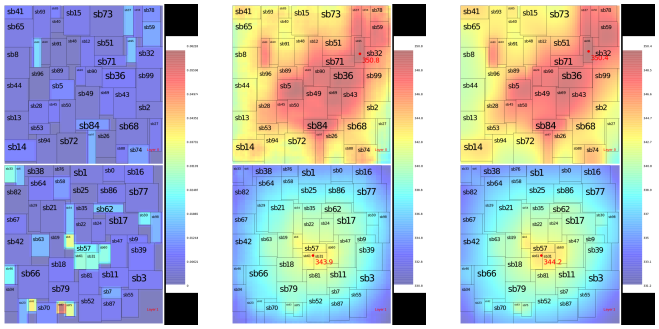
The grid size used for both the power map and temperature map is 64×64. The power of each block ranges from 0.05W to 1W.

B. Accuracy and Runtime of the Model

To train the neural network model, we generate 10,000 random legal floorplans as the dataset. The dataset is split into a training set with 7,000 samples, a validation set with 1,500 samples, and a test set with 1,500 samples. To balance runtime and accuracy, we fine-tune the number of intermediate channels in the model to $[C_1, C_2, C_3, C_4] = [4, 8, 16, 24]$, as this configuration achieves the best performance on the validation set with relatively few parameters. The model is evaluated on the test set using root mean square error (RMSE) and mean absolute error (MAE). The test results and runtime are shown in Table I. Compared with *HotSpot* simulation, our model achieves approximately 3700× speedup with acceptable accuracy loss. More importantly, the significant runtime reduction enables peak temperature evaluation during the floorplanning iteration process. Fig. 5 shows a comparison between the thermal map predicted by our U-Net model and the groundtruth thermal map generated by *HotSpot*.

TABLE I
Model Accuracy and Runtime Comparison

Metrics	Our model	HotSpot 7.0 [1]
RMSE	0.56K	groundtruth
MAE	0.39K	
RMSE (peak temperature)	1.34K	
MAE (peak temperature)	0.91K	
runtime (speed up)	0.87ms (3687.4x)	3208ms (1x)



(a) Power map (b) Predicted (c) Groundtruth

Fig. 5. Comparison between the predicted temperature map and groundtruth temperature map of n100.

C. Result of our Thermal-driven 3D Floorplanning Framework

To demonstrate the effectiveness of our algorithm, we compare the 3D floorplan results with 1) a basic wirelength-driven

method and 2) the vertical-heat-flow (VHF) method [14], [15]. The VHF method does not perform full thermal analysis; instead, it estimates peak temperature by leveraging the power map and thermal resistance. While this approach significantly reduces runtime, it may compromise accuracy. For each method, we run 10 trials and report the average to mitigate the randomness inherent in the simulated annealing algorithm. The results are presented in Table II.

TABLE II
AVERAGE WIRELENGTH, PEAK TEMPERATURE AND RUNTIME OF WIRELENGTH-DRIVEN, VERTICAL-HEAT-FLOW(VHF) AND OUR METHOD.

Cases		n30	n50	n100	n200	n300
Wirelength Driven	WL	68292	106814	177384	361051	550347
	Peak T	333.50K	341.06K	355.93K	392.71K	436.28K
	RT	1.38s	10.52s	22.17s	104.05s	151.68s
VHF	WL	73157 (+4865)	108719 (+1878)	181094 (+3710)	367146 (+6095)	558405 (+8058)
	Peak T	330.74K (-2.76K)	337.36K (-3.70K)	353.30K (-2.63K)	388.63K (-4.08K)	429.63K (-6.65K)
	RT	1.72s	12.27s	24.11s	108.20s	157.93s
Our method	WL	71756 (+3464)	108429 (+1615)	181628 (+4244)	368015 (+6964)	560186 (+9839)
	Peak T	330.83K (-2.67K)	337.27K (-3.79K)	350.76K (-5.17K)	384.26K (-8.45K)	425.02K (-11.26K)
	RT	11.62s	62.94s	69.87s	207.32s	279.26s

Our learning-based method effectively reduces peak temperature with minimal wirelength overhead. For small cases, the VHF method can achieve similar temperature reduction as our method. While in larger cases such as n100, n200 and n300, due to the lack of horizontal thermal coupling consideration, the VHF method failes to separate high power density blocks from each other, thereby degrading the solution quality. In contrast, our method demonstrates superior performance, as the neural network effectively captures thermal coupling effects via convolution operations in the U-Net architecture.

Regarding overall runtime, neural network inference accounts for a significant portion. However, the inference overhead remains constant as circuit size increases. Thus, for larger circuits, the relative runtime overhead becomes smaller (e.g., 8.42× for n30, 1.84× for n300).

IV. CONCLUSION

This paper proposed a novel fixed-outline thermal-aware floorplanning algorithm. By introducing a nerual network based thermal evaluation method, accurate thermal map is available for the floorplanner in each iteration, which greatly improves the solution quality. To incorporate this evaluation method into the floorplanner with fixed-outline constraint, we presented an adaptive cost function to help the convergence of the stochastic searching. Experiment result shows that our method can effectively reduce the peak temperature with little wirelength overhead.

REFERENCES

[1] J.-H. Han, X. Guo, K. Skadron, and M. R. Stan, "From 2.5D to 3D chiplet systems: Investigation of thermal implications with HotSpot 7.0," in *IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, 2022, pp. 1–6.

[2] Y. Zhao, L. Zou, and B. Yu, "Physical design for advanced 3D ICs: Challenges and solutions," in *Proceedings of the International Symposium on Physical Design (ISPD)*, 2025, pp. 209–216.

[3] S. S. Sapatnekar, "Addressing thermal and power delivery bottlenecks in 3d circuits," in *Proceedings of Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2009, pp. 423–428.

[4] T. Lu, C. Serafy, Z. Yang, S. K. Samal, S. K. Lim, and A. Srivastava, "TSV-based 3-D ICs: design methods and tools," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 36, no. 10, pp. 1593–1619, 2017.

[5] A. Todri-Sanial and C. S. Tan, *Physical Design for 3D Integrated Circuits*. CRC Press, 2017.

[6] H. Sultan, A. Chauhan, and S. R. Sarangi, "A survey of chip-level thermal simulators," *ACM Comput. Surv.*, vol. 52, no. 2, 2019.

[7] S. K. Samal, S. Panth, K. Samadi, M. Saeidi, Y. Du, and S. K. Lim, "Adaptive regression-based thermal modeling and optimization for monolithic 3-D ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 35, no. 10, pp. 1707–1720, 2016.

[8] J.-M. Lin, W.-Y. Chang, H.-Y. Hsieh, Y.-T. Shyu, Y.-J. Chang, and J.-M. Lu, "Thermal-aware floorplanning and tsv-planning for mixed-type modules in a fixed-outline 3-D IC," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 29, no. 9, pp. 1652–1664, 2021.

[9] P. Zhou, Y. Ma, Z. Li, R. P. Dick, L. Shang, H. Zhou, X. Hong, and Q. Zhou, "3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits," in *Proceedings of International Conference on Computer-Aided Design (ICCAD)*, 2007, pp. 590–597.

[10] S. S.-Y. Liu, R.-G. Luo, S. Aroonsantidecha, C.-Y. Chin, and H.-M. Chen, "Fast thermal aware placement with accurate thermal analysis based on green function," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 22, no. 6, pp. 1404–1415, 2014.

[11] W. Guan, X. Tang, H. Lu, Y. Zhang, and Y. Zhang, "Thermal-aware fixed-outline 3-D IC floorplanning: An end-to-end learning-based approach," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 31, no. 12, pp. 1882–1895, 2023.

[12] A. Ziabari, J.-H. Park, E. K. Ardestani, J. Renau, S.-M. Kang, and A. Shakouri, "Power blurring: Fast static and transient thermal analysis method for packaged integrated circuits and power devices," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 22, no. 11, pp. 2366–2379, 2014.

[13] G. Luo, Y. Shi, and J. Cong, "An analytical placement framework for 3-D ICs and its extension on thermal awareness," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 32, no. 4, pp. 510–523, 2013.

[14] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-aware 3D IC placement via transformation," in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2007, pp. 780–785.

[15] L. Xiao, S. Sinha, J. Xu, and E. F. Young, "Fixed-outline thermal-aware 3d floorplanning," in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2010, pp. 561–567.

[16] L. Chen, J. Lu, W. Jin, and S. X.-D. Tan, "Fast full-chip parametric thermal analysis based on enhanced physics enforced neural networks," in *Proceedings of the International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–8.

[17] Z. Liu, Y. Li, J. Hu, X. Yu, S. Shiau, X. Ai, Z. Zeng, and Z. Zhang, "DeepOHeat: Operator learning-based ultra-fast thermal simulation in 3D-IC design," in *Proceedings of the Design Automation Conference (DAC)*, 2023, pp. 1–6.

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

[19] A. Sridhar, A. Vincenzi, M. Ruggiero, and D. Atienza, "Neural network-based thermal simulation of integrated circuits on gpus," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 31, no. 1, pp. 23–36, 2012.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[21] R. Otten, "Automatic floorplan design," in *Proceedings of the Design Automation Conference (DAC)*, 1982, pp. 261–267.

[22] http://vlsicad.eecs.umich.edu/BK/GSRCbench.