

FALCON-3D: Full-Chip Analytical Thermal Simulation with Lateral CONvection for 3D-Stacked ICs

Tsung-Lin Lu, Yu-Min Lee, Pei-Yu Huang, Ching-Hsiang Wang

Institute of Communications Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
johnny8dada.ee12@nycu.edu.tw, yumin@nycu.edu.tw, pyhunagmtk@gmail.com, elbarto.ee13@nycu.edu.tw

Abstract—As the power density and complexity of modern chips continue to increase, thermal analysis has become an essential step in the design process. While existing analytical approaches assume purely vertical heat flow, lateral heat transfer becomes significant when chip thickness increases, as in 3D ICs, and cooling capability is limited, as in mobile devices. Though commercial numerical tools can capture these effects, they are too computationally intensive for use in early design stages.

This work proposes FALCON-3D, a high-performance and full-chip analytical thermal solver tailored for early-stage design analysis, which explicitly models lateral surface heat transfer. Experimental results demonstrate not only the computational efficiency of FALCON-3D but also that ignoring lateral heat transfer introduces notable errors in temperature prediction, underscoring the importance of incorporating lateral effects.

I. INTRODUCTION

As integrated circuits (ICs) continue to advance in performance, three-dimensional (3D) ICs have emerged as a key technology by enabling the vertical integration of multiple functional layers. However, this stacking architecture introduces substantial thermal management challenges. The accumulated heat in stacked layers becomes increasingly difficult to dissipate, particularly in the absence of efficient cooling mechanisms. Elevated chip temperatures may increase leakage power and signal delay, ultimately degrading system performance and reliability.

To effectively evaluate the thermal impact of such designs, accurate thermal simulators are indispensable. Numerical methods, such as the finite difference method (FDM), finite element method, and finite volume method, can handle complex geometries and boundary conditions, but need heavy computational load for large-scale problems [1]–[7]. Analytical approaches, while generally limited to simpler structures, offer superior efficiency and are especially attractive in early design stages which rapid evaluations are critical [8]–[14]. Green’s function-based approaches leverage system response and convolution to compute temperature distributions [8]–[12]. Zhan and Sapatnekar combined Green’s functions with discrete cosine transform to generate two-dimensional (2D) maps [8], while Oh *et al.* [9] and Wang and Mazumder [10] extended this technique to model multilayer heat transfer. Wang and Pang [11] presented a model that is capable of non-ideal inter-layer contacts. Sultan and Sarangi [12] accounted for

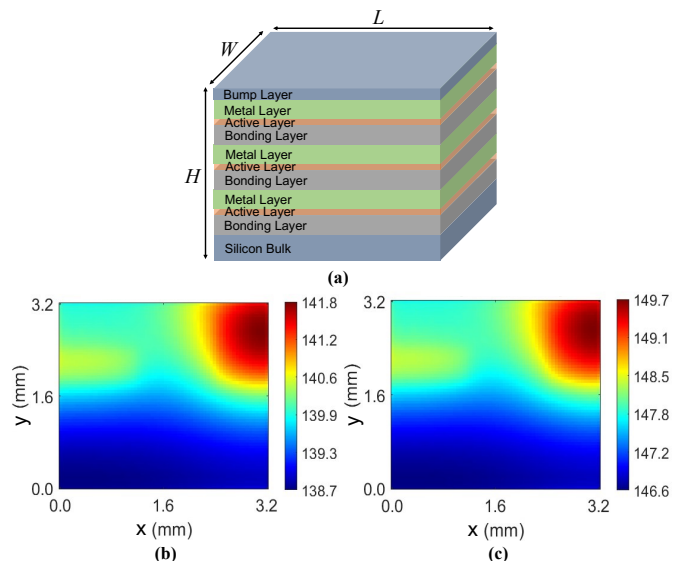


Fig. 1: (a) 3D IC structure. (b) Temperature map considering lateral convection. (c) Temperature map ignoring lateral convection.

process variation, leakage power, and temperature-dependent conductivity. The Power Blurring method [13] employs thermal masks extracted from commercial tools and applies the method of images to address edge effects. Alternatively, Huang and Lee [14] solved the Sturm–Liouville equation via basis expansion for solving the 3D steady-state temperature distribution. Their approach achieves high accuracy with fewer terms compared to traditional Green’s function methods.

Despite effectiveness, existing analytical methods assume adiabatic side boundaries, ignoring lateral heat dissipation. It is reasonable as vertical heat dissipation dominates. However, in space-constrained environments such as mobile devices [15], conventional cooling solutions like heat sinks are often impractical. In such scenarios, lateral heat dissipation through sides of the chip becomes relatively significant. Ignoring this effect may compromise the accuracy of thermal simulation.

To demonstrate the influence of lateral heat dissipation, we employ Ansys Icepak [2] to simulate a representative 3D IC structure [1]. As shown in Fig. 1(a), the chip consists

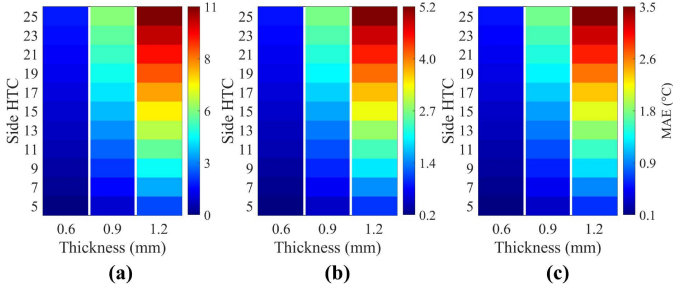


Fig. 2: MAE for varying thickness and side HTC. (a) area = 9 mm². (b) area = 49 mm². (c) area = 121 mm².

of eleven stacked layers, including three active layers, with the dimension of $3.3 \times 3.3 \times 0.78$ mm³. The top surface is attached to a printed circuit board, while the bottom is exposed to air. Boundary conditions are configured as follows [16]. The lateral surfaces are assigned a heat transfer coefficient (HTC) of 10 W/m²·°C, while the top and bottom surfaces are assigned 120 and 10 W/m²·°C, respectively. Fig. 1(b)–(c) show the temperature of the central active layer considering and ignoring lateral convection. Ignoring side cooling leads to an error up to 7.98 °C, which becomes more pronounced in thicker structures or under weaker vertical cooling.

Inaccurate estimation may result in overly conservative thermal solutions and excessively pessimistic reliability assessments, which in turn propagate unnecessary design margins throughout the design process. To address this issue, we propose FALCON-3D, an analytical framework that, to the best of our knowledge, represents the first attempt to explicitly incorporate lateral heat dissipation in both steady-state and transient simulations, while maintaining computational efficiency. FALCON-3D achieves maximum error of only 0.295 °C in steady-state simulation and 0.235 °C in transient simulation, thereby establishing it a practical solution for reliable early-stage thermal prediction.

The rest of this paper is organized as follows. Section II extends the case study analysis on lateral heat dissipation across diverse chip configuration. Section III presents the problem formulation. Section IV details the proposed analytical thermal simulator, FALCON-3D. Section V demonstrates experimental validation. Finally, Section VI concludes this paper.

II. CASE STUDY ON LATERAL HEAT DISSIPATION

To further quantify the importance of lateral heat dissipation, we conduct controlled experiments across diverse chip geometries and cooling configurations. The power map is defined over a nominal 3.3×3.3 mm² chip and geometrically rescaled for each test case while preserving its spatial distribution. The total power is adjusted to maintain consistent local heat generation.

Chip thickness is varied from 0.6 mm to 1.2 mm by stacking one to three identical die, including silicon bulk, active layer, and metal layer, while chip areas of 9, 49, and 121 mm² represent designs of increasing footprint. Convective boundary conditions are configured to reflect the equivalent environment of a mobile device, with a fixed top HTC of 120 W/m²·°C,

bottom HTC of 10 W/m²·°C, and side HTC ranging from 5 to 25 W/m²·°C.

We quantify thermal impact by computing the mean absolute error (MAE) between simulations considering and ignoring lateral convection. Fig. 2 shows the results. Across configurations, increasing side HTC consistently intensifies the error as lateral heat transfer is ignored. Thicker chips exacerbate this effect due to additional stacked power-dissipating layers, while smaller chips exhibit higher relative error since lateral surfaces form a larger fraction of thermal dissipation path.

These results confirm that neglecting lateral dissipation can severely undermine prediction accuracy, particularly in space-constrained 3D ICs, motivating the development of analytical methods that explicitly incorporate lateral convection.

III. PROBLEM FORMULATION

To analyze the temperature distribution of a 3D IC, we consider a multilayer chip structure. It contains silicon bulk, bounding layer, active layer, metal layer, and bump layer. Fig. 1(a) shows its physical model [1]. Its steady-state rising temperature $T(\mathbf{r}) = \bar{T}(\mathbf{r}) - T_a$ is governed by the heat transfer equation,

$$\nabla \cdot (\kappa(\mathbf{r})\nabla T(\mathbf{r})) = -p(\mathbf{r}); \quad \mathbf{r} \in \Omega_L, \quad (1)$$

where $T(\mathbf{r})$ is the rising temperature at location $\mathbf{r} = (x, y, z)$, $\bar{T}(\mathbf{r})$ is the temperature at $\mathbf{r} = (x, y, z)$, T_a is the ambient temperature, Ω_L is the interior domain of the chip, $p(\mathbf{r})$ is the power density in Ω_L , and $\kappa(\mathbf{r})$ is the thermal conductivity.

At the interfaces between stacked layers, the continuity conditions hold as

$$T(\mathbf{r})|_{z=\epsilon_l^+} = T(\mathbf{r})|_{z=\epsilon_l^-}, \quad (2)$$

$$\kappa_l \frac{\partial T(\mathbf{r})}{\partial z} \Big|_{z=\epsilon_l^+} = \kappa_{l+1} \frac{\partial T(\mathbf{r})}{\partial z} \Big|_{z=\epsilon_l^-}, \quad (3)$$

where ϵ_l is the interface between the l -th and $(l+1)$ -th layers, and κ_l is the thermal conductivity of the l -th stacked layer.

All external surfaces of chip are assumed to be subject to convective boundary conditions,

$$\kappa(\mathbf{r}) \frac{\partial T(\mathbf{r})}{\partial \vec{n}_b} = h_b T(\mathbf{r}); \quad \mathbf{r} \in \Omega_B, \quad (4)$$

where Ω_B is the boundary surface domain of chip that has effective HTC h_b , and \vec{n}_b is the outward surface normal vector.

IV. FALCON-3D

The flow of FALCON-3D is summarized in Fig. 3. Given the chip structure and boundary condition, the eigenfunctions $\phi_{il}(x, y)$'s with convective boundary, used as the spatial basis functions, are derived in Section IV-A. The spatial basis is then expanded by Fourier series (FS) presented in Section IV-B, which is a critical step in the overall flow for computational efficiency. The infinite FS coefficients are truncated into a finite number of modes according to an energy-based truncation scheme, and the results are stored in an offline lookup table.

Next, the given power map of each active layer is transformed into its power spectrum $\mathbf{p}_{jk,t}$'s by applying FFT. Using $\mathbf{p}_{jk,t}$'s with the pre-calculated FS coefficients, FALCON-3D

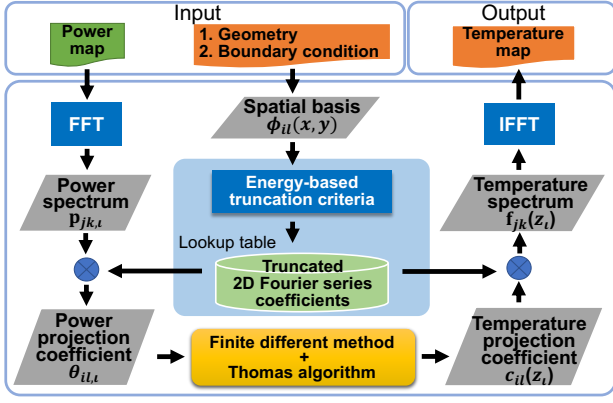


Fig. 3: Flow of FALCON-3D.

computes the power projection coefficients θ_{il} 's along the z -direction detailed in Section IV-C. Within the Galerkin framework, these power projection coefficients are then used to solve a z -directional subproblem formulated in Section IV-D via the FDM and Thomas algorithm, yielding the corresponding temperature projection coefficients $c_{il}(z_i)$'s.

Combining the temperature projection coefficients with the pre-calculated FS coefficients, we have the temperature spectrum $\mathbf{f}_{jk}(z_i)$'s. Finally, the temperature distribution $T_{mn}(z_i)$ of each grid cell is recovered by applying the inverse FFT (IFFT) to $\mathbf{f}_{jk}(z_i)$'s shown in Section IV-E. To capture the temperature waveform $T_{mn}(z_i, t)$ of dynamic workload. The capability of transient simulation is also described in Section IV-F.

A. Temperature Integral Transform

Under the Galerkin framework, $T(\mathbf{r})$ can be approximated by using a set of orthogonal spatial basis functions as

$$T(\mathbf{r}) \approx \hat{T}(\mathbf{r}) = \sum_{l=0}^{N_{By}-1} \sum_{i=0}^{N_{Bx}-1} c_{il}(z) \frac{\phi_{il}(x, y)}{\eta_{il}}, \quad (5)$$

where $c_{il}(z)$ is the coefficient for the z -direction, $\phi_{il}(x, y)$ is the lateral spatial basis function, and $\eta_{il} = \|\phi_{il}\|_{L^2(\Omega_L)}$ is the L^2 norm of the basis function over domain Ω_L . The numbers of basis functions in the x - and y -directions are N_{Bx} and N_{By} , respectively.

To construct $\phi_{il}(x, y)$, a homogeneous eigenvalue problem can be formulated to satisfy (1)–(4) as stated in [17], which can be written as the following Sturm-Liouville problem.

$$\nabla^2 \phi_{il}(x, y) + \lambda_{il}^2 \phi_{il}(x, y) = 0; \quad (x, y) \in \Omega_L, \quad (6)$$

$$\kappa \frac{\partial \phi_{il}(x, y)}{\partial \vec{n}_b} = h_b \phi_{il}(x, y); \quad (x, y) \in \Omega_B, \quad (7)$$

where λ_{il} is the eigenvalue.

To solve the 2D Sturm-Liouville problem, we employ the separation of variables by assuming a product solution,

$$\phi_{il}(x, y) = \rho_i(x) u_l(y), \quad (8)$$

where $\rho_i(x)$ and $u_l(y)$ are eigenfunctions.

Substituting (8) into (6) allows separating (6) into two independent 1D Sturm-Liouville problems in the x - and y -directions. After solving both problems independently, the

eigenfunctions, $\rho_i(x)$ and $u_l(y)$, and their corresponding eigenvalues, μ_i and ν_l , can be obtained as

$$\rho_i(x) = \kappa \cos(\mu_i x) + \bar{h}_{x_0} / \mu_i \cdot \sin(\mu_i x), \quad (9)$$

$$u_l(y) = \kappa \cos(\nu_l y) + \bar{h}_{y_0} / \nu_l \cdot \sin(\nu_l y). \quad (10)$$

Here, $\lambda_{il}^2 = \mu_i^2 + \nu_l^2$ and μ_i satisfies

$$\frac{(\kappa \mu_i)^2 - \bar{h}_{x_0} \bar{h}_{x_1}}{\kappa \mu_i (\bar{h}_{x_0} + \bar{h}_{x_1})} = \cot(\mu_i W), \quad (11)$$

where \bar{h}_{x_0} and \bar{h}_{x_1} are HTC's on the boundary surfaces x_0 ($x = 0$) and x_1 ($x = W$), respectively, and W is the length in the x -direction.

ν_l follows a similar formulation, with $(\bar{h}_{x_0}, \bar{h}_{x_1}, W)$ replaced by $(\bar{h}_{y_0}, \bar{h}_{y_1}, L)$. \bar{h}_{y_0} and \bar{h}_{y_1} are HTC's on the boundary surfaces y_0 ($y = 0$) and y_1 ($y = L$), respectively, and L is the length in the y -direction. These eigenvalues are obtained numerically via the Newton-Raphson method [18], which converges efficiently due to the monotonic structure of cotangent function. The results of μ_i and ν_l are written as

$$\mu_i = \frac{i\pi}{W} + \Delta\mu_i; \quad \nu_l = \frac{l\pi}{L} + \Delta\nu_l. \quad (12)$$

B. Periodic Continuation of Spatial Basis Functions

In [8], [14], the FFT is exploited for its high computational efficiency because the basis functions in the x - and y -directions are cosine-periodic. However, the basis functions, $\rho_i(x)$ and $u_l(y)$, in (9)–(10) are non-periodic within the analysis domain, which rules out the direct use of FFT. To overcome this limitation, $\rho_i(x)$ and $u_l(y)$ are periodically extended by means of Fourier-series expansion over the intervals $[0, W]$ and $[0, L]$, respectively. The resulting periodic surrogates $\hat{\rho}_i(x)$ and $\hat{u}_l(y)$ are expressed as

$$\rho_i(x) \approx \hat{\rho}_i(x) = \sum_{j=0}^{N_x^i-1} a_j^i \cos\left(\frac{2j\pi x}{W}\right) + b_j^i \sin\left(\frac{2j\pi x}{W}\right), \quad (13)$$

$$u_l(y) \approx \hat{u}_l(y) = \sum_{k=0}^{N_y^l-1} c_k^l \cos\left(\frac{2k\pi y}{L}\right) + d_k^l \sin\left(\frac{2k\pi y}{L}\right). \quad (14)$$

Here, the coefficients, a_j^i , b_j^i , c_k^l , and d_k^l , are obtained in closed form by projecting the basis functions onto orthogonal trigonometric functions. j and k are non-negative integer mode indices, and N_x^i and N_y^l denote the truncation numbers retained for $\rho_i(x)$ and $u_l(y)$, respectively.

The selection of these truncation numbers dictates both computational cost and approximation accuracy, and is therefore guided by an energy-based criterion. According to Parseval's theorem, the total energy of a signal equals the sum of the squared magnitudes of its Fourier coefficients. By tracking the cumulative energy captured by the leading Fourier modes, the series can be truncated once a prescribed fraction of the total energy has been retained. Specifically, when constructing the i -th basis function $\hat{\rho}_i(x)$, FALCON-3D successively evaluates a_j^i 's and b_j^i 's. The cumulative energy for the first Q terms is

$$Enrg^Q = \sum_{j=0}^Q \left(a_j^i \right)^2 + \left(b_j^i \right)^2. \quad (15)$$

The iteration terminates as the incremental contribution of the next mode satisfies

$$(a_{Q+1}^i)^2 + (b_{Q+1}^i)^2 < \gamma \cdot \text{Enrg}^{Q+1}, \quad (16)$$

where $\gamma \in (0, 1)$ is the user-defined truncation threshold.

The same criterion can be applied on c_k^l and d_k^l for $\hat{u}_l(y)$.

As shown in (12), each eigenvalue consists of a periodic component ($i\pi/W$ and $l\pi/L$) plus a deviation term ($\Delta\mu_i$ and $\Delta\nu_l$). The presence of deviation terms prevents the direct use of FFT. Due to the right-hand sides of (11) are periodic in $i\pi/W$ and $l\pi/L$, and the left-hand sides are quasi-linear with respect to i and l , these deviations decrease as the indices i and l increase. Since the deviations diminish at higher frequencies, fewer FS terms are required to accurately approximate the corresponding basis functions.

In practice, we observe that when the deviation falls below 0.001% compared to the periodic component, a single Fourier term suffices for expansion. When the deviation exceeds 0.001%, we empirically set the truncation threshold γ as $2e-6$, which provides an accurate approximation of the temperature distribution. Although $\gamma = 2e-6$ makes the truncation numbers larger than one in the low-frequency bases, very few bases (typically 2 to 4 bases) are required to apply this strict threshold.

This energy-guided stopping rule offers a balance between spectral accuracy and computational cost. After the eigenfunctions in x - and y -directions have been approximated, we substitute (13) and (14) into (8) to obtain the 2D Fourier-series expansion of $\phi_{il}(x, y)$ as

$$\phi_{il}(x, y) \approx \varphi_{il}(x, y) = \sum_{j=0}^{N_x^i-1} \sum_{k=0}^{N_y^l-1} \langle \mathbf{a}_{jk}^{il}, \boldsymbol{\alpha}_{jk} \rangle, \quad (17)$$

where $\langle \mathbf{a}_{jk}^{il}, \boldsymbol{\alpha}_{jk} \rangle$ is the inner product of \mathbf{a}_{jk}^{il} and $\boldsymbol{\alpha}_{jk}$, and

$$\boldsymbol{\alpha}_{jk} = \begin{bmatrix} \cos\left(\frac{2j\pi x}{W}\right) \\ \sin\left(\frac{2j\pi x}{W}\right) \end{bmatrix} \otimes \begin{bmatrix} \cos\left(\frac{2k\pi y}{L}\right) \\ \sin\left(\frac{2k\pi y}{L}\right) \end{bmatrix}; \quad \mathbf{a}_{jk}^{il} = \begin{bmatrix} a_j^i \\ b_j^i \end{bmatrix} \otimes \begin{bmatrix} c_k^l \\ d_k^l \end{bmatrix}.$$

Here, the operator \otimes is the Kronecker product and \mathbf{a}_{jk}^{il} denotes the coefficients of 2D FS.

Once the material properties and boundary conditions are specified, $\rho_i(x)$ and $u_i(y)$ and their associated eigenvalues μ_i and ν_l are uniquely determined. These eigenvalues in turn determine \mathbf{a}_{jk}^{il} . Because all of these quantities depend solely on material and boundary configuration, they can be computed once, stored in lookup tables, and reused throughout subsequent simulation steps, thereby to accelerate the overall analysis.

C. Power Integral Transform

For the representation of source plane in ι -th stacked layer. The power map $p_\iota(x, y)$ is represented as a piecewise-constant function [8], [14] over a uniform $M \times N$ grid.

$$p_\iota(x, y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} p_{mn,\iota} \Pi_{mn}(x, y), \quad (18)$$

where m and n are the non-negative integer indices of grid cell, $\Pi_{mn}(x, y)$ is an indicative function with nonzero value being 1

only when (x, y) is in $[m \cdot \Delta x, (m+1) \cdot \Delta x] \times [n \cdot \Delta y, (n+1) \cdot \Delta y]$, $\Delta x = W/M$ and $\Delta y = L/N$ are the sizes of grid cell in x - and y - directions, respectively, and $p_{mn,\iota}$ is the power of grid cell (m, n) in the ι -th layer.

We apply the basis expansion on the power map with basis functions obtained in (17) and each $p_\iota(x, y)$ can be projected into the spatial basis space as $\hat{p}_\iota(x, y)$,

$$\hat{p}_\iota(x, y) = \sum_{l=0}^{N_{By}-1} \sum_{i=0}^{N_{Bx}-1} \theta_{il,\iota} \varphi_{il}(x, y) / \eta_{il}. \quad (19)$$

Here, $\theta_{il,\iota}$ is the projection coefficient of $p_\iota(x, y)$ and is determined as

$$\theta_{il,\iota} = \sum_{j=0}^{N_x^i-1} \sum_{k=0}^{N_y^l-1} \langle \mathbf{a}_{jk}^{il}, \mathbf{p}_{jk,\iota} \rangle, \quad (20)$$

with

$$\mathbf{p}_{jk,\iota} = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} p_{mn,\iota} \boldsymbol{\beta}_{jk}^{mn}, \quad (21)$$

where

$$\boldsymbol{\beta}_{jk}^{mn} = S_{jk} \begin{bmatrix} \cos\left(\frac{j\pi(2m+1)}{M}\right) \\ \sin\left(\frac{j\pi(2m+1)}{M}\right) \end{bmatrix} \otimes \begin{bmatrix} \cos\left(\frac{k\pi(2n+1)}{N}\right) \\ \sin\left(\frac{k\pi(2n+1)}{N}\right) \end{bmatrix},$$

and

$$S_{jk} = \begin{cases} \Delta x \Delta y, & j = 0 \text{ and } k = 0, \\ \frac{N \Delta x}{k\pi} \sin\left(\frac{k\pi}{N}\right), & j = 0 \text{ and } k \neq 0, \\ \frac{M \Delta y}{j\pi} \sin\left(\frac{j\pi}{M}\right), & j \neq 0 \text{ and } k = 0, \\ \frac{MN}{jk\pi^2} \sin\left(\frac{j\pi}{M}\right) \sin\left(\frac{k\pi}{N}\right), & j \neq 0 \text{ and } k \neq 0. \end{cases}$$

Here, $\mathbf{p}_{jk,\iota}$ represents the power spectrum of the power map.

By expanding the non-periodic basis functions into FS shown in Section IV-B, (21) can be efficiently computed using 2D FFT.

D. 1D Sub-problem in the z -Direction

To determine $c_{il}(z)$ in (5), we substitute $\hat{T}(\mathbf{r})$ and $\hat{p}_\iota(x, y)$ into (1), and apply the Galerkin scheme using $\phi_{il}(x, y)$ together with the conditions (2)–(4). Then, the 1D sub-problem in the z -direction is

$$\kappa_\iota \left(\frac{\partial^2 c_{il}(z)}{\partial z^2} - \lambda_{il}^2 c_{il}(z) \right) = -\theta_{il,\iota}, \quad (22)$$

$$c_{il}(z) \Big|_{z=\epsilon_\iota^+} = c_{il}(z) \Big|_{z=\epsilon_\iota^-}, \quad (23)$$

$$\kappa_\iota \frac{\partial c_{il}(z)}{\partial z} \Big|_{z=\epsilon_\iota^+} = \kappa_{\iota+1} \frac{\partial c_{il}(z)}{\partial z} \Big|_{z=\epsilon_\iota^-}, \quad (24)$$

$$\kappa_0 \frac{\partial c_{il}(z)}{\partial z} \Big|_{z=0} = \bar{h}_{z_0} c_{il}(0) \quad (25)$$

$$\kappa_{N_z-1} \frac{\partial c_{il}(z)}{\partial z} \Big|_{z=H} = \bar{h}_{z_1} c_{il}(H), \quad (26)$$

where κ_ι is the thermal conductivity in the ι -th layer, λ_{il} is the eigenvalue defined in (6)–(7), N_z is the number of stacked layers, and H is the thickness of chip.

(23) and (24) indicate the continuity between layer interface ϵ_ι , and (25) and (26) are the convective boundary conditions

on the top and bottom surfaces, respectively. This 1-D partial differential equation can be calculated by discretizing the spatial domain using the FDM. The resulting linear system is

$$\mathbf{G}\mathbf{x} = \boldsymbol{\theta}, \quad (27)$$

where \mathbf{G} is a tridiagonal thermal conductance matrix, \mathbf{x} is the vector containing sampling values of $c_{il}(z_t)$'s, which z_t is the coordinate at the middle of l -th layer, and $\boldsymbol{\theta}$ is the vector of $\theta_{il,t}$'s.

Since \mathbf{G} is tridiagonal, we can apply Thomas algorithm [18] to solve (27) in linear time.

E. Full-Chip Steady-state Thermal Simulation

Once the coefficients $c_{il}(z_t)$'s and the Fourier-series expansion of each $\phi_{il}(x, y)$ from (17) have been obtained, the rising temperature at each grid cell of l -th layer, denoted as $T_{mn}(z_t)$, is expressed to be the average of $\hat{T}(\mathbf{r})$ over the domain of grid cell (m, n) .

$$\begin{aligned} T_{mn}(z_t) &= \frac{1}{\Delta x \Delta y} \sum_{l=0}^{N_{By}-1} \sum_{i=0}^{N_{Bx}-1} \frac{c_{il}(z_t)}{\eta_{il}} \int_{n\Delta y}^{(n+1)\Delta y} \int_{m\Delta x}^{(m+1)\Delta x} \varphi_{il}(x, y) dx dy \\ &= \sum_{j=0}^{N_x^i-1} \sum_{k=0}^{N_y^i-1} \langle \mathbf{f}_{jk}(z_t), \boldsymbol{\beta}_{jk}^{mn} \rangle. \end{aligned} \quad (28)$$

Here, $\mathbf{f}_{jk}(z_t)$ is the temperature spectrum and is computed as

$$\mathbf{f}_{jk}(z_t) = \sum_{l=0}^{N_{By}-1} \sum_{i=0}^{N_{Bx}-1} \frac{c_{il}(z_t)}{\eta_{il}} \mathbf{a}_{jk}^{il}. \quad (29)$$

The above summation can be efficiently evaluated using pre-calculated Fourier coefficients and normalization constants stored in lookup tables. Finally, $T_{mn}(z_t)$ for each grid cell of each layer can be obtained by (28) using 2D FFT.

F. Full-Chip Transient Thermal Simulation

The transient simulation of FALCON-3D builds directly upon the steady-state formulation described from Section IV-A to Section IV-E. The rising temperature distribution and power density in (1)–(4) are replaced by $T(\mathbf{r}, t)$ and $p(\mathbf{r}, t)$, respectively, and the governing equation (1) with the time-derivative term becomes

$$\sigma(\mathbf{r}) \frac{\partial T(\mathbf{r}, t)}{\partial t} + \nabla \cdot (\kappa(\mathbf{r}) \nabla T(\mathbf{r}, t)) = -p(\mathbf{r}, t), \quad (30)$$

where $\sigma(\mathbf{r})$ is the product of material density and specific heat and $p(\mathbf{r}, t)$ is the time-varying power density.

The spatial expansion introduced in (5) remains unchanged, with the coefficients $c_{il}(z, t)$'s now evolving over time,

$$T(\mathbf{r}, t) \approx \hat{T}(\mathbf{r}, t) = \sum_{l=0}^{N_{By}-1} \sum_{i=0}^{N_{Bx}-1} c_{il}(z, t) \frac{\varphi_{il}(x, y)}{\eta_{il}}. \quad (31)$$

Since $\varphi_{il}(x, y)$'s and η_{il} 's depend only on the lateral dimensions, \mathbf{a}_{jk}^{il} 's are identical to the steady-state case and can be reused without recomputation at each time step.

Along the z -direction, the 1D subproblem is extended by including the time derivative in (22). Its right-hand side is characterized by the time-varying projection coefficients $\theta_{il,t}(t)$'s,

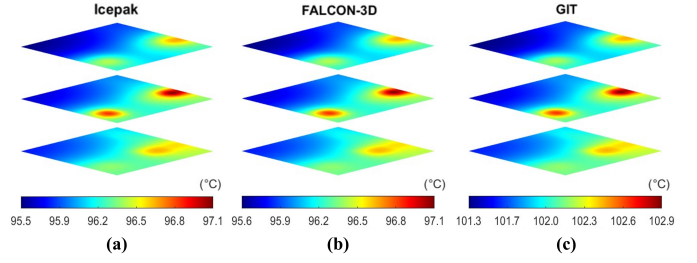


Fig. 4: Thermal profiles of case 3D-C in three active layers. (a) Icepak. (b) FALCON-3D. (c) GIT [14].

which are defined similar to (19)–(21) but with instantaneous powers $p_t(x, y, t)$'s.

We then discretize the simulation period by employing the backward Euler scheme for time marching procedure and $c_{il}(z_t, t_n)$'s, at the sampling points z_t 's and sampling time $t_n = n\Delta t$, can be obtained by solving

$$\left(\mathbf{G} + \frac{\mathbf{C}}{\Delta t} \right) \mathbf{x}^n = \boldsymbol{\theta}^n + \frac{\mathbf{C}}{\Delta t} \mathbf{x}^{n-1}, \quad (32)$$

where \mathbf{C} is a diagonal thermal capacitance matrix, Δt is the time step size, \mathbf{x}^n is the vector of $c_{il}(z_t, t_n)$'s, and $\boldsymbol{\theta}^n$ is the vector of $\theta_{il,t}(t_n)$'s.

Because of the tri-diagonal form of $\mathbf{G} + \mathbf{C}/\Delta t$, (32) can also be solved efficiently in linear time.

Finally, the transient temperature waveform can be obtained through the same reconstruction procedure as in (28)–(29) at each time step.

V. EXPERIMENTAL RESULTS

FALCON-3D is implemented in C++ language, and all experiments are conducted on a PC equipped with an Intel Core i5-14400F 2.50 GHz CPU and 128 GB of RAM. To evaluate the accuracy and effectiveness of FALCON-3D, we compare it with GIT [14] and use the commercial thermal simulation tool Icepak [2] as the reference solution. The GIT solver is obtained from the authors of [14]. The FFT operations are performed using the Intel Math Kernel Library (MKL) [19], which serves as the computational kernel for FFT in our implementation.

The simulation structure is the stacked-layer 3D IC shown in Fig. 1(a). The thermal conductivities (W/m \cdot °C) for different materials are set as, metal: 194, bonding: 50, silicon: 148, and substrate: 10 [20]. The ambient temperature is 25 °C.

A. Accuracy and Efficiency of FALCON-3D for Steady-state Simulation

To evaluate the performance of FALCON-3D for steady-state simulation, we construct three test cases, 3D-A, 3D-B, and 3D-C with one, two, and three active layers, respectively. The boundary conditions are the same as shown in Section II and the power distribution in each case is derived from a realistic power density map. The power density values are constrained within the range of 4.4×10^{-2} W/mm 3 to 15 W/mm 3 , reflecting practical operating conditions.

TABLE I: Comparison of Thermal Simulation Performance

| Case | #Grids | FALCON-3D | | | | | GIT [14] | | | | | Icepak |
|------|----------|-------------|-------------|----------|----------|---------------|-------------|-------------|----------|----------|---------------|-------------|
| | | Pre-Cal (s) | Solving (s) | MAE (°C) | MAPE (%) | Max.Err. (°C) | Pre-Cal (s) | Solving (s) | MAE (°C) | MAPE (%) | Max.Err. (°C) | Solving (s) |
| 3D-A | 786432 | 0.242 | 0.240 | 0.025 | 0.05 | 0.104 | 0.002 | 0.006 | 1.070 | 1.94 | 1.078 | 355.5 |
| | 3145728 | 0.892 | 0.962 | 0.025 | 0.04 | 0.120 | 0.029 | 0.080 | 1.070 | 1.94 | 1.077 | 1753.9 |
| | 6220800 | 1.744 | 1.713 | 0.025 | 0.04 | 0.150 | 0.050 | 0.131 | 1.070 | 1.94 | 1.077 | 3177.9 |
| | 12582912 | 3.411 | 4.595 | 0.025 | 0.04 | 0.189 | 0.099 | 0.306 | 1.070 | 1.94 | 1.076 | 6179.1 |
| 3D-B | 1048576 | 0.248 | 0.321 | 0.036 | 0.04 | 0.160 | 0.004 | 0.007 | 2.546 | 3.00 | 2.559 | 536.1 |
| | 4194304 | 0.902 | 1.258 | 0.035 | 0.04 | 0.187 | 0.036 | 0.101 | 2.546 | 3.00 | 2.557 | 2577.5 |
| | 8294400 | 1.725 | 2.320 | 0.036 | 0.04 | 0.234 | 0.070 | 0.178 | 2.546 | 3.00 | 2.557 | 5336.3 |
| | 16777216 | 3.380 | 6.293 | 0.036 | 0.04 | 0.295 | 0.130 | 0.399 | 2.546 | 3.00 | 2.557 | 9077.2 |
| 3D-C | 1310720 | 0.345 | 0.409 | 0.029 | 0.03 | 0.127 | 0.013 | 0.033 | 5.883 | 6.12 | 5.900 | 412.0 |
| | 5242880 | 1.275 | 1.636 | 0.029 | 0.03 | 0.127 | 0.046 | 0.128 | 5.883 | 6.12 | 5.900 | 3706.4 |
| | 10368000 | 2.377 | 2.926 | 0.029 | 0.03 | 0.126 | 0.083 | 0.226 | 5.883 | 6.12 | 5.900 | 9172.4 |
| | 20971520 | 4.668 | 7.444 | 0.029 | 0.03 | 0.146 | 0.158 | 0.501 | 5.883 | 6.12 | 5.900 | 12672.4 |

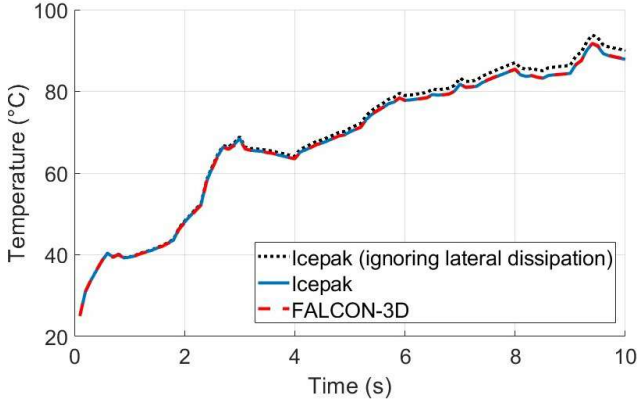


Fig. 5: Transient temperature waveform of the hotspot in the central active layer of 3D-C from FALCON-3D, Icepak, and Icepak ignoring lateral dissipation.

Table I compares the steady-state performance across cases 3D-A to 3D-C. “#Grids” column reports the total number of grid cells, calculated as the product of lateral spatial resolution and the number of stacked layers in each case. In addition to the MAE, the table also reports the maximum percentage error (MAPE) and the maximum absolute error (Max Err.). Both FALCON-3D and GIT significantly demonstrate much higher efficiency than Icepak, achieving over 1000× speedup in all cases. Among them, GIT achieves the shortest runtime, primarily due to its neglect of lateral heat dissipation.

However, as the number of stacked layers increases from 3D-A to 3D-C, representing progressively deeper 3D IC integration, the error of GIT increases substantially. Specifically, its MAE rises from 1.070 °C and reaches 5.883 °C in 3D-C. In contrast, FALCON-3D, which explicitly models lateral convection, maintains MAE below 0.036 °C across all cases, highlighting its robustness for accurate thermal simulation.

Fig. 4(a), (b), and (c) present the temperature maps of case 3D-C in three active layers obtained by Icepak, FALCON-3D, and GIT, respectively. Compared with Icepak, it demonstrates the high accuracy of FALCON-3D and shows the significant error of GIT.

B. Performance of FALCON-3D for Transient Simulation

In transient simulation, three time-varying power maps are applied to 3D-C. The power waveforms are generated by as-

signing random on/off durations to emulate realistic workload variations for each grid cell with the power density range defined previously. The simulation period is 10 s, the time step size is 0.1 s, and the grid number is 5,242,880.

Fig. 5 plots the temperature waveform of hotspot in the central active layer of 3D-C obtained from FALCON-3D, Icepak, and Icepak ignoring lateral dissipation. The waveform simulated by FALCON-3D (orange dashed line) shows good consistency with Icepak (blue solid line) throughout the simulation period, with MAE and maximum absolute error below 0.068 °C, and 0.235 °C, respectively. In contrast, the Icepak result ignoring lateral dissipation (black dotted line) deviates significantly, with errors up to 2.27 °C. Furthermore, under the same workload and grid resolution, Icepak requires 19 hours and 9 minutes for simulation, whereas FALCON-3D completes in 221.7 seconds, yielding a 312× speedup without compromising accuracy.

These results establish FALCON-3D as an efficient and scalable thermal simulator that preserves accuracy while capturing the essential effects of lateral dissipation.

VI. CONCLUSION

We have represented a novel analytical framework for full-chip thermal simulation, FALCON-3D, capable of capturing lateral heat dissipation with high efficiency. By leveraging an eigenfunction expansion with convective boundary conditions and approximating solutions via Fourier series and the truncation scheme, the method enables efficient computation through FFT, while precomputed lookup tables further accelerate simulation.

Moreover, FALCON-3D inherently allows parallelization, as the independence in Fourier projections, FFT operations, and z-direction solvers can be mapped efficiently onto multi-core CPUs and GPUs for further acceleration. These capabilities position FALCON-3D as a scalable, accurate, and practical solution for integration into early-stage thermal-aware design flows for modern 3D ICs.

ACKNOWLEDGEMENT

This work was partially supported by the National Science and Technology Council (NSTC) in Taiwan, 114-2640-E-A49-001.

REFERENCES

- [1] S. S. Salvi and A. Jain, "A review of recent research on heat transfer in three-dimensional integrated circuits (3-D ICs)," *IEEE Trans. Compon. Packag. Manuf. Technol. (TCPMT)*, vol. 11, no. 5, pp. 802–821, 2021.
- [2] ANSYS Icepak, <https://www.ansys.com/products/electronics/ansys-icepak>.
- [3] S. Liu, C. Wang, Z. Yu, W. Tang, and W. Zhuang, "Thermal-WLP: A transient thermal simulation method based on weighted laguerre polynomials for 3-D ICs," *IEEE Trans. Compon. Packag. Manuf. Technol. (TCPMT)*, vol. 7, no. 3, pp. 405–411, 2017.
- [4] T.-Y. Wang and C. C.-P. Chen, "3-D Thermal-ADI: a linear-time chip level transient thermal simulator," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. (TCAD)*, vol. 21, no. 12, pp. 1434–1445, 2002.
- [5] B.-W. C. Chen, Y.-H. Lin, C.-Y. Lin, and Y.-M. Lee, "ThPA: Thermal simulation for advanced ICs," in *ASP Design Autom. Conf. (ASP-DAC)*, 2026.
- [6] Y.-M. Lee, T.-H. Wu, P.-Y. Huang, and C.-P. Yang, "NUMANA: A hybrid numerical and analytical thermal simulator for 3-D ICs," in *Proc. Design. Autom. Test Europe Conf. Exhib. (DATE)*, pp. 1379–1384, 2013.
- [7] Y.-M. Lee, C.-W. Pan, P.-Y. Huang, and C.-P. Yang, "LUTSim: A look-up table-based thermal simulator for 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. (TCAD)*, vol. 34, no. 8, pp. 1250–1263, 2015.
- [8] Y. Zhan and S. S. Sapatnekar, "High-efficiency Green function-based thermal simulation algorithms," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. (TCAD)*, vol. 26, no. 9, pp. 1661–1675, 2007.
- [9] D. Oh, C. C. P. Chen, and Y. H. Hu, "3DFFT: Thermal analysis of non-homogeneous IC using 3D FFT Green function method," in *Int. Symp. on Qual. Electron. Design (ISQED)*, pp. 567–572, 2007.
- [10] B. Wang and P. Mazumder, "Accelerated chip-level thermal analysis using multilayer Green's function," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. (TCAD)*, vol. 26, no. 2, pp. 325–344, 2007.
- [11] K. Wang and Z. Pan, "An analytical model for steady-state and transient temperature fields in 3-D integrated circuits," *IEEE Trans. Compon. Packag. Manuf. Technol. (TCPMT)*, vol. 6, no. 7, pp. 1026–1039, 2016.
- [12] H. Sultan and S. R. Sarangi, "Varsim: A fast and accurate variability and leakage aware thermal simulator," in *57th Design Autom. Conf. (DAC)*, pp. 1–6, 2020.
- [13] 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 Trans. Very Large Scale Integr. Syst. (TVLSI)*, vol. 22, no. 11, pp. 2366–2379, 2014.
- [14] P.-Y. Huang and Y.-M. Lee, "Full-chip thermal analysis for the early design stage via generalized integral transforms," *IEEE Trans. Very Large Scale Integr. Syst. (TVLSI)*, vol. 17, no. 5, pp. 613–626, 2009.
- [15] Y. Cho, H. Choi, H. Lee, Y. Im, H. Lee, and Y. Shin, "Thermal aware 3-D floorplanning on multi-stacked board of smart phone," in *IEEE Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 636–641, 2020.
- [16] C.-W. Wu, Y.-M. Lee, P.-Y. Huang, B.-J. Yang, T.-Y. Chen, T.-C. Huang, and Y.-L. Lee, "SpotLight: A hotspot-greedy, light-weighted, and automated thermal modeling framework for early smartphone design," in *Int. Symp. on Qual. Electron. Design (ISQED)*, pp. 1–8, 2024.
- [17] M. D. Mikhailov and M. N. Ozisik, *Unified Analysis and Solutions of Heat and Mass Diffusion*. Dover, 1994.
- [18] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, *Numerical Recipes in C++: the Art of Scientific Computing*. USA: Cambridge University Press, 2nd ed., 2001.
- [19] Intel Corporation, "Intel Math Kernel Library FFT Functions." <https://www.intel.com/content/www/us/en/docs/onemkl/developer-reference-c/2023-2/fft-functions.html>, 2023.
- [20] K. O. Petrosyants and N. I. Ryabov, "Quasi-3D Thermal Simulation of Integrated Circuit Systems in Packages," *Energies*, vol. 13, no. 12, 2020.