ATPlace2.5D: Analytical Thermal-Aware Chiplet Placement Framework for Large-Scale 2.5D-IC

Qipan Wang[1,2], Xueqing Li[1], Tianyu Jia[1,4], Yibo Lin[1,3,4*], Runsheng Wang[1,3,4], Ru Huang[1,3,4]
[1]School of Integrated Circuits, [2]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing; [3]Institute of Electronic Design Automation, Peking University, Wuxi;
[4]Beijing Advanced Innovation Center for Integrated Circuits
{qpwang, xueqing_li, tianyuj, yibolin, r.wang, ruhuang}@pku.edu.cn

ABSTRACT

The surge in consumer electronics is catalyzing the evolution of 2.5D integrated circuits (2.5D-IC). As these systems expand in scale and integrate more chiplets, the significance of chiplet design tools, particularly automatic chiplet placement, is increasingly apparent. Yet, previous studies did not sufficiently consider the distinctive features of chiplets, encountering challenges related to low quality of wirelength and poor scalability. Moreover, the pronounced high temperatures in 2.5D-ICs have not been thoroughly addressed, indicating a lack of thermal-aware design exploration. In response, this paper presents ATPlace2.5D, an analytical thermal-aware chiplet placement framework for large-scale 2.5D-ICs. It can deliver solutions that balance wirelength and temperature, residing on the optimal Pareto frontier, in collaboration with an innovative, physics-based compact thermal model. Experimental results show that ATPlace2.5D can handle more than 60 chiplets in minutes, and outperforms TAP-2.5D in both maximum temperature and total wirelength by 5% and 42% in thermal-aware placement, with a 23× acceleration. This advancement holds promise for promoting the maturity and widespread application of 2.5D-ICs.

1 INTRODUCTION

Recent years have witnessed an increasing demand for cost-effective and scalable chips in various markets, such as processors, automotive electronics, and AI [1, 2]. Yet, as the development pace of advanced technology nodes slows down, the cost of designing Systems on Chip (SoC) has been on an upward trend. Against this backdrop, 2.5D integration is being increasingly recognized and explored as a means to develop cost-efficient and large-scale chip systems. 2.5D integration, as shown in Fig. 1, involves the assembly of multiple integrated circuits (ICs) that contain a well-defined subset of functionality (*a.k.a. chiplets*) on a single interposer, which serves as a bridge facilitating high-speed and high-bandwidth communications. It offers several advantages compared to traditional SoC [3]. Firstly, it facilitates reduced costs across the design and manufacturing stages and higher yield. Secondly, 2.5D-IC enables the seamless integration of heterogeneous technologies and nodes within a single package, *i.e.*, System in Package (SiP). Lastly, it supports the re-using of pre-manufactured chiplets, paving the way for more sustainable and complex systems.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICCAD '24, October 27–31, 2024, New York, NY, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1077-3/24/10...$15.00
https://doi.org/10.1145/3676536.3676648

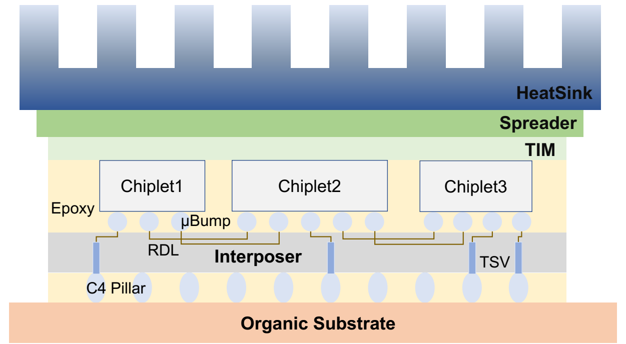


Figure 1: Illustration of the 2.5D-IC structure, fabricated with a passive interposer on an organic substrate.

Table 1: Comparison of different chiplet placement research.

Algorithm	Work	Wirelength	Scalability	Thermal-Aware
SA	Ho and Chang [4]	Medium	Low	×
	Seemuth et al. [5]		Medium	×
	Ma et al. [6]		Medium	✓
Enumeration	Osmolovskyi et al. [7]	**Low**	Medium	×
	Chiou et al. [8]			✓
RL	Duan et al. [9]	Medium	Medium	✓
	Deng et al. [10]			✓
Analytical	ATPlace2.5D (this)	**Low**	**High**	✓

To develop versatile and large-scale 2.5D-IC systems efficiently, leveraging composable chiplets (e.g., XPUs, memory, and analog modules), dedicated design automation tools are indispensable [11, 12]. Among them, we focus on the critical issue that *how to arrange the placement of the chiplets to achieve optimal performance* [13, 14]. Previous researches can be categorized into three types: simulated annealing (SA)-based [4–6, 15, 16], enumeration-based [7, 8, 17], and reinforcement-learning (RL)-based [9, 10]. The first type represents the layout in various manners, including vanilla layout [5], occupation chiplet matrix [6], and hierarchical B*-tree [4]. It can tackle multiple performance metrics beyond the wirelength, but often consumes significant runtime and comes to solutions of low quality. In contrast, the enumeration-based method can obtain better solutions for the placement of a few dies (typically less than a dozen), utilizing certain pruning (branch-and-bound (B&B)) and parallelization techniques [7, 17, 18]. For the last type, RL agents place chiplets one by one according to the reward function.

However, as summarized in Table 1, we recognize that these methods demonstrate poor scalability when dealing with large-scale 2.5D-ICs. They usually account for fewer than a dozen chiplets, whereas the scale of future 2.5D-IC systems is growing quickly beyond dozens of chiplets [19–21]. Additionally, prior efforts also suffer from low efficiency. Our study reveals that methods based on SA, enumeration, and RL all necessitate several hours to process systems comprising ten or more chiplets. This is highly time-consuming, particularly for tasks requiring iterative optimization of placement such as early-stage chiplet design space exploration [22, 23].

What's worse, prior research often focuses on just reducing the area and total wirelength, which will result in compact placement results and bring about high power density, making the large-scale

systems prone to thermal failure [24]. To this end, several recent research studies about thermal-aware placement for 2.5-D ICs. Coskun et al. [15] incorporate temperature constraints during SA, while TAP-2.5D [6] proposes to add a term related to the worst-case temperatures in the SA cost function. SP-CP [8] introduces a post-placement procedure after the B&B search to reduce the operating temperatures by refinement. However, they often conduct thermal simulations based on numerical methods [25, 26] during the iterative optimization, which significantly augments the overall runtime. Furthermore, they typically treat temperature as a constraint, handled by controlling either the maximum temperature or minimum distance between chiplets [6, 8], lacking exploration in thermal design space, which is significant for future large-scale 2.5D-IC systems.

In this work, we aim to address these deficiencies and set up an analytical thermal-aware placement framework for 2.5D-IC. To polish up the long runtime and poor scalability, we adopt an orientation-aware analytical placement algorithm [27]. In pursuit of accurate and efficient thermal evaluation, we develop a physics-based analytical compact thermal model and integrate it to optimize the overall temperature profile. Key contributions are summarized as follows:

- We propose an analytical chiplet placement framework named ATPlace2.5D, able to optimize the total wirelength and temperature simultaneously.
- We propose a new physics-based analytical compact thermal model for fast thermal simulation and optimization. It achieves a mean absolute error of $\sim 1.2\,°\mathrm{C}$ and a speedup of $2575\times$ during thermal evaluation compared to *HotSpot*.
- We propose a non-linear formulation that simultaneously optimizes wirelength and temperature as an objective for both positions and orientations of chiplets.
- In a benchmark suite with a maximum case of more than 60 chiplets, ATPlace2.5D yields solutions that surpass TAP-2.5D in both maximum temperature and total wirelength by 5% and 42% in thermal-aware placement, with a 23× acceleration.

The rest of the paper is organized as follows. Section 2 describes the background and preliminary; Section 3 explains the framework; Section 4 demonstrates the results; Section 5 concludes the paper.

2 PRELIMINARIES

In this section, we first introduce the considered 2.5D-IC configuration in Section 2.1. Then we summarize the chiplet interface and thermal models in Section 2.2 and Section 4.2. Finally, we formulate the problem in Section 2.4.

2.1 2.5D-IC Configuration

Fig. 1 shows the simplified 2.5D-IC structure discussed in this work. The bottom layer is the organic substrate, connected to the interposer through the C4 pillar (bump) layer. There exist multiple interconnect options for 2.5D-ICs, including active and passive interposer [28]. Here we choose the transistor-free passive interposer for its lower cost and flexibility for placement, while the framework can be generalized. TSVs penetrate through the interposer to connect the chiplets to the exterior. Chiplets are located above, and treated as a silicon block with certain powers. They are connected to the redistribution layers (RDL) in the interposer by arrays of microbumps, which are modeled to be homogeneous with an effective thermal conductivity considering their small size. The microbump layer and chiplet layer feature heterogeneous materials by filling the space between microbump arrays and chiplets with epoxy.

We assume the availability of various pre-manufactured and tested 'on-the-shelf' chiplets. Each chiplet is equipped with one or more interface modules for data communication. These interfaces are connected to the interposer via microbumps, which then link the chiplets needed together, forming systems of various applications.

2.2 Chiplet Interface

To enable seamless integration of diverse chiplets, interface standards are needed to foster an ecosystem [29]. Serial [30, 31] and parallel [32–34] interface are two common types of inter chiplet communication interfaces. The former only requires a few pairs of differential connections to transmit data in the physical layer, while the latter uses multiple (tens to hundreds) connections. However, performance metrics vary between these interfaces, and they have different demands on the package technologies. Therefore, there is still a long way to go to accommodate multiple chiplets by heterogeneous integration for various applications, which is out of the scope of this work [35]. For demonstration purposes, we utilize the Universal Chiplet Interconnect Express (UCIe) standard for die-to-die (D2D) interconnect, paving the way for subsequent studies, while our placement framework can be generalized to other interface standards easily.

The D2D links between chiplets are shown in the left of Fig. 2, while the right part gives the bump map inside a D2D module, which is a small portion of reference bump matrices in UCIe. A pair of signals flow out from the blue bumps, representing the transmitter (Tx) and receiver (Rx), and connect to the corresponding Rx and Tx on the other side, forming a single lane. To address diverse communication requirements, additional costs and communication overhead may sometimes be incurred, facilitated by a hub chiplet [35] interconnecting two or more chiplets.

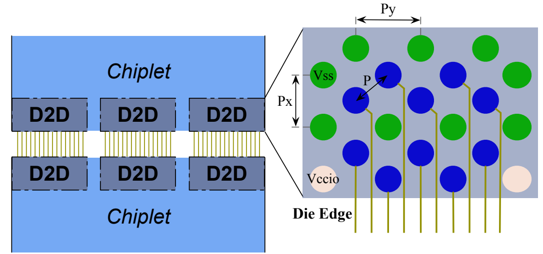


Figure 2: Illustration of the D2D link between chiplets (left), and a portion of the detailed package bump map of a ×16 module (right). The seashell and green circles are the I/O supply (Vccio) and ground reference (Vss) bump, while the blue circles represent the Tx/Rx bumps, respectively.

2.3 Thermal Models

In this work, we focus on the steady-state temperature profiles under worst-case power distribution. The governing equation reads: $\nabla \cdot (\kappa(\mathbf{r},T)\nabla T(\mathbf{r})) = -\mathbf{P}(\mathbf{r},T)$, subject to certain boundary conditions. $\kappa(\mathbf{r},T)$ is the heterogeneous conductivity, $\mathbf{P}(\mathbf{r},T)$ is the power density of each chiplet. In this work, we ignore the temperature dependence of all parameters and the above equation is reduced to be:

$$\nabla \cdot (\kappa(\mathbf{r})\nabla T(\mathbf{r})) = -\mathbf{P}(\mathbf{r}). \tag{1}$$

As a common practice [6, 8], we assume that heat is dissipated into the ambient by convection through the primary path composed of thermal interface material (TIM), heat spreader, and heatsink. Namely, adiabatic boundary condition is imposed on each lateral surface, ignoring intricate mechanisms and the secondary path of substrate and PCB. We use an air-forced heatsink as the cooling technique, and the ambient temperature is $45°C$. The edge size of the heat spreader and heatsink are both 2× that of the interposer and spreader. We inherit the properties (such as layer thickness, materials, dimensions

of bumps, and TSVs) from [6, 36]. The convective resistance of the heatsink is set to be 0.1K/W for all cases [26].

Various thermal simulators have been proposed, and mainstream methods encompass numerical and analytical approaches [37]. Numerical methods that mesh the system and solve linear equations form the foundation of a majority of simulators, whether commercial (*COMSOL* [38], *Celcius* [39]) or academic (*HotSpot* [25], *ATSim3D* [36], *etc.*). Among them, we choose the *HotSpot* 6.0 [40] as the golden simulator, and the resolution of thermal grids g is 64×64.

However, while numerical methods can offer high precision, they also demand large computation time. Hence, analytical methods are favored during the design stage, where iterative simulations are required. Such methods can provide rapid solutions through closed-form approximate expressions, either obtaining exact solutions based on simplified models [41] or constructing approximate expressions based on accurate simulation results [42–44]. The latter approach often relies on Green's function, system's impulse response concerning input power given specified initial or boundary conditions.

Table 2: Some notations used in this paper.

Notation	Meaning
$\mathbb{C}, \mathbb{A}, \mathbb{E}$	Sets of chiplets, connected chiplet pairs, and nets.
$i, j; C_i, C_j$	Index and symbol of a chiplet $\in \mathbb{C}$.
x_i, y_i, θ_i	X and Y coordinates of the center, and rotation angle of the chiplet C_i.
w_i, h_i, t_i	Width, height, and thickness of chiplet C_i.
w_i', h_i'	Width and height of chiplet C_i after rotation.
$\mathbb{B}_i, \mathbb{B}_e$	Set of microbumps belongs to the chiplet C_i or net e.
x_{p_i}, y_{p_i}	X- and Y-offsets from the center of the chiplet C_i for a microbump $p_i \in \mathbb{B}_i$.
w_{gap}	Minimum spacing between two chiplets, set to be $100\mu m$ as in [6].
W, H	Width and height of the placement region (interposer).

2.4 Problem Formulation

We summarize the notations involved in this paper in Table 2. The goal of this work is to determine the position and orientation (counterclockwise direction) of the chiplets, minimizing the total wirelength and worst-case temperature meanwhile. Note that the orientation can not be arbitrary according to design rules, but must adhere to certain legal values, denoted by $\Theta = \{\theta^0, \theta^1, \theta^2, \theta^3\}$, where $\theta^{0\sim3}$ are $0°, 90°, 180°, 270°$ respectively.

In this work, inter-chiplet communications, a pivotal factor affecting system performance, is measured by the total wirelength, following the convention of previous work [6, 17]. Here we consider only the signal nets (between Tx and Rx in the D2D links), under the premise that power and ground connections and external I/O connectivity can be addressed in subsequent design stages. Considering that signal nets in chiplets are almost two-pin nets (transmission lines in the interposer), we do not refer to the common half-perimeter wirelength and smooth approximation functions as in previous chip-level placement works [45]. Instead, we define the total wirelength $WL(\mathrm{x}, \mathrm{y}, \theta)$ following its exact definition:

$$WL = \Sigma_{e\in\mathbb{E}}\Sigma_{p_i,p_j\in\mathbb{B}_e}\left(\|X_{p_i} - X_{p_j}\| + \|Y_{p_i} - Y_{p_j}\|\right), \tag{2}$$

where p_i and p_j are the two microbumps belonging to net e and

$$X_{p_{i(j)}} = x_{i(j)} + x_{p_{i(j)}} \cdot \cos\theta_{i(j)} - y_{p_{i(j)}} \cdot \sin\theta_{i(j)},$$
$$Y_{p_{i(j)}} = y_{i(j)} + x_{p_{i(j)}} \cdot \sin\theta_{i(j)} - y_{p_{i(j)}} \cdot \cos\theta_{i(j)},$$

for $i(j)$ representing the corresponding chiplet of microbump $p_i(p_j)$.

Different from the prior practice of treating the worst-case temperature by constraints, we propose a formulation that seeks to optimize the overall thermal distribution $T_g(\mathrm{x}, \mathrm{y}, \theta)$. Here the temperature map is calculated according to Eq. (1), with the power sources to be chiplets under certain worst-case workloads. To this end, we define a new temperature penalty term $\mathcal{R}\left[T\right](\mathrm{x}, \mathrm{y}, \theta)$:

$$\mathcal{R}[T] = \Sigma_g\left[T_g(\mathrm{x}, \mathrm{y}, \theta) - T_{th}\right]^\gamma \tag{3}$$

where threshold temperature T_{th} and positive integer γ controls the strength of penalty. A γ greater than 1 makes the penalty grow at an increasing rate as the temperature rises, which can be used to strongly discourage solutions with high temperatures.

To avoid overlapping between chiplets, we employ the bell-shaped density function as in [27, 45]. After the placement region (interposer) has been divided into uniform bin grids, the density function $D_b(\mathrm{x}, \mathrm{y}, \theta)$ in bin b is defined as:

$$D_b(\mathrm{x}, \mathrm{y}, \theta) = \sum_{C_i\in\mathbb{C}}\left(D_i \times P_x(b, C_i)P_y(b, C_i)\right), \tag{4}$$

where D_i is the normalization factor, $P_x(b, C_i)$, $P_y(b, C_i)$ are the overlap functions of bin b and chiplet C_i along the x and y directions. Since their plain expressions are neither smooth nor differentiable, the bell-shaped potential function is used to smooth P_x and P_y.

With the preparations above, we can model the chiplet placement as a constrained optimization problem:

$$\begin{aligned} \min \quad & WL(\mathrm{x}, \mathrm{y}, \theta) + \lambda_{therm}\mathcal{R}\left[T_g\right](\mathrm{x}, \mathrm{y}, \theta) \\ \text{s.t.} \quad & D_b(\mathrm{x}, \mathrm{y}, \theta) \le M_b, \quad \text{for each bin } b, \end{aligned}$$

where M_b is the maximum allowable area inside b, defined as: $M_b = t_{max}(w_b h_b)$, where t_{max} is a parameter representing target density value for each bin, and w_b (h_b) is the width (height) of bin b. Utilizing the quadratic penalty method, it is further translated into an unconstrained minimization problem:

$$\min_{\mathrm{x},\mathrm{y},\theta}\{WL + \lambda_{therm}\mathcal{R}\left[T\right] + \lambda_{dens}\Sigma_b\left(D_b - M_b\right)^2\}, \tag{5}$$

where $\lambda_{therm}, \lambda_{dens}$ are balancing factors for temperature and density.

Finally, we formulate the problems in this work as follows.

Problem I (Compact Thermal Model): *Given the thermal configuration, including system geometry, material parameters, and power, the goal is to generate a compact thermal model that can predict the temperature map as accurately as possible compared to* HotSpot.

Problem II (Thermal-Aware Optimization): *Given the design specifications and chiplet information, including geometrical sizes, location of microbumps, and connection relations, the goal is to derive the chiplet placement with both the total wirelength and maximum temperature as small as possible.*

3 ATPLACE2.5D ALGORITHM

In this section, we illustrate the framework of ATPlace2.5D. We first introduce the whole framework in Section 3.1, and then the proposed compact thermal model in Section 3.2. The thermal-aware chiplet placement algorithm is detailed in Section 3.3 (initialization), Section 3.4 (optimization), and Section 3.5 (legalization).

3.1 Whole Framework

The flowchart of our framework is shown in Fig. 3. We first train a compact thermal model (Section 3.2) with the thermal simulator HotSpot based on the thermal configuration, including system geometry, material parameters, and power of chiplets. Then comes the placement process. Given the placement input, including the size of chiplets, the location and connection map of microbumps, and information of the interposer, we establish a mixed integer linear

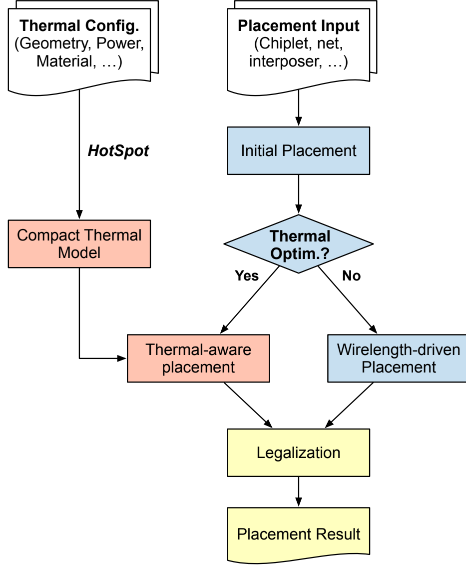


Figure 3: Illustration of the ATPlace2.5D framework.

programming (MILP) formulation to generate initial placement (Section 3.3). Then the physical location and orientation of chiplets are optimized by an analytical optimization algorithm (Section 3.4), employing the compact thermal model constructed before. Eventually, the legalization step provides legalized placement solutions (Section 3.5) that adhere to constraints on chiplet positions and fixed-outline specifications.

3.2 Compact Thermal Model

For thermal-aware optimization, the accurate gradient of the temperature w.r.t. the locations and orientations of chiplets (power sources) is necessary but hard to calculate in most cases [46]. To accelerate the acquisition of temperature and gradient information, mainstream methods [42, 43] often calculate the approximate Green's function for each grid inside the discrete system. For a system divided into $M \times M$ grids in the active layer with N chiplets, the computational complexity is generally $\mathcal{O}(M^2 \cdot M^2) = \mathcal{O}(M^4)$. These methods treat the power sources as small units and ignore their sizes, which cannot adapt well to large-size sources such as chiplets. In addition, they often omit the heterogeneous nature of the system components, which include not only silicon but also underfill (epoxy) and others.

In this work, we propose a novel compact thermal model, calculating an approximate Green's function for each chiplet, taking the chiplet size and piece-wise homogeneous materials into consideration. It features a computational complexity of $\mathcal{O}(N \cdot M^2)$, thus tends to be more efficient than previous models since the number of chiplets is much less than the number of grids.

We start from the quasi-homogeneous approximation of Eq. (1):

$$\nabla^2 T(\mathbf{r}) = -\mathbf{P}(\mathbf{r})/\kappa(\mathbf{r}), \tag{6}$$

and introduce the 3D Green's function for solving this Poisson equation: $\nabla^2 G(\mathbf{r}, \mathbf{r}_0) = \delta(\mathbf{r} - \mathbf{r}')$, where δ is the Dirac delta function, and $\mathbf{r}'$ is the source point in $\mathbb{R}^3$. It is well known that in a infinite space, $G(\mathbf{r}, \mathbf{r}_0) = \frac{1}{4\pi\|\mathbf{r}-\mathbf{r}'\|}$. Then according to Green's theorem, we can derive the general solution for Eq. (6) in free space:

$$T(\mathbf{r}) = -\iiint_\Omega G(\mathbf{r}, \mathbf{r}')\frac{\mathbf{P}(\mathbf{r}')}{\kappa(\mathbf{r}')}d\mathbf{r}', = -\sum_{C_i}\frac{\mathbf{P}_i}{4\pi\kappa_i}\iiint_{\Omega_i}\frac{1}{\|\mathbf{r}-\mathbf{r}'\|}d\mathbf{r}'. \tag{7}$$

This expression considers the impact of chiplet size by a volume integral, which is calculated over the whole source domain, that is, all the chiplets. The infinite space assumption is valid since the size of the interposer is much larger than that of chiplets. We introduce an auxiliary variable u by: $\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \frac{2}{\sqrt{\pi}}\int_0^{+\infty} du\, e^{-u^2(\mathbf{r}-\mathbf{r}')^2}$ following [47] to solve this integral, then each term in the summation reads:

$$\frac{2}{\sqrt{\pi}}\int_0^{+\infty} du \iiint_{\Omega_i} e^{-u^2((x-x')^2+(y-y')^2+(z-z')^2)}dx'dy'dz', \tag{8}$$
$$(x', y', z') \in [x_i - w_i'/2, x_i + w_i'/2] \times [y_i - h_i'/2, y_i + h_i'/2] \times \in [0, t_i].$$

Here $w_i' = \|w_i * \cos(\theta_i) + h_i * \sin(\theta_i)\|$ and h_i' are the width and height of chiplet C_i after rotation. Utilizing the error function $\mathbf{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$, the integral with respect to x' can be transformed as:

$$\frac{\sqrt{\pi}}{2u}\left[\mathbf{erf}\left(u(\frac{w_i'}{2} - (x - x_i))\right) + \mathbf{erf}\left(u(\frac{w_i'}{2} + (x - x_i))\right)\right]. \tag{9}$$

A similar expression can be derived for y' also. For the integral concerning z': $\int_0^{t_i} dz' e^{-u^2(z-z')^2}$, considering that the thickness of chiplets are much smaller than the scale in the x and y directions, we approximate it to be $t_i e^{-u^2a^2}$, where a is a fitting parameter ranging between 0 and t_i. Now Eq. (8) is transformed to be:

$$\frac{\sqrt{\pi}t_i}{2}\int_0^{+\infty}\frac{e^{-u^2a^2}du}{u^2}\cdot\left[\mathbf{erf}\left(u(\frac{w_i'}{2} - (x - x_i))\right) + \mathbf{erf}\left(u(\frac{w_i'}{2} + (x - x_i))\right)\right]$$
$$\cdot\left[\mathbf{erf}\left(u(\frac{h_i'}{2} - (y - y_i))\right) + \mathbf{erf}\left(u(\frac{h_i'}{2} + (y - y_i))\right)\right].$$

Although seems daunting at first glance, this integral can be solved analytically. It involves the calculation of four integrals, all of which can be calculated by the integral:

$$F(a, b, c) = \int_0^{+\infty} dx\, \mathrm{e}^{-a^2x^2}\frac{\mathbf{erf}(bx)\cdot\mathbf{erf}(cx)}{x^2}, \tag{10}$$

and it has a compact expression [47]:

$$F = \frac{2}{\sqrt{\pi}}\left[b\ln\left(\frac{c+\Delta}{\sqrt{a^2+b^2}}\right) + c\ln\left(\frac{b+\Delta}{\sqrt{a^2+c^2}}\right) - a\tan^{-1}\left(\frac{bc}{a\Delta}\right)\right], \tag{11}$$

here $\Delta = \sqrt{a^2+b^2+c^2}$ and $\tan^{-1}(x)$ is the inverse tangent function.

However, the above formulation is derived based on the assumption of the quasi-homogeneous condition (Eq. (6)), ignoring the heterogeneous material interface. In view of this drawback, our compact thermal model introduces two length normalization factors $l_{x,i}, l_{y,i}$ for each chiplet to capture the effects of heterogeneous conductivities. To conclude, in our model, the overall temperature is given by:

$$T_c(x, y, \theta) = \sum_i \mathrm{A}\mathbf{P}_i\Bigg[F\left(a, \frac{\frac{w_i'}{2} - (x - x_i)}{l_{x,i}}, \frac{\frac{h_i'}{2} - (y - y_i)}{l_{y,i}}\right)$$
$$+F\left(a, \frac{\frac{w_i'}{2} - (x - x_i)}{l_{x,i}}, \frac{\frac{h_i'}{2} + (y - y_i)}{l_{y,i}}\right)$$
$$+F\left(a, \frac{\frac{w_i'}{2} + (x - x_i)}{l_{x,i}}, \frac{\frac{h_i'}{2} - (y - y_i)}{l_{y,i}}\right)$$
$$+F\left(a, \frac{\frac{w_i'}{2} + (x - x_i)}{l_{x,i}}, \frac{\frac{h_i'}{2} + (y - y_i)}{l_{y,i}}\right) + \mathrm{B}\Bigg], \tag{12}$$

here A, a are the general factors of the amplitude and effective thickness for all chiplets, and B is a bias term. To sum up, we use $2N + 3$ parameters here.

This compact model is easy to construct. As shown in Fig. 3, given the thermal configuration, we first generate several legalized placement results (5-10 layouts are enough according to experiments) and

simulate the ground truth temperature map T_{label} with the thermal simulator HotSpot. We implement the model and fit the parameters (denoted by $\boldsymbol{\beta}$) by Pytorch with a mean squared loss function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|T_c(\boldsymbol{\beta}) - T_{\text{label}}\|_2^2. \tag{13}$$

Thanks to the simple closed-form formulation of Eq. (12), the fitting process is easy to converge and exhibits little dependence on the choice of initial parameters.

3.3 Initialization

The initial placement has a great impact on the quality of final results, according to previous works in chip placement [48]. However, the orientation of macros and the location of pins are often omitted, with the quadratic wirelength model considered only. In this work, we propose a MILP-based initialization algorithm, tackling the orientation of chiplets and the location of microbumps.

We begin by aggregating all nets belonging to pairs of connected chiplets, denoted as $A_{i,j}$ for $C_i, C_j \in \mathbb{A}$. Then we collect the microbumps connected to $A_{i,j}$ belonging to C_i and C_j respectively, forming two clumps on each chiplet. The center positions of the clumps are calculated as the connecting positions on C_i and C_j for nets $A_{i,j}$. The offset values of the clump in chiplet C_i connected to chiplet C_j is denoted by (O_{ij}^x, O_{ij}^y). This is reasonable since the microbumps of connections tend to be localized spatially rather than distributed, although there may be numerous connections, ranging from dozens to hundreds, between chiplets. Then we introduce two binary variables $u_i,\ v_i$ for each chiplet to describe the clump position (X_{ij}, Y_{ij}) after rotation:

$$\begin{aligned} X_{ij} &= x_i + O_{ij}^x * (1 - u_i - v_i) - O_{ij}^y * (v_i - u_i) \\ Y_{ij} &= y_i + O_{ij}^x * (v_i - u_i) + O_{ij}^y * (1 - u_i - v_i). \end{aligned}$$

Here the variables $u_i,\ v_i$ with values $(0,0), (0,1), (1,1), (1,0)$ indicate rotation angles of $0°, 90°, 180°, 270°$. After the aforementioned process, the total wirelength between the clumps of microbumps reads:

$$WL^{(init)}(\mathrm{x}, \mathrm{y}, u, v) = \Sigma_{i,j\in\mathbb{A}} A_{ij} \left(\|X_{ij} - X_{ji}\| + \|Y_{ij} - Y_{ji}\|\right). \tag{14}$$

To ensure that all chiplets are inside the placement region, we impose the following constraint for each chiplet C_i:

$$w_i'/2 \le x_i \le W - w_i'/2, \quad h_i'/2 \le y_i \le H - h_i'/2, \tag{15}$$

here $w_i' = w_i * \|1-u_i-v_i\| + h_i * \|v_i - u_i\|$ and $h_i' = w_i * \|v_i - u_i\| + h_i * \|1 - u_i - v_i\|$ are the width and height of chiplet C_i after rotation. To minimize overlapping between chiplets, we further introduce four binary variables $\delta^{(1\sim4)}$ for any pair of C_i and C_j:

$$\begin{aligned} x_i + (w_i' + w_j') * \varepsilon \le x_j + W\delta_{ij}^{(1)}, \quad x_j + (w_i' + w_j') * \varepsilon \le x_i + W\delta_{ij}^{(2)} \\ y_i + (h_i' + h_j') * \varepsilon \le y_j + H\delta_{ij}^{(3)}, \quad y_j + (h_i' + h_j') * \varepsilon \le y_i + H\delta_{ij}^{(4)} \end{aligned} \tag{16}$$

here ε is a parameter ranging between 0 and 0.5 that controls the extent of non-overlapping, and

$$\Sigma_{k=0}^{3} \delta^{(k)} <= 3 \tag{17}$$

to ensure that at least one inequality holds.

To conclude, the initial placement is obtained by solving the MILP:

$$\min WL^{(init)}(\mathrm{x}, \mathrm{y}, u, v)\ (Eq.(14)), s.t.\ \text{Eq. (15-17) holds.} \tag{18}$$

Algorithm 1 Optimization Flow based on CGD

Input: $F(\mathbf{X})$: objective function; $\mathbf{X} = (x_i, y_i, \theta_i)$: placement variables; *Max_iter*: number of iterations; $\mathbf{lr} = (lr_{pos}, lr_{ang})$: learning rates
Output: optimal $\mathbf{X}^*$
1: initialize $\mathbf{X}_0$ by solving the MILP (Eq. (18));
2: initialize λ_{dens}, $\mathbf{g}_0 = \mathbf{0}$, and $\mathbf{d}_0 = \mathbf{0}$;
3: **for** $k = 1$ to *Max_iter* **do**
4: compute function value $F(\mathbf{X}_k)$;
5: compute gradient $\mathbf{g}_k = \nabla F(\mathbf{X}_k) = (\mathbf{g}_{k,\ pos}, \mathbf{g}_{k,\ ang})$;
6: compute Polak-Ribiere parameter $\beta_k = \frac{\mathbf{g}_k^T(\mathbf{g}_k - \mathbf{g}_{k-1})}{\|\mathbf{g}_{k-1}\|_2}$ ($\beta_1 = 1$);
7: compute the conjugate direction $\mathbf{d}_k = -\mathbf{g}_k + \beta_k \mathbf{d}_{k-1}$;
8: compute the step size $\alpha_k = \frac{\mathbf{lr}}{\|\mathbf{d}_k\|_2} = (\alpha_{k,\ pos}, \alpha_{k,\ ang})$;
9: update the solution $\mathbf{X}_k = \mathbf{X}_{k-1} + \alpha_k \mathbf{d}_k$;
10: **if** *OVFL Stuck or Converge* **then**
11: **Add noise to $\mathbf{X}_k$**
12: **end if**
13: **end for**
14: legalize (Section 3.5) and derive the final result $\mathbf{X}^*$

3.4 Thermal-Aware Placement

After the initialization, the system has reached a state of reduced overlap and optimized wirelength, then it enters the subsequent stage. Our orientation-aware optimization algorithm is based on the model proposed by Lin et al. [27]. Considering that chiplets can only be rotated to certain discrete orientations, a projection function is applied to rotate the continuous angles into legal orientations when calculating the wirelength and density. In this way, the formulation of the wirelength and density in Eq. (2) and (4) are transformed to be:

$$\begin{aligned} WL' = \sum_{e\in\mathbb{E}} \sum_{p_i,p_j\in\mathbb{B}_e} \Bigg(&\|\sum_{\theta^k\in\Theta} B_z(\theta^k,\theta_i) X_{p_i}(\theta^k) - \sum_{\theta^k\in\Theta} B_z(\theta^k,\theta_j) X_{p_j}(\theta^k)\| \\ &+ \|\sum_{\theta^k\in\Theta} B_z(\theta^k,\theta_i) Y_{p_i}(\theta^k) - \sum_{\theta^k\in\Theta} B_z(\theta^k,\theta_j) Y_{p_j}(\theta^k)\| \Bigg), \end{aligned} \tag{19}$$

$$D_b' = \sum_{C_i\in\mathbb{C}} \sum_{\theta^k\in\Theta} \left(D_i \times B_z(\theta^k,\theta_i) P_x(b, C_i, \theta^k) P_y(b, C_i, \theta^k)\right), \tag{20}$$

here $B_z(\theta^k, \theta_i) = e^{\frac{R_z(\theta^k,\theta_i)}{\eta}} / \left(\sum_{\theta^k} e^{\frac{R_z(\theta^k,\theta_i)}{\eta}}\right)$ is the projection function that calculates the probability of the chiplet C_i to be rotated to each legal orientation θ^k, where η is a controlling parameter, and

$$R_z\left(\theta^k, \theta_i\right) = \begin{cases} 1 - 2|\Theta|^2 \left|\Delta\theta_i^k\right|^2, & 0 < \Delta\theta_i^k \le \frac{1}{2|\Theta|} \\ 2|\Theta|^2 \left(\Delta\theta_i^k - \frac{1}{|\Theta|}\right)^2, & \frac{1}{2|\Theta|} < \Delta\theta_i^k \le \frac{1}{|\Theta|} \\ 0, & \text{otherwise,} \end{cases}$$
$$\Delta\theta_i^k = \left|0.5 - \left|0.5 - \|\theta_i - \theta^k\|/360\right|\right|. \tag{21}$$

Instead of the direct subtraction $(\theta_i - \theta^k)/360$ in [27, 45], which can not accurately calculate the precise angular difference (such as between $350°$ and $0°$), we calculate the angular deviation using $\Delta\theta_i^k$. With this strategy, the function B_z will project the rotation angle of chiplets to the nearest legal orientation when evaluating the wirelength and density. Moreover, after optimization the chiplets are often rotated to near legal orientations, avoiding the legalization of orientations in [45].

The temperature penalty $\mathcal{R}[T_c]$ is calculated by the trained compact thermal model $T_c(\hat{\boldsymbol{\beta}})$. This term is differentiable w.r.t. x_i, y_i and θ_i through the chain rule. Then the algorithm can spread the chiplets to optimize the wirelength and alleviate the hotspot during iteration.

In our framework, the minimization of Eq. (5) is solved by the conjugate gradient descent (CGD) method. The algorithm is summarized in Algorithm 1, where the dynamic step size α_k is calculated in a heuristic manner following Chen et al. [49]. The objective function is $F = WL' + \lambda_{dens} \sum_b (D'_b - M_b)^2 + \lambda_{therm}\mathcal{R}[T_c]$. The multipliers λ_{dens} and λ_{therm} are initialized according to the strength of the gradients of wirelength and density and thermal penalty: $\lambda_{dens,\ init} = \lambda_{dens,0} \cdot \sum |\partial WL'| / \sum \left|\partial D'_b\right|$, $\lambda_{therm,\ init} = \lambda_{therm,0} \cdot \sum |\partial WL'| / \sum |\partial \mathcal{R}|$, with hyper parameter $\lambda_{dens,0}, \lambda_{therm,0}$, while the thermal multiplier $\lambda_{therm} = 0$ for wirelength-driven optimization. In light of the significant differences between positional and angular variables, we treat them separately and use two learning rates lr_{pos} and lr_{ang} in step 8.

Finally, according to experiments, the system is prone to get stuck in local optima during the nonlinear optimization. Inspired by Xue et al. [50], we perturb the placement by adding random noise when the overflow of the system is stuck at high values or converged. The overflow metric is defined to measure the evenness of chiplet distribution during optimization:

$$OVFL = \Sigma_b \max\left(D_b - M_b, 0\right) / total_chiplet_area. \tag{22}$$

3.5 Legalization

After optimization, the chiplets are roughly uniformly distributed across the interposer. But there are still some overlapping areas at this time. Additionally, sometimes minor temperature optimization can lead to excessive increases in total wirelength. Thus, a legalization stage is indispensable to eliminate all overlaps and reduce the total wirelength while keeping the changes in chiplet positions minimal, essentially maintaining the temperature distribution. In this stage, given the optimized solution $\left(x_i^{(opt)}, y_i^{(opt)}, \theta_i^{(opt)}\right)$, we fix the orientations and optimize the positions of chiplets by MILP.

The optimization objective includes two parts. The first term seeks to minimize the total wirelength:

$$WL^{(legal)} = \Sigma_{e \in \mathbb{E}} \Sigma_{p_i, p_j \in \mathbb{B}_e} \Big(\|(x_i + x_{p_i}) - (x_j + x_{p_j})\| + \|(y_i + y_{p_i}) - (y_j + y_{p_j})\| \Big), \tag{23}$$

While the second term constraints the displacement:

$$DSP = \Sigma_{C_i} \left(\|x_i - x_i^{(opt)}\| + \|y_i - y_i^{(opt)}\| \right). \tag{24}$$

Combined with the non-overlapping conditions formulated before (Eq. (16) with $\varepsilon = 1$) with the chiplet width and height modified concerning orientations and the minimum spacing between chiplets, the final MILP reads:

$$\min(DSP + \lambda_w \cdot WL^{(legal)}), s.t.\ \text{Eq. (15-17) holds}, \tag{25}$$

where λ_w is a parameter that controls the extent to which the total wirelength is optimized in this stage. Since it may be time-consuming or sometimes even infeasible to optimize the wirelength, the algorithm will set λ_w to 0 when the time limit of 100 seconds is reached, and optimize the displacement only.

Ending on a note, experiments have shown that the quality of the placement results heavily relies on the choice of parameters. Thus, inspired by AutoDMP [48], which puts forth to optimize the hyperparameters through Bayesian optimization to improve overall performance, we also carry out hyperparameter optimization to find the most favorable parameters.

4 EXPERIMENTAL RESULTS

In this section, we first introduce the experimental setup (Section 4.1) and then demonstrate our proposed compact thermal model (Section 4.2). Details of the placement results are analyzed in Section 4.3.

4.1 Experimental Setup

Due to the fact that real-world cases of large-scale 2.5D-IC systems are not available, and the test cases from previous studies overlook the die-to-die links of chiplets, we have constructed ten test cases here. The configurations of the cases are concluded in Table 3, featuring a die count from 6 to 61, a net count reaching several thousands, and a whitespace ratio within the range of 0.35 to 0.65. The first five cases consist of high-performance computing systems composed of CPU, GPU, HBM, and DRAM, adapted from the cases in [6]. The latter five cases incorporate other modules of analog and micro-electromechanical system (MEMS) [11, 22] to demonstrate the potential for heterogeneous integration. The power densities of the chiplets range from $2 \times 10^5\ \mathrm{W\,m^{-2}}$ to $3 \times 10^6\ \mathrm{W\,m^{-2}}$.

We assign several D2D link modules between chiplets for interconnect. Both standard package modules with 16 lanes (×16 module) and advanced standard package modules with 32 lanes (×32 module) are employed for data transferring in our cases. The former can provide a long channel reach while the latter has a larger bandwidth. Some of their physical parameters are summarized in Table 4, inherited from UCIe, with the geometrical meaning of different pitches shown in Fig. 2.

The ATPlace2.5D framework was developed in Python with PyTorch [51] and Optuna [52] for optimization [53]. We solve the MILPs of initialization (18) and legalization (25) by Gurobi [54]. The experiments were conducted on a Linux server with Intel Xeon 2.10 GHz processors with a maximum of 80 cores and 128 GB RAM. For all algorithms, particularly enumeration-based ones, a maximum of 80 cores and a time limit of 12 hours are set.

Table 3: Benchmark configurations.

Case	BumpType	Dies	Nets	Interposer		
				Width (mm)	Height (mm)	Whitespace/%
1	×32	6	3168	42.0	42.0	0.4
2	×32	6	3520	55.0	52.0	0.65
3	×32	8	8448	39.0	39.0	0.6
4	×32	11	7040	57.0	59.0	0.4
5	×32	12	7392	37.0	37.0	0.35
6	×16	20	5632	49.0	53.0	0.55
7	×16	28	2816	30.0	25.0	0.55
8	×16	36	2948	26.0	23.0	0.6
9	×16	44	7656	59.0	61.0	0.45
10	×16	61	5280	47.0	47.0	0.5

Table 4: Characteristics of UCIe interfaces used.

Package	Cols	Lanes	Bump Pitch (μm)	Pitch$_x$ (μm)	Pitch$_y$ (μm)
Standard (×16)	12	16	100	180	90
Advanced (×32)	16	32	25	27	42

4.2 Compact Thermal Model

We first validate the accuracy of our compact thermal model. For each case in the benchmark, we generate 5 random floorplans for training and another 10 for test. We calculate the mean absolute error (MAE), mean absolute percentage error (MAPE), and Pearson correlation factor. The correlation factor serves a dual purpose: it can evaluate the overall match between predicted and actual temperature maps while also capturing the similarity in gradients, which is essential for our gradient-based optimization. The results are summarized in Table 5. We can see that our compact model achieves an average MAE of 1.16 °C, a 3.23% relative error, and a correlation factor close to 0.99, providing sufficiently accurate temperature information. The speedup of our model compared to HotSpot

is 2575×. We further show the temperature map in Case **1** in Fig. 4. The predicted distribution is very close to the golden truth, while the error concentrates in two areas: (1) the edge of chiplets, a natural result of our quasi-homogeneous approximation (Eq. (6)); (2) the boundary of the interposer, originating from our infinite space approximation. The latter can be resolved by the mirror method, at the cost of higher computational overhead, exemplified by Lin et al. [55]. In addition, the construction of the compact model is highly efficient, since both the dataset generation and the model training process consume just several minutes (see Section 4.3.4).

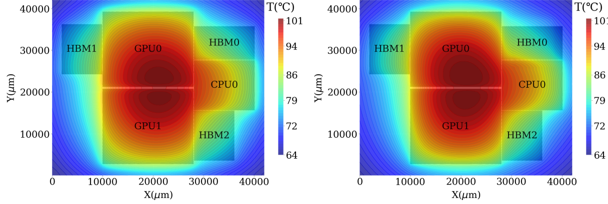


Figure 4: Thermal map of the placement by SP-CP in Case 1. (left) golden truth, (right) predicted by our compact model.

Table 5: Performance of the proposed thermal model.

Case	HotSpot Time/s	Ours			
		Time/ms	MAPE/%	MAE/°C	Correlation
1	22	5.3	2.42	0.96	0.994
2	21	5.4	5.08	0.65	0.992
3	18	6.8	3.27	1.25	0.992
4	20	6.2	3.14	1.72	0.986
5	29	7.1	2.52	1.50	0.991
6	23	9.8	3.69	1.25	0.984
7	15	9.3	3.29	0.72	0.975
8	13	11.8	3.45	0.82	0.977
9	29	14.4	3.11	1.67	0.989
10	22	17.1	2.29	1.06	0.991
Avg.	**2575×**	**1×**	**3.23**	**1.16**	**0.987**

Table 6: Comparison of different placement algorithms for thermal-aware optimization. 'RT' represents the total Runtime.

Case	TAP-2.5D [6]			ATPlace2.5D		
	RT/min	TWL/m	T_{max}/°C	RT/min	TWL/m	T_{max}/°C
1	198	56.81	**93.0**	7	52.89	94.1
2	234	145.06	75.0	8	123.52	**74.9**
3	186	185.20	94.1	6	142.42	**92.3**
4	250	202.87	128.4	11	169.80	**125.9**
5	222	167.87	114.4	10	108.31	**111.0**
6	192	169.82	93.6	7	106.97	**92.6**
7	126	46.84	71.9	6	42.49	**67.3**
8	138	44.31	67.5	8	23.55	**66.3**
9	340	295.49	130.6	16	148.95	**124.1**
10	300	141.44	103.9	23	105.23	**104.0**
Avg.	23×	1.42×	1.05×	**1×**	**1×**	**1×**

4.3 Placement Results

In this section, we evaluate various chiplet placement works, including the SA-based TAP-2.5D [6], enumeration-based SP-CP [8], and our ATPlace2.5D. RL-based methods like [9, 10] are not compared since the executable files are not yet available after contacting the authors. And according to the paper, their performance is expected to be slightly better than that of TAP-2.5D. We adapt the open-source code of TAP-2.5D to accommodate our problem setting while inheriting their choice of hyperparameters. The binary of SP-CP for wirelength optimization is compared, while the thermal optimization part is not available due to commercial reasons. The open-sourced binary of Osmolovskyi et al. [7] fails in our case since

Table 7: Comparison of different placement algorithms for wirelength-driven optimization.

Case	SP-CP [8]			TAP-2.5D [6]			ATPlace2.5D		
	RT/min	TWL/m	T_{max}/°C	RT/min	TWL/m	T_{max}/°C	RT/min	TWL/m	T_{max}/°C
1	0.02	**11.90**	101.4	3.1	49.44	100.2	0.3	12.01	102.0
2	0.01	16.11	78.7	3.5	63.16	77.5	0.3	**15.35**	78.7
3	0.5	**32.20**	117.3	7.3	110.80	114.6	0.7	34.09	116.9
4	84	**54.31**	141.9	6.3	153.32	133.4	2.6	54.47	128.3
5	>12h	52.05	N/A†	6.3	147.99	120.0	3.5	**47.89**	131.9
6	>12h	N/A	N/A	5.0	81.95	102.2	1.1	**29.27**	116.2
7	>12h	N/A	N/A	2.8	32.23	73.3	2.7	**12.29**	75.3
8	>12h	N/A	N/A	3.0	24.51	70.2	3.2	**8.41**	71.1
9	>12h	N/A	N/A	7.0	183.36	137.4	4.1	**43.21**	148.0
10	>12h	N/A	N/A	4.9	115.24	107.9	7.5	**25.74**	123.4
Avg.	8.3×	1.01×‡	1.02×	4.6×	3.43×	0.90×	**1×**	**1×**	1×

† The binary of SP-CP only reports the temporary minimal wirelength at time limit
‡ Compared only for the first five cases

it may not handle cases with too many nets. All of these works are tested on the 10 cases above, but the enumeration-based SP-CP can not derive the solutions for cases with more than 12 chiplets within the time limit.

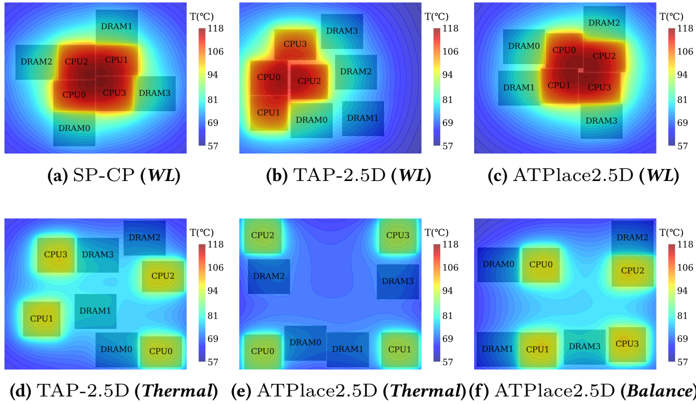
4.3.1 Thermal-Aware Optimization. To begin with, we show the results for thermal-aware optimization in Table 6. Our method can obtain the minimal total wirelength (42% improvement in average compared to that of TAP-2.5D) and lowest maximum temperature (5% improvement) for almost all cases. ATPlace2.5D provides a noticeable temperature reduction when the overall temperature is high (exceeding 100 °C at it worst). Additionally, when the number of chiplets increases, our method exhibits a more significant improvement in total wirelength compared to SA-based TAP-2.5D, possibly due to the limited exploration of high-dimensional placement spaces of SA. When it comes to the runtime, our method can achieve an efficiency gain of 23× compared to the TAP-2.5D, thanks to our efficient analytical optimization algorithm and compact thermal model.

4.3.2 Wirelength-driven Optimization. Table 7 summarizes the results of wirelength-driven optimization. The table illustrates several insights. Firstly, for scenarios with a few chiplets (<12), the enumeration-based SP-CP can achieve minimal total wirelength for most cases (Case **1,3,4**) at the cost of the highest maximum temperature. Even so, in certain scenarios, such as case **2**, our method can still achieve solutions with smaller total wirelength and similar temperature than SP-CP. Besides, our results are marginally better than SP-CP when looking at the average *TWL* (0.6% improvement) for the first five cases. Secondly, both the number of chiplets and the number of nets influence the quality of the placement results of TAP-2.5D. The larger the number of chiplets and nets, the inferior the total wirelength by TAP-2.5D compared to ATPlace2.5D. Ultimately, deducing from Table 6 and 7, there is a trade-off between temperature and wirelength, indicating further exploration of the interplay between these two metrics.

4.3.3 Multi-objective Optimization. Practical designs often need to trade off between wirelength and temperature, necessitating multi-objective optimization. Fig. 6 provides the Pareto front of Case **3** and **9**, respectively. The Pareto front delineates a collection of superior results that balance wirelength and temperature effectively. As shown in Fig. 6 (a), our work can not only deliver solutions with the smallest total wirelength and maximum temperature, but also allow for thorough exploration between the two limits. For a more intuitive understanding, in Fig. 5 we show placement layouts in the Pareto front. We can see that in the wirelength-driven optimization, our result (Fig. 5(c)) is similar to the optimal solution (Fig. 5(a)), concentrating the CPUs together, surrounded by the DRAMs around. For thermal-aware optimization, our result (Fig. 5(e)) exhibits a more dispersed distribution of chiplets compared to that of TAP-2.5D

(Fig. 5(d)), keeping the CPUs (featuring high power density) as far away from each other as possible, thereby significantly reducing the worst-case temperature. We also present results that balance between wirelength and temperature, as depicted in Fig. 5 (f), where significant thermal optimization (20 °C) is achieved with an acceptable wirelength increase. Fig. 6 (b) indicates that for cases with more chiplets, the solutions obtained by TAP-2.5D tend to be of lower quality due to the super-linear growth of the placement space w.r.t. the number of chiplets. In contrast, our method can provide a much better Pareto front. In summary, ATPlace2.5D enables designers to freely select placement solutions to meet diverse needs.

Moreover, it should be noted that the quality of the placement heavily depends on the parameters in the framework. It was found that the overlapping parameter ε in Eq. (16) has the major impact in most cases. This is reasonable since the initial solution is crucial for subsequent optimization. What's more, lr_{pos} and lr_{ang} have distinctive impacts on the final results, validating our decision to treat them separately.



(a) SP-CP (*WL*) (b) TAP-2.5D (*WL*) (c) ATPlace2.5D (*WL*)

(d) TAP-2.5D (*Thermal*) (e) ATPlace2.5D (*Thermal*) (f) ATPlace2.5D (*Balance*)

Figure 5: Chiplet placement results of different algorithms for Case **3**. The abbreviation *WL* and *Thermal* represent results for wirelength-driven and thermal-aware optimization, respectively, and *Balance* means balanced results. The dimensions of a single CPU and DRAM measure $9 \times 8mm^2$, and $9 \times 9mm^2$.

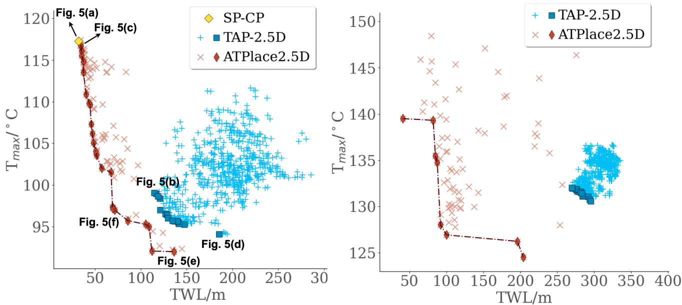


Figure 6: Pareto fronts of the multi-objective optimization in Case **3** (Up) and **9** (Down). Diamond/Square-shaped markers indicate solutions located at the Pareto frontier by each method, and the dashed line represents the complete Pareto front.

4.3.4 *Runtime Analysis.*

In this section, we break down the runtime of ATPlace2.5D for the thermal-aware optimization in Case **3** and **10** in Fig. 7. We can observe that the majority of the time, ranging from ~ 60% to ~ 80%, is spent on constructing the compact thermal model. This expenditure is acceptable in practice if multiple optimization iterations are required to find the Pareto front since the model can be reused once trained. In the remaining portion, initialization is typically rapid, while the optimization and legalization processes take up a significant amount of time, which is reasonable.

Finally, we show the progression of the wirelength and temperature during thermal-aware optimization in Case **3** in Fig. 8. In the initial stages, wirelength tends to increase while temperature decreases, as the initial solutions are clustered together. Soon, both of them will stabilize, and the perturbations (in Section 3.4) are observed, which can motivate the exploration of a larger solution space.

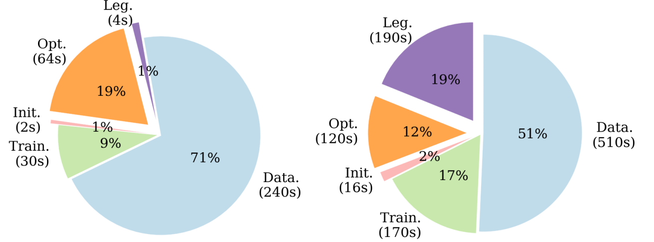


Figure 7: Runtime breakdown of ATPlace2.5D in Case 3 (left) and 9 (right). 'Data.', 'Train.', 'Init.', 'Opt.', and 'Leg.' represent the process of dataset generation, compact model training, initialization, optimization, and legalization, with the time consumed also annotated.

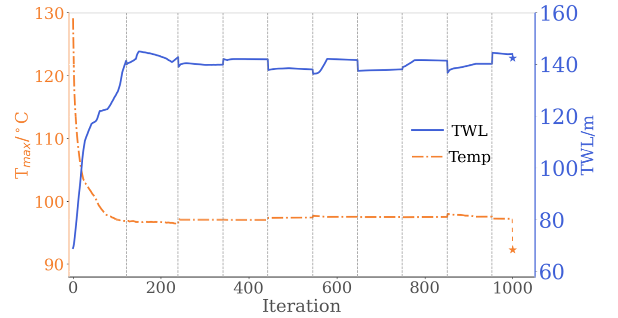


Figure 8: Curves of the total wirelength and maximum temperature during thermal-aware optimization in Case 3. The dashed line represents the moment of introducing perturbations, and the star represents the final results after legalization.

5 CONCLUSION

As both academia and industry pay increasing attention to the research and development of 2.5D-ICs, the advent of large-scale 2.5D-IC systems is inevitable. However, current 2.5D-IC placement studies are inadequate in scalability, efficiency, and thermal-aware optimization, posing a barrier to further development. Confronting the challenges, this paper presents an analytical thermal-aware 2.5D-IC placement framework ATPlace2.5D. With the analytical optimization algorithms, it achieves solutions with wirelength as small as enumeration-based methods. What's more, we propose a novel physics-based compact thermal model that provides a fast and accurate temperature map compared to *HotSpot* (1.2 °C mean error and 2575× acceleration) during thermal optimization. We construct large-scale 2.5D-IC cases conforming to UCIe standards as the benchmark for thermal-aware wirelength optimization. ATPlace2.5D demonstrates superior placement solutions in terms of both total wirelength and maximum temperature by 5% and 42% compared to TAP-2.5D in thermal-aware optimization. We envision that it will foster the growth of sustainable and versatile automatic design of large-scale 2.5D-ICs.

ACKNOWLEDGE

This work was supported in part by the National Science Foundations of China (Grant No. 62125401, 62034007), the Natural Science Foundation of Beijing, China (Grant No. Z230002) and the 111 project (B18001).

REFERENCES

[1] S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, and S. White, "Pioneering chiplet technology and design for the amd epyc™ and ryzen™ processor families: Industrial product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 57–70.

[2] G. H. Loh and R. Swaminathan, "The next era for chiplet innovation," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–6.

[3] A. Sangiovanni-Vincentelli, Z. Liang, Z. Zhou, and J. Zhang, "Automated design of chiplets," in *Proceedings of the 2023 International Symposium on Physical Design*, 2023, pp. 1–8.

[4] Y.-K. Ho and Y.-W. Chang, "Multiple chip planning for chip-interposer codesign," in *Proceedings of the 50th Annual Design Automation Conference*, 2013, pp. 1–6.

[5] D. P. Seemuth, A. Davoodi, and K. Morrow, "Automatic die placement and flexible i/o assignment in 2.5 d ic design," in *Sixteenth International Symposium on Quality Electronic Design*. IEEE, 2015, pp. 524–527.

[6] Y. Ma, L. Delshadtehrani, C. Demirkiran, J. L. Abellan, and A. Joshi, "Tap-2.5 d: A thermally-aware chiplet placement methodology for 2.5 d systems," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 1246–1251.

[7] S. Osmolovskyi, J. Knechtel, I. L. Markov, and J. Lienig, "Optimal die placement for interposer-based 3d ics," in *2018 23rd Asia and South Pacific design automation conference (ASP-DAC)*. IEEE, 2018, pp. 513–520.

[8] H.-W. Chiou, J.-H. Jiang, Y.-T. Chang, Y.-M. Lee, and C.-W. Pan, "Chiplet placement for 2.5 d ic with sequence pair based tree and thermal consideration," in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, 2023, pp. 7–12.

[9] Y. Duan, X. Liu, Z. Yu, H. Wu, L. Shao, and X. Zhu, "Rlplanner: Reinforcement learning based floorplanning for chiplets with fast thermal analysis," 2024.

[10] Z. Deng, Y. Duan, L. Shao, and X. Zhu, "Chiplet placement order exploration based on learning to rank with graph representation," *arXiv preprint arXiv:2404.04943*, 2024.

[11] J. Nasrullah, Z. Luo, and G. Taylor, "Designing software configurable chips and sips using chiplets and zglue," in *International Symposium on Microelectronics*, vol. 2019, no. 1. International Microelectronics Assembly and Packaging Society, 2019, pp. 000 027–000 032.

[12] P. Ehrett, T. Austin, and V. Bertacco, "Chopin: Composing cost-effective custom chips with algorithmic chiplets," in *2021 IEEE 39th International Conference on Computer Design (ICCD)*. IEEE, 2021, pp. 395–399.

[13] S. Osmolovskyi and J. Lienig, "Physical design challenges and solutions for interposer-based 3d systems," in *Reliability by Design; 9. ITG/GMM/GI-Symposium*. VDE, 2017, pp. 1–8.

[14] M. A. Kabir and Y. Peng, "Chiplet-package co-design for 2.5 d systems using standard asic cad tools," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2020, pp. 351–356.

[15] A. Coskun, F. Eris, A. Joshi, A. B. Kahng, Y. Ma, and V. Srinivas, "A cross-layer methodology for design and optimization of networks in 2.5 d systems," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.

[16] F. Eris, A. Joshi, A. B. Kahng, Y. Ma, S. Mojumder, and T. Zhang, "Leveraging thermally-aware chiplet organization in 2.5 d systems to reclaim dark silicon," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1441–1446.

[17] W.-H. Liu, M.-S. Chang, and T.-C. Wang, "Floorplanning and signal assignment for silicon interposer-based 3d ics," in *Proceedings of the 51st Annual Design Automation Conference*, 2014, pp. 1–6.

[18] X. Yang, Z. Liu, K. Tang, X. Yin, C. Zhuo, Q. Wei, and F. Qiao, "Breaking the energy-efficiency barriers for smart sensing applications with "sensing with computing" architectures," *Science China Information Sciences*, vol. 66, no. 10, p. 200409, 2023.

[19] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, and G. Rodgers, "Achieving exascale capabilities through heterogeneous computing," *IEEE Micro*, vol. 35, no. 4, pp. 26–36, 2015.

[20] Y. S. Shao, J. Cemons, R. Venkatesan, B. Zimmer, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina *et al.*, "Simba: scaling deep-learning inference with chiplet-based architecture," *Communications of the ACM*, vol. 64, no. 6, pp. 107–116, 2021.

[21] H. Jiang, "Intel's ponte vecchio gpu: Architecture, systems & software," in *2022 IEEE Hot Chips 34 Symposium (HCS)*. IEEE Computer Society, 2022, pp. 1–29.

[22] F. Li, Y. Wang, Y. Cheng, Y. Wang, Y. Han, H. Li, and X. Li, "Gia: A reusable general interposer architecture for agile chiplet integration," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.

[23] S. Chen, S. Li, Z. Zhuang, S. Zheng, Z. Liang, T.-Y. Ho, B. Yu, and A. L. Sangiovanni-Vincentelli, "Floorplet: Performance-aware floorplan framework for chiplet integration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[24] J.-H. Han, X. Guo, K. Skadron, and M. R. Stan, "From 2.5 d to 3d chiplet systems: Investigation of thermal implications with hotspot 7.0," in *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*. IEEE, 2022, pp. 1–6.

[25] M. R. Stan, K. Skadron, M. Barcella, W. Huang, K. Sankaranarayanan, and S. Velusamy, "Hotspot: A dynamic compact thermal model at the processor-architecture level," *Microelectronics Journal*, vol. 34, no. 12, pp. 1153–1165, 2003.

[26] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-d multicore systems with stacked dram under power and thermal constraints," in *DAC Design Automation Conference 2012*, 2012, pp. 648–655.

[27] J.-M. Lin, T.-C. Tsai, and R.-T. Shen, "Routability-driven orientation-aware analytical placement for system in package," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 2023, pp. 1–8.

[28] D. Stow, I. Akgun, and Y. Xie, "Investigation of cost-optimal network-on-chip for passive and active interposer systems," in *2019 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*. IEEE, 2019, pp. 1–8.

[29] D. Sharma *et al.*, "Universal chiplet interconnect express (ucie)," *MEPTEC: Road to Chiplets*, pp. 10–12, 2022.

[30] B. Dehlaghi, N. Wary, and T. C. Carusone, "Ultra-short-reach interconnects for die-to-die links: Global bandwidth demands in microcosm," *IEEE Solid-State Circuits Magazine*, vol. 11, no. 2, pp. 42–53, 2019.

[31] K. Ma, "Introducing acc 1.0: Advanced cost-driven chiplet interface standard," in *The 3rd HiPChips Conference at ISCA*, 2023.

[32] M.-S. Lin, C.-C. Tsai, C.-H. Hsieh, W.-H. Huang, Y.-C. Chen, S.-C. Yang, C.-M. Fu, H.-J. Zhan, J.-Y. Chien, S.-Y. Li *et al.*, "A 16nm 256-bit wide 89.6 gbyte/s total bandwidth in-package interconnect with 0.3 v swing and 0.062 pj/bit power in info package," in *2016 IEEE Hot Chips 28 Symposium (HCS)*. IEEE, 2016, pp. 1–32.

[33] S. Ardalan, R. Farjadrad, M. Kuemerle, K. Poulton, S. Subramaniam, and B. Vinnakota, "An open inter-chiplet communication link: Bunch of wires (bow)," *IEEE Micro*, vol. 41, no. 1, pp. 54–60, 2020.

[34] Intel. (2022) Advanced interface bus (aib) specification, revision 2.0.3. Update Date: 2022/06/17.

[35] Z. Yang, S. Ji, X. Chen, J. Zhuang, W. Zhang, D. Jani, and P. Zhou, "Challenges and opportunities to enable large-scale computing via heterogeneous chiplets," *arXiv preprint arXiv:2311.16417*, 2023.

[36] Q. Wang, T. Zhu, Y. Lin, R. Wang, and R. Huang, "Atsim3d: Towards accurate thermal simulator for heterogeneous 3d ic systems considering nonlinear leakage and conductivity," in *2024 International Symposium of Electronics Design Automation (ISEDA)*, 2024, pp. 1–6.

[37] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in vlsi circuits: Principles and methods," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1487–1501, 2006.

[38] "Comsol Multiphysics," http://www.comsol.com/products/multiphysics/.

[39] "Cadence Celcius Thermal Solver," https://www.cadence.com/en_US/home/tools/system-analysis/thermal-solutions/celsius-thermal-solver.html.

[40] R. Zhang, M. R. Stan, and K. Skadron, "Hotspot 6.0: Validation, acceleration and extension," *University of Virginia, Tech. Rep*, 2015.

[41] B. Wang and P. Mazumder, "Accelerated chip-level thermal analysis using multilayer green's function," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 2, pp. 325–344, 2007.

[42] C. Torregiani, H. Oprins, B. Vandevelde, E. Beyne, and I. De Wolf, "Compact thermal modeling of hot spots in advanced 3d-stacked ics," in *2009 11th Electronics Packaging Technology Conference*, 2009, pp. 131–136.

[43] J. Kung, I. Han, S. Sapatnekar, and Y. Shin, "Thermal signature: A simple yet accurate thermal index for floorplan optimization," in *Proceedings of the 48th Design Automation Conference*, 2011, pp. 108–113.

[44] 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 (VLSI) Systems*, vol. 22, no. 11, pp. 2366–2379, 2014.

[45] M.-K. Hsu and Y.-W. Chang, "Unified analytical global placement for large-scale mixed-size circuit designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 9, pp. 1366–1378, 2012.

[46] H. Park and O. Chung, "An inverse natural convection problem of estimating the strength of a heat source," *International journal of heat and mass transfer*, vol. 42, no. 23, pp. 4259–4273, 1999.

[47] O. Ciftja, "Electrostatic potential of a uniformly charged square plate at an arbitrary point in space," *Physica Scripta*, vol. 95, no. 9, p. 095802, 2020.

[48] A. Agnesina, P. Rajvanshi, T. Yang, G. Pradipta, A. Jiao, B. Keller, B. Khailany, and H. Ren, "Autodmp: Automated dreamplace-based macro placement," in *Proceedings of the 2023 International Symposium on Physical Design*, 2023, pp. 149–157.

[49] T.-C. Chen, Z.-W. Jiang, T.-C. Hsu, H.-C. Chen, and Y.-W. Chang, "Ntuplace3: An analytical placer for large-scale mixed-size designs with preplaced blocks and density constraints," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 7, pp. 1228–1240, 2008.

[50] K. Xue, X. Lin, Y. Shi, S. Kai, S. Xu, and C. Qian, "Escaping local optima in global placement," *ArXiv*, vol. abs/2402.18311, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268041751

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[52] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[53] Y. Lin, S. Dhar, W. Li, H. Ren, B. Khailany, and D. Z. Pan, "Dreamplace: Deep learning toolkit-enabled gpu acceleration for modern vlsi placement," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[54] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com

[55] J.-M. Lin, T.-T. Chen, Y.-F. Chang, W.-Y. Chang, Y.-T. Shyu, Y.-J. Chang, and J.-M. Lu, "A fast thermal-aware fixed-outline floorplanning methodology based on analytical models," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.